MASTER

Visualization of data from longitudinal studies concerning multi-variant diseases

van Wezel, B.J.P.A.

Award date:
2016

Link to publication
Visualization of Data from Longitudinal Studies Concerning Multi-Variant Diseases

Master Thesis

B.J.P.A. van Wezel

Supervisors:
Prof. dr.ir. Jarke J. van Wijk
L. Soetens MSc

Final version

Eindhoven, November 2016
Abstract

The Center for Infectious Disease Control is the national knowledge center and coordinator in the infectious disease domain. One way of reducing health problems related to infectious diseases is by vaccination. For some infectious diseases, such as the Human papilloma virus (HPV) infection and pneumococcal disease, persons can get infected by multiple variants at the same time. Because vaccination only works against some variants of these diseases, the effect of vaccination on these diseases are still unclear. Studies are being conducted to investigate this. The information in those studies is quite complex.

Therefore we designed an interactive exploratory visualization tool, ComboVis. ComboVis can help epidemiologist answer questions about the data from longitudinal studies concerning multi-variant diseases. Multiple visualization techniques that are being used for data in a similar format have been discussed with multiple epidemiologists. Based on the feedback, we combined different visualization techniques. During the design phase, an early user evaluation was conducted. The feedback from the early user evaluation is processed in the next version of ComboVis. Finally, another user evaluation with another group of epidemiologists was conducted, which resulted in enthusiastic and positive reactions and the final version of ComboVis.
Preface

After many years of hard work in college, I conclude my studies with this thesis. Therefore, I would like to take this opportunity to thank all the people who have supported me in these years.

First I want to thank my TUe supervisor Jarke J. van Wijk. Jack was an enthusiastic supervisor, who came up with useful comments, suggestions, and advice every week. The meetings on Mondays provided me with motivation for the whole week.

I want to thank my supervisors at the RIVM, Loes Soetens and Irene Man. Loes and Irene provided me with useful weekly feedback and insights in the problems of longitudinal multi-variant diseases. I want to thank you for the smooth cooperation and the freedom that I got during my internship.

I would also like my third committee member Joos Buijs, for completing the assessment committee and providing this TUe LaTeX Master Thesis template.

Furthermore, I would like to thank all my family members, friends, and colleagues. Finally, I would like to thank my girlfriend for supporting and motivating me during my studies.

Bart van Wezel
## Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Objectives</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Structure</td>
<td>2</td>
</tr>
<tr>
<td>2 Background</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Human papillomavirus</td>
<td>3</td>
</tr>
<tr>
<td>2.2 Domain problems</td>
<td>3</td>
</tr>
<tr>
<td>3 Problem statement</td>
<td>5</td>
</tr>
<tr>
<td>3.1 Data representation</td>
<td>5</td>
</tr>
<tr>
<td>3.2 Tasks</td>
<td>8</td>
</tr>
<tr>
<td>3.3 Requirements</td>
<td>10</td>
</tr>
<tr>
<td>4 State of art</td>
<td>13</td>
</tr>
<tr>
<td>4.1 Related work</td>
<td>13</td>
</tr>
<tr>
<td>4.1.1 Euler-based</td>
<td>13</td>
</tr>
<tr>
<td>4.1.2 Node link diagrams</td>
<td>13</td>
</tr>
<tr>
<td>4.1.3 Aggregation based methods</td>
<td>15</td>
</tr>
<tr>
<td>4.1.4 Matrix</td>
<td>16</td>
</tr>
<tr>
<td>4.1.5 Linear diagram</td>
<td>18</td>
</tr>
<tr>
<td>4.2 Comparison</td>
<td>19</td>
</tr>
<tr>
<td>5 Design</td>
<td>24</td>
</tr>
<tr>
<td>5.1 Tasks</td>
<td>25</td>
</tr>
<tr>
<td>5.1.1 Higher level reasoning about the sets</td>
<td>25</td>
</tr>
<tr>
<td>5.1.2 Interrelation between set memberships and attributes</td>
<td>27</td>
</tr>
<tr>
<td>5.1.3 Development of the sets and combinations over time</td>
<td>29</td>
</tr>
<tr>
<td>5.2 Visualization Techniques</td>
<td>31</td>
</tr>
<tr>
<td>5.2.1 Brushing and linking</td>
<td>31</td>
</tr>
<tr>
<td>5.2.2 Selection</td>
<td>32</td>
</tr>
<tr>
<td>5.2.3 Details</td>
<td>33</td>
</tr>
<tr>
<td>6 Evaluation</td>
<td>35</td>
</tr>
<tr>
<td>6.1 Structure of the evaluations</td>
<td>35</td>
</tr>
<tr>
<td>6.2 Early user evaluation</td>
<td>35</td>
</tr>
<tr>
<td>6.2.1 Observations</td>
<td>35</td>
</tr>
<tr>
<td>6.2.2 Results</td>
<td>38</td>
</tr>
<tr>
<td>6.2.3 Improvements</td>
<td>39</td>
</tr>
<tr>
<td>6.3 Final user evaluation</td>
<td>41</td>
</tr>
<tr>
<td>6.3.1 Observations</td>
<td>41</td>
</tr>
<tr>
<td>6.3.2 Results</td>
<td>41</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>6.3.3 Improvements</td>
<td>43</td>
</tr>
<tr>
<td>7 Discussion</td>
<td>45</td>
</tr>
<tr>
<td>7.1 Scalability</td>
<td>45</td>
</tr>
<tr>
<td>7.1.1 Sets</td>
<td>45</td>
</tr>
<tr>
<td>7.1.2 Combinations</td>
<td>45</td>
</tr>
<tr>
<td>7.1.3 Elements</td>
<td>46</td>
</tr>
<tr>
<td>7.2 Other applications</td>
<td>47</td>
</tr>
<tr>
<td>7.2.1 Movies</td>
<td>48</td>
</tr>
<tr>
<td>7.2.2 Actors</td>
<td>49</td>
</tr>
<tr>
<td>7.3 Future work</td>
<td>50</td>
</tr>
<tr>
<td>8 Conclusions</td>
<td>53</td>
</tr>
<tr>
<td>Bibliography</td>
<td>55</td>
</tr>
<tr>
<td>Appendix</td>
<td>59</td>
</tr>
<tr>
<td>A Statistics</td>
<td>59</td>
</tr>
<tr>
<td>A.1 Deviance</td>
<td>59</td>
</tr>
<tr>
<td>A.2 Fisher's test</td>
<td>61</td>
</tr>
<tr>
<td>B Evaluation Questions</td>
<td>62</td>
</tr>
<tr>
<td>B.1 Early user evaluation</td>
<td>62</td>
</tr>
<tr>
<td>B.1.1 Tasks</td>
<td>62</td>
</tr>
<tr>
<td>B.1.2 Questionnaire</td>
<td>63</td>
</tr>
<tr>
<td>B.2 Final user evaluation</td>
<td>64</td>
</tr>
<tr>
<td>B.2.1 Tasks</td>
<td>64</td>
</tr>
<tr>
<td>B.2.2 Questionnaire</td>
<td>64</td>
</tr>
<tr>
<td>C Implementation</td>
<td>66</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

How can we keep ourselves and our environment healthy? That is the challenge government authorities face at all levels, from local to international. The Dutch National Institute for Public Health and the Environment (RIVM) carries out independent research and provides policy advice to assist them in this task. Driven by public health challenges, they collect data and carry out research in close collaboration with universities and research centers worldwide. The RIVM organization consists of three domains with specific knowledge and expertise: Infectious Diseases and Vaccinology (Center for Infectious Disease Control), Environment and Safety and Public Health and Health Services.

The Center for Infectious Disease Control (CIb) is the national knowledge center and coordinator in the infectious disease domain. They provide policy advice to the Dutch Government and support professionals in health care and public health. At the CIb the control of infectious diseases is coordinated, including effective prevention, close vigilance and quick response in the event of an outbreak. Therefore the CIb contributes to reducing health problems related to infectious diseases.

One way of reducing health problems related to infectious diseases is by vaccination. With vaccination, the immune system is stimulated to develop immunity against the corresponding infectious disease. Studies are being conducted to monitor the effects of vaccination at a population level and are mainly focused at the effectiveness of vaccination.

For some infectious diseases, such as human papilloma virus (HPV) infection and pneumococcal disease, persons can get infected by multiple variants at the same time. Currently vaccination against these diseases only protects against some of the variants. This makes it difficult to estimate the effect of vaccination for these diseases at a population level, as we cannot predict the interaction effects between those variants.

Possible interaction effects between multiple variants are competition and synergy. When two variants compete for the same resources, type replacement is likely, since vaccination against one variant makes the resources available for the other variant. With type replacement another variant replaces the variant against which is vaccinated, and hence the effect of the vaccination is neutralized. In some cases the effect is not only neutralized, but even worsened, as non-vaccine variants might grow in prevalence. Synergy occurs when one variant likes to grow together with another variant: if people are vaccinated against one variant, then the other variant is likely to decrease in prevalence as well. Therefore, synergy and type replacement can both influence the effectiveness of the vaccination.

Much research has been done to investigate the interactions between different variants of these diseases. To investigate interactions, researchers mostly study the co-occurrence of variants. When two variants do not occur as much together as expected, it might be a sign for competitive interaction, and thus a sign for type replacement. When two variants occur more frequently together than expected, it might be a sign for synergy. This is not only applicable for the co-occurrence of two variants, but also for the co-occurrence of three or more variants.

However, the number of possible combinations of variants increases rapidly, when the number of variants increases. It is not very likely that all variants reside in one person, but with 25 variants over 30 million possible combinations of variants exists, since for \( n \) variants, \( 2^n \) possible combinations of variants exists. When we want to compare attributes among those combinations, such as the vaccina-
tion status of persons or different timestamps, the number of unique parts only increases. Interactive visualizations can help with the exploratory analysis and hypothesis formulation in the data from longitudinal studies concerning multi-variant infectious diseases.

### 1.1 Objectives

The focus of this thesis is on the interactive visualization of the data from longitudinal studies concerning multi-variant infectious diseases. The goal is to help researchers understand the collected data more efficiently. To provide some guidelines, we formulated the following research questions for this Thesis.

1. Can we use visualizations to display deviating combinations? For example: Can we find all combinations of variants that occur significantly more than expected?

2. Can we use interactive visualization to analyze combinations in the context of other attributes? For example: Can we find all combinations of variants that occur more often in non-vaccinated persons compared to vaccinated persons?

3. Can we use interactive visualization to display longitudinal infectious disease data? For example: Can we visualize which variant people had preceding or following a specific combination of multiple variants?

### 1.2 Structure

In Chapter 2, the background is discussed. The problem formulation and the requirements are covered in Chapter 3. In Chapter 4, the state of art is discussed. Our solution is presented in Chapter 5. The evaluation of our solution is described in Chapter 6. The limitations and possibilities are discussed in Chapter 7. In Chapter 8, the conclusions are presented.
Chapter 2

Background

2.1 Human papillomavirus

A persistent infection with human papilloma virus (HPV) is the cause of almost all cases of cervical cancer[21]. 80% of the sexually active women are infected with HPV at one point during their lives. Fortunately, in most cases the body clears the virus properly. Therefore, being infected with HPV does not always result in cervical cancer. There are multiple variants of the HPV, called the HPV-types. Not all the HPV-types have the same risk on progression to cancer. For around 15 types, it has been demonstrated that they increase the risk of cancer[20]. These types are called the high-risk HPV (hr-HPV). The other types are called the low-risk HPV (lr-HPV).

The vaccination against HPV only protects persons against a subset of all types. The hr-HPVs type 16 and type 18 are the most occurring hr-HPV types. In The Netherlands young girls are being vaccinated against those two types within the national immunization program since 2010. Because little is known regarding the interaction effects between different HPV-types, it is hard to predict the complete effect of the vaccination. The vaccination against HPV started recently and thus only the short-term effects of vaccination can be measured. Several of the HPV studies, such as [7, 14, 25, 15], try to predict the effect of vaccination by looking at the co-occurrence of HPV-types. Multiple studies are investigating this interaction between different HPV-types, such as [4, 24].

HAVANA study The HAVANA (HPV Amongst Vaccinated And Non-vaccinated Adolescents) study has been designed to assess the effects of vaccination [19, 23, 18]. Girls between the age of 14 and 16 were randomly asked to join the study in 2009. A year later another group of girls was included in the study in order to increase the precision of the study. In total 1800 girls joined the study. Of those girls, 1037 were vaccinated. They are asked to supply a vaginal self-swab to test for different HPV-types each year. The presence of HPV-types is assessed by the SPF10 testing method[11]. The SPF10 testing method tests for 12 hr-HPV types and 12 lr-HPV types. In addition, they fill out a small questionnaire concerning possible risk factors and background characteristics.

2.2 Domain problems

Competition or synergy When two variants do not occur as much together as expected, it might be a sign for competition between the two types. When two variants occur more frequently together than expected, it might be a sign for synergy between the two types. The deviation from the expected value can be an indicator for both competition and synergy. The expected value of a combination of multiple types depends on the occurrence of the HPV-types and the total number of people included within the study. When a combination occurs more frequently than expected, it does not always mean that there is synergy between the two types, and that cross protection will occur when persons are vaccinated against one of those types. The combination of types can also occur more frequently than expected because of other variables, which act as confounding factors. For instance, if both types occur more
frequently in women who had a relation longer than six months, those two types also occur more frequently together, compared with other types that occur more frequently in women without a relation.

Another problem that arises when analyzing the data in this study, is that the same people are followed-up every year and hence the samples are not independent. For example, someone whose sample contains HPV-types 16 and 18 are more likely to have HPV-types 16 and 18 in her sample next year, as compared to someone who had no HPV-types in her sample.

**Infection development over time** The infection development over time is still being investigated, especially concerning possible attributes that can influence the clearance of HPV-types. For example, are HPV-types 16 and 18 cleared faster by the body in vaccinated persons compared to non vaccinated persons? In addition, it might also be possible that certain HPV-types influence the probability of being infected with another HPV-type in the future, even if that HPV-type is not longer present. For example, by being infect with HPV-type 16, persons can have a higher or lower probability on acquiring another HPV-type in the future. Finally, certain HPV-types could also affect the clearance rate of other HPV-types that are present at the same time. For example, in persons that have both HPV-type 16 and 18, is the probability on clearance of one of those types different from the clearance rate in persons who have only HPV-type 16 or only HPV-type 18? Such questions are hard to answer, as following a large group of people over time is difficult, due to the costs and the drop out of persons.
Chapter 3

Problem statement

In this Chapter we describe the data and the tasks that we want to perform on the data. In Section 3.1 we start with a description of the structure of the data in the HAVANA study. Thereafter, we describe how this data translates to a more abstract data format, such that we can apply our solutions to other data sets with similar problems as well. In Section 3.2 we describe the tasks we would like to perform. We conclude the Chapter with the requirements for a solution in Section 3.3.

3.1 Data representation

Data in the study  The data format from the HAVANA study is shown in Table 3.1. We show HPV-type 1 till HPV-type 25 as example, however in the study we use the HPV-types that are detected by the SPF10 testing method. Each row represents a sample from a person at a certain timestamp. In the example, person 1 did send in two samples. Only HPV-type 4 was found in both samples. Only HPV-type 4 was found in both samples. The rows represent all information belonging to the measured sample, corresponding to a specific person and a specific timestamp. Since we have additional information about the persons that sent in the sample, we have other information about the samples as well, for instance, such as the vaccination status of the person the sample belongs to. All these attributes are stored in one row, as some attributes can change over time. For example, the number of sex partners in the last six months can be different on each timestamp for the same person.

<table>
<thead>
<tr>
<th>ID</th>
<th>Person</th>
<th>Vaccination</th>
<th>Time</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
<th>...</th>
<th>Type 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>1</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m2</td>
<td>1</td>
<td>Yes</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m3</td>
<td>2</td>
<td>No</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m4</td>
<td>2</td>
<td>No</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m5</td>
<td>2</td>
<td>No</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m6</td>
<td>3</td>
<td>Yes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: A sample from the HAVANA data

Set-Typed Data  The data format of the HAVANA study can be translated into a Set Typed data format. The Set-typed data format represents each set as a collection of elements. In case of the HAVANA study this means that each sample is an element, so each set is a collection of samples. We can create a set of all samples in which a certain HPV-type was found. For instance the set of samples in which HPV-type 1 was found in Table 3.1 is \{m3, m4, m5, m6\}. Since there are also attributes in the Set-Typed data format, we do not have to translate these. Figure 3.1 shows the translation from the HAVANA study data into the Set-typed data format.
CHAPTER 3. PROBLEM STATEMENT

Figure 3.1: How the data of the HAVANA study translates to the Set-Typed data format

We can transform the samples from the data in Table 3.1 into the Set-Typed data format. The result of this transformation is shown in Table 3.2. The 0 or 1 beneath each set represents whether or not that element is in that set. Here $S_3$ simply denotes the set of samples in which type 3 was found. Since we also want to analyze the infection development over time we defined two attributes of which we assume they are always present in the data, namely source and time. The source attribute identifies the person that sent in the sample in the HAVANA study. This source attribute is needed to investigate the samples from one or more persons. Beside the source attribute, we also need an attribute for time. The time attribute is needed to analyze the infection development over time, as we have to know which sample from a person was preceding or following that sample.

Sets are a universal concept in scientific data analysis. There are many data sets to which the Set-Typed data format can be applied. In soil samples, collections of samples with similar bacterial species can be treated as a set. In movies, collections of movies with a similar genre can be treated as a set. Multiple relations and operations between sets are possible. The following notations are useful to reason about relations between sets and to obtain new insights.

- Containment $A \subseteq B$: All elements that are in set $A$, are also in set $B$. $S_2$ contains the elements $\{e_4, e_5, e_6\}$, $S_1$ contains the elements $\{e_3, e_4, e_5, e_6\}$, so $S_2 \subseteq S_1$.

- Union $A \cup B$: All elements that are in set $A$ or in set $B$. $S_3$ contains the elements $\{e_3\}$, $S_4$ contains the elements $\{e_1, e_2\}$, so $S_3 \cup S_4 = \{e_1, e_2, e_3\}$.

- Intersection $A \cap B$: All elements that are in both set $A$ and set $B$. $S_3$ contains the element $\{e_3\}$, $S_1$ contains the elements $\{e_3, e_4, e_5, e_6\}$, so $S_3 \cap S_1 = \{e_3\}$.

- Complement $A \setminus B$: All elements that are in set $A$, but not in set $B$. $S_2$ contains the element $\{e_4, e_5, e_6\}$, $S_1$ contains the elements $\{e_3, e_4, e_5, e_6\}$, so $S_1 \setminus S_2 = \{e_3\}$.

- Cardinality $|A|$: The number of elements that are in set $A$. $S_1$ contains the elements $\{e_3, e_4, e_5, e_6\}$, since $S_1$ contains four elements, so $|S_1| = 4$

Table 3.2: General Set-Typed data

<table>
<thead>
<tr>
<th>Element</th>
<th>Source attribute</th>
<th>Attribute 1</th>
<th>Time attribute</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>...</th>
<th>$S_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>1</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$e_2$</td>
<td>1</td>
<td>Yes</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$e_3$</td>
<td>2</td>
<td>No</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$e_4$</td>
<td>2</td>
<td>No</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$e_5$</td>
<td>2</td>
<td>No</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$e_6$</td>
<td>3</td>
<td>Yes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
CHAPTER 3. PROBLEM STATEMENT

Data notation  We expand the formal representation as described in The State-of-the-Art of Set Visualization by Alsallakh et al.[2]. We use this representation in the remainder of the Thesis, since it allows us to define precisely which parts of the data we want to see. We give each notation a name, an informal and a formal definition, followed by an example based on Table 3.2, and finally how the definition applies to the HAVANA data set. We define the following notations:

**Element**  An element in the data set.
\( e_i \), where \( i \) is the identifier of the element
\( e_2 \) represents the element in the second row
A sample from a person in the HAVANA study.

**Elements**  The elements in the data set.
\( M = \{ e_1 \ldots e_n \} \), where \( n \) is the number of elements in the data set
All elements in the data set are \{ \( e_1 \), \( e_2 \), \( e_3 \), \( e_4 \), \( e_5 \), \( e_6 \) \}
All samples in the HAVANA study.

**Set**  A collection of elements in the data set that share a defining property. In our case all samples in which a certain HPV-type is found.
\( S_i \subseteq M \), for \( 1 \leq i \leq m \), where \( m \) is the number of sets in the data set
All elements in \( S_1 \) are \{ \( e_3 \), \( e_4 \), \( e_5 \), \( e_6 \) \}
All samples that contain HPV-type 16.

**Set identifiers**  The identifiers for the sets.
\( F_S = \{1 \ldots m\} \), where \( m \) is the number of sets.
The set identifiers are \( \{1,2,3,4\} \).
\( F_S = \{6 \ldots 74\} \) denotes all HPV-types in the HAVANA study.

**Combination identifier**  An identifier for a combination of sets, which is a collection of set identifiers.
\( I \subseteq F_S \), which represents a subset of set identifiers.
A possible subset of the set identifiers is \{1,3\}.
A combination of HPV-types.

**Combination Identifiers**  All possible identifiers for the combinations
\( F_C = \{ I | I \subseteq F_S \wedge |I| > 1 \} \)
The possible combinations are \( \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\} \ldots \{1,2,3,4\} \)
All possible combinations of different HPV-types.

**Combination**  All elements that are in all sets of the combination identifier.
\( C_I = \bigcap_{i \in I} S_i \), for an \( I \in F_C \)
All elements that are in the combination of \( S_1, S_2 \) are \( \{e_4, e_5, e_6\} \)
All samples that have both HPV-type 16 and 18.

**Element value for an attribute**  The value of an attribute for an element
\( e_{iA} \), which represents the value of an attribute \( A \) for an element \( e_i \)
\( e_{1A} = v_{A,1} \), since the value for \( A_1 \) for \( e_1 \) is yes, which is the first value of \( A_1 \)
The age of the person of the sample.

**Range of an attribute**  The possible values for an attribute in the data set, which must be discrete.
\( R_A = \{v_{A,1} \ldots v_{A,q}\} \), where \( q \) is the number of different values for attribute \( A \).
\( R_{A,1} = \{v_{A,1}, v_{A,2}\} \), since attribute 1 has two different answers.
All possible answers for an attribute, such as yes or no for vaccinated, which are represented by \( v_{vac,1} \) and \( v_{vac,2} \).
CHAPTER 3. PROBLEM STATEMENT

**Identifiers for the values** The identifiers for the possible values for an attribute in the data set.

\( Q_A = \{1 \ldots q \} \), where \( q \) is the number of different values for attribute \( A \).

\( Q_A = \{1, 2\} \), since attribute 1 has two different answers.

The identifiers for the values of an attribute, such as \( Q_{vac} = 1, 2 \), where the 1 represents yes and 2 represents no.

**Elements of a value** All elements that share the same value for an attribute.

\( E_A, i = \{e \mid e \in M \land e_A = v_A, i\} \)

\( E_A, 1 = \{e_1, e_2, e_6\} \) and \( E_A, 2 = \{e_3, e_4, e_5\} \).

All samples that share the same value for an attribute, such as all samples from vaccinated persons.

**Time attribute** The attribute that defines the timestamp of an element.

\( R_T = \{v_{T,1} \ldots v_{T,q}\} \)

\( E_T, i = \{e \mid e \in M \land e_T = v_{T,i}\} \)

\( v_{T,1} = 1 \), since the first time point is 1 and \( V_{t,1} = \{e_1, e_3, e_6\} \), since those elements have value 1 for the time attribute.

The attribute that defines the timestamp of a sample.

**Source attribute** The attribute that defines the source of an element.

\( R_P = \{v_{P,1} \ldots v_{P,q}\} \)

\( E_P, i = \{e \mid e \in M \land e_T = v_{P,i}\} \)

\( v_{P,1} = 1 \), since the first element has id 1 and \( E_P, 1 = \{e_1, e_2\} \), since those elements have value 1 for the source id.

The attribute that defines the person of a sample.

Since we only want one element per source per time stamp, we can assume the following always holds: \( |E_i \cap E_j| \leq 1 \), for all \( 1 \leq i \leq r \) and \( 1 \leq j \leq s \). In the HAVANA study this means that for each person there is at most one sample at each timestamp.

We also defined the deviation from the expected value of a combination as \( D_f \). More details about the computation of this deviation can be found in the Appendix A.

### 3.2 Tasks

In the survey by Alsallakh et al.\[2\] common tasks for Set-Typed data are defined. The tasks are divided into three groups:

1. Membership of the elements in the set. For example, which HPV-types does sample 1230 have?

2. Higher level reasoning about the sets without taking individual elements into account. For example, how many samples contain HPV-type 16?

3. Interrelation between set memberships and attributes. For example, are the samples that contain HPV-type 16 from persons that were vaccinated or not?

In the HAVANA study, the focus is on the tasks about the higher level reasoning and how the element memberships and attributes are interrelated. In addition we defined another group of tasks:

4. Development of the sets and combinations over time. For example, do samples from persons with preceding samples containing HPV-type 16, still contain HPV-type 16?

If we combine this new group of tasks with the already defined tasks in the survey, we can deduce the following tasks. For each task, we describe the general task, the information we want to see to perform the task, and how the tasks translates to the HAVANA study.
CHAPTER 3. PROBLEM STATEMENT

1. Tasks concerning higher level reasoning about the sets without taking individual elements into account:
   (a) **Determine the number of sets that are in the set family.**
       Show $|F_s|$.
       How many HPV-types are measured in the study?
   (b) **Analyze and compare set frequencies.**
       Show $|S_i|$ for every $i \in F_s$.
       How often does each HPV-type occur in the study?
   (c) **Analyze and compare combination frequencies.**
       Show $|C_I|$ for every $I \in F_C$.
       How often does each combination of HPV-types occur in the study?
   (d) **Analyze and compare the deviance of combination frequencies.**
       Show $D_I$ for every $I \in F_C$.
       What is the Pearson value for each of the combinations of HPV-types in the study?
   (e) **Identify the sets that constitute a certain combination.**
       Show every $i \in I$ for a combination $C_I$.
       Which HPV-types belong to a certain combination of HPV-types?
   (f) **Identify combinations which contain a specific set.**
       Show every $C_I$ where $i \in I$ for a set $i \in F_S$.
       Which combinations of HPV-types contain a certain HPV-type?

2. Tasks concerning interrelation between set memberships and attributes:
   (a) **Determine the distribution of an attribute in a certain set.**
       Show $|E_{A,j} \cap S_i|$ for every value $j \in Q_A$ for a set $i \in F_S$.
       How many samples of a certain HPV-type originate from vaccinated and non-vaccinated persons?
   (b) **Compare the attribute values between multiple sets.**
       Show $|E_{A,j} \cap S_i|$ for every value $j \in Q_A$ for every set $i \in F_S$.
       Which HPV-types have most the most samples from vaccinated persons?
   (c) **Compare the distribution of an attribute between multiple sets.**
       Show $|E_{A,j} \cap S_i|$ for every value $j \in Q_A$ for every set $i \in F_S$.
       For which HPV-types do more samples originate from vaccinated persons than from non-vaccinated persons?
   (d) **Find out the distribution of an attribute in a certain combination.**
       Show $|E_{A,j} \cap C_I|$ for every value $j \in Q_A$ for a combination $I \in F_C$.
       How many samples of a certain combination of HPV-types originate from vaccinated and non-vaccinated persons?
   (e) **Compare the attribute values between multiple combinations.**
       Show $|E_{A,j} \cap C_I|$ for a value $j \in Q_A$ for every combination $I \in F_C$.
       Which combination of HPV-types has many samples that originate from vaccinated persons?
   (f) **Compare the distribution of an attribute between multiple combinations.**
       Show $|E_{A,j} \cap C_I|$ for every value $j \in Q_A$ for every combination $I \in F_C$.
       For which combination of HPV-types do more samples originate from vaccinated persons than from non-vaccinated persons?

3. Tasks concerning development of the sets and combinations over time:
   (a) **Analyze the changing cardinality over time for every set.**
       Show $|E_{A,j} \cap C_I|$ for every value $j \in Q_T$ for every set $i \in F_S$.
       Does the frequency of HPV-type 51 remain the same over the course of the study?
(b) Analyze how often the elements preceding and following the elements in a set, also occur in that set.
   
   Show $e_T$ and $e \in S_i$ for every $e$ where there exists a $e_j \in S_i$ for which $e_P = e_{j,P}$ for a set $i \in F_S$.
   
   Do the samples preceding and following samples with HPV-type 6 also contain HPV-type 6?

(c) Analyze the values of an attribute in the elements preceding and following the elements in a set.
   
   Show $e_T$ and $e \in A$ for every $e$ where there exists a $e_j \in S_i$ for which $e_P = e_{j,P}$ for a set $i \in F_S$ and an attribute $A$.
   
   Does the vaccination status influence whether or not samples preceding and following samples with HPV-type 6 also contain HPV-type 6?

(d) Analyze the changing cardinality over time for every combination.
   
   Show $|E_{T,j} \cap C_i|$ for every value $j \in Q_T$ for every combination $I \in F_C$.
   
   Does the frequency of the combination of HPV-types 6, 66 remain the same over the course of the study?

(e) Analyze how often the elements preceding and following the elements in a combination, also occur in that combination.
   
   Show $e_T$ and $e \in C_I$ for every $e$ where there exists a $e_j \in C_I$ for which $e_P = e_{j,P}$ for a combination $I \in F_C$.
   
   Are HPV-types 6 and 66 found in samples preceding and following samples with HPV-types 6 and 66?

(f) Analyze the values of an attribute in the elements preceding and following the elements in a combination.
   
   Show $e_T$ and $e \in A$ for every $e$ where there exists a $e_j \in C_I$ for which $e_P = e_{j,P}$ for a combination $I \in F_C$ and an attribute $A$.
   
   How many persons are vaccinated in samples containing HPV-type 6 or HPV-type 66 preceding and following samples containing HPV-type 6 and 66?

### 3.3 Requirements

After discussion with several epidemiologists we formulated the desired properties of a possible solution. After which we suggested the following requirements.

**Visualizations**

V1 **Visible Tasks** Enable users to do the following tasks, without any interaction: all tasks concerning higher level reasoning about the sets without taking individual elements into account

V2 **Interaction Tasks** Enable users to do the following tasks: all tasks concerning interrelation between set memberships and attributes and all tasks concerning progression of the sets and combinations over time

**Usability**

U1 **Consistent** To make the system more usable, we require all interactions to behave similarly. This way users know what to expect before they are hovering over or clicking on a part of the visualization. When users know what to expect they do not need as much time to think about what is happening in the views.

U2 **Uncluttered** Every part of the visualization should have a clear purpose and every purpose should only be shown on the screen once.

U3 **Responsive** Actions on the tool should give a response within 300 milliseconds. When actions take too long, it might confuse the users. When the actions take more time, feedback should be given to the user. This way the users know the tool is still processing the action.
Technology Constraints

T1  **Software** The tool should work on default Windows 7 computers without any installation procedures. For the web-browser we recommend Google Chrome version 34 and above. Furthermore, Internet explorer version 11 and above and Mozilla Firefox version 45 and above could be supported.

T2  **Size of the data** The tool should at least be able to handle data sets up to 10,000 samples. The number of visible sets should be at least 30. The number of visible combinations should be at least 100. The number of different time-stamps that are supported should be at least 6.

T3  **Confidential** Since the data for the tool can contain confidential information, the data should not be accessible by outsiders.
Chapter 4

State of art

4.1 Related work

In this Chapter, the current visualization techniques for Set-Typed data are discussed. The relevant techniques are discussed in more detail. A comprehensive overview of all techniques can be found in The State-of-the-Art of Set Visualization by Alsallakh et al[2].

4.1.1 Euler-based

Venn and Euler diagrams  Venn and Euler diagrams are one of the oldest techniques that have been used for the visualization of set-typed data. Both diagrams are easy to understand and appealing to look at. Each set is displayed as an object, such as a circle. Combinations are displayed by overlapping objects. A Venn diagram is a diagram that shows all possible logical relations between a finite collection of different sets. Even when the combination of two sets contains no elements, that combination is still shown in the Venn diagram. In an Euler diagram only the non empty combinations are shown. Figure 4.1 shows a Venn and Euler diagram for four sets. In the Euler diagram in Figure 4.1a there is no area for the combination of only the blue and yellow circle. In the Venn diagram in Figure 4.1b there is an area for each combination of sets.

![Figure 4.1: Difference between a Euler diagram and a Venn diagram](image)

4.1.2 Node link diagrams

Figure 4.2 shows multiple node-link diagrams that are used for the visualization of Set-typed data applied to the data in the Euler diagram in Figure 4.2a.

Set and element nodes  Figure 4.2b shows a node-link diagram in which nodes are drawn for each set and element. If an element is in a set, a link is drawn between the element and the set. The element that is both in set A and set B, is drawn between the nodes of set A and B with a link to both nodes. This
technique is applied by Anchored maps [17], Eye Diagram [3] and PivotPaths [5]. The difference between the techniques is in the orientation of the nodes and the possible interactions.

Set nodes  It is also possible to draw nodes for sets only, as is shown in Figure 4.2c. In this diagram, links are drawn between sets when the intersection between those sets contains at least one element. Since one element is in both set A and set B, a link is drawn between the nodes of set A and B. This method is applied in the Ecology Network by Milns et al. [16] to analyze relations between different ecological objects. Milns et al. display both positive and negative relations between the ecological objects by changing the color or the links.

![Euler diagram with 6 elements.](image1)

![Method applied by Anchored maps[17].](image2)

![Method applied by Milns et al.[16]](image3)

![Method applied by Alsallakh et al. [1]](image4)

Figure 4.2: Different node-link based methods for the visualization of Set-Typed data.

Radial sets  Figure 4.2d shows a node-link diagram in which nodes are drawn for each set and for each combination. Since there are two combinations in this example, two nodes are drawn for the combinations, namely AC and AB. Furthermore, a node is drawn for each set. There are links between the combinations and the corresponding sets.

In Figure 4.3 Radial sets is applied to a movie dataset. Each set represents a collection of movies with the same genre. All sets are drawn in a circle. For each set, a degree distribution is shown in the outer circle. For example, in Figure 4.3 most of the movies with the genre adult, do not have other genres.

For each combination, a circle is drawn in one of the inner circles. The size of the combination represents the number of movies which have that combination of genres.

Each combination is sorted by size of the combination and the degree of the combination. The higher the degree of the combination, the closer to the center the combination is drawn. This makes sure that the combinations with a higher degree do not get lost within the usually larger combinations with degree two. The color of each circle represents the median release year of those movies, but it is also possible to select another attribute such as the average rating of the movies in the combination.

Since there are many combinations and each combination has at least two links, Alsallakh et al. decided to remove all links to reduce the visual clutter. Users can hover on combinations to view which sets are in that combination. Radial sets combines the node-link diagram with other views such as the histogram of each set, the histogram of the degree of the elements and the histogram of each attribute. Interaction on those histograms allows the user to formulate complex queries. For example, by selecting all movies with three genres and a rating of four and higher.
4.1.3 Aggregation based methods

Aggregation based methods focus on sets and combinations instead of individual elements. In Figure 4.4, we applied three different aggregation methods to the Euler diagram in Figure 4.4a. The three different aggregation methods are explained in the following paragraphs.

Sets 'o' gram  Figure 4.4b shows the method Freiler et al.[6] apply in their tool Sets 'o' gram. Freiler et al. draw a bar for each set, where the height of the bar represents the number of elements in that set. Inside each bar, all combinations in which that set occurs are drawn. Since one element in set A also occurs in set B, Freiler et al. draw a bar with height one inside the bar of set A. Sets 'o' gram improves this visualization technique by supporting interactions on the bars, such as selections and brushing to further investigate the sets and their combinations.
CHAPTER 4. STATE OF ART

Double-Decker plots  Figure 4.4c shows the technique that is applied by Hoffman et al.[8] in their Double-Decker plots tool. For each set, the data is divided into two parts, namely the elements that are in this set and the elements that are not in this set. The lowest layer is divided in two parts between the elements that are in set A and the elements that are not in set A. On the next layer, both areas of element in set A and elements not in set A are divided into two parts, namely the elements in set B and the elements that are not in set B. This way one can generate bars for each combinations and sets and directly compare them.

Furthermore, Double-Decker plots fills each bar with a boolean attribute as shown in Figure 4.5a. Information about the persons that were on the Titanic is shown. On the first layer, the persons are divided between adults and children. The second layer divides both parts in a group of male and female persons. Finally, on the third layer, each group is divided between the ticket they bought, for example first class or crew member. The figure shows that the number of persons that were adult male crew members is larger than the number of persons that were adult female crew members. The red parts of each bar display the survivors of that group of persons. Female crew members had a much higher surviving rate than the male crew members. Double-Decker plots provides interaction with the bars for more information about each combination.

![Double-Decker plot of set typed data by [8]](image1)

![Parallel Sets visualization of set typed data by [12]](image2)

Figure 4.5: Double-Decker plots and Parallel sets

Parallel Sets  Another aggregation based method is shown in Figure 4.4d, which represents the technique used by Kosara et al.[12] in Parallel Sets. Parallel Sets also divides each set in multiple parts in the same way as Double-Decker plots. However, instead of drawing them on top of each other, they leave some space between the layers. Parallel Sets draws links between each layer, depending on their set memberships. Since three elements in set A are not in set B, the link between A and not B is three elements wide. There is one element in both set A and set B, so the link between A and B is one element wide. Between the second and the third layer, the links between the first and second layer are continued. The link between elements in set A and not in set B, is divided between the elements in set C and the elements that are not in set C. Since there is one element that occurs in set A and set C, there is a link of width one coming from the link of elements in set A and not in set B. There is also a link of width two, going the elements not in C, since there are two elements that are only in set A.

Figure 4.5b shows an example of Parallel Sets. In this example, the same data is shown as in the Double-Decker plot in Figure 4.5a. In this example of Parallel Sets, survivors are shown on the third layer. The second layer shows which parts of the survivors were male and which were male. The first layer represents the class in which the persons on board were traveling. Since most of the crew members were male, there goes a wide link to the male part of the second layer and only a small link to the female part of the second layer.

4.1.4 Matrix

There are different ways to utilize matrices for set visualization. We show several methods that are applied in literature in Figure 4.6 and explain them in the following paragraphs. All examples in Figure 4.6 are applied to the data in the Euler diagram in Figure 4.6a.
CHAPTER 4. STATE OF ART

(a) Euler diagram with 6 elements.

(b) Method applied by J. Huo [9]

(c) Method applied by Sadana et al [22]

(d) Co-occurrence matrix used by Yalcin et al. [26]

Figure 4.6: Different methods used to utilize matrices for set visualization

**KMVQL (Boolean Query Specification and Query Result Visualization)** KMVQL is a tool that can be used to visualize sets and is created by J. Huo. [9]. Figure 4.6b shows the technique they use to show information about the sets and combinations. They divide each set in the same way as the aggregated methods; all elements in a set and all element that are not in that set. Instead of placing the layers only on the horizontal axis, the layers are placed on both the vertical and horizontal axis. For example, the upper-left corner represents the elements that are in both set A and set C, but not in set B. Each cell can be used to display information about that combination, such as the number of elements or a scatterplot of the elements.

**OnSet** OnSet by Sadana et al [22] applies the method displayed in Figure 4.6c. Each set is represented as a matrix. Inside the matrix, each cell represents an element. When the element is present, the cell is filled. Since the first four elements are in set A, the first four cells are filled with a dot. The third element is also in set C, so the third cell is also filled in the matrix of set C.

As shown in Figure 4.7, users can also view intersections by dragging the set matrices on top of each other. This way users can find which elements are in all of the selected sets. Furthermore a similarity measure is defined. Sets that share a lot of elements are similar to each other. This enables users to display the similarity with links between each set, where the thickness represents the similarity of the two sets.

Figure 4.7: An example of displaying intersections between sets in OnSet by Sadana et al. [22]
AggreSet  AggreSet uses a set matrix as shown in Figure 4.6d. Each cell represents a combination of two sets, namely the set of the column and the set of the row. The color represents the frequency of the combination. For instance, since there is only one element that is in both set A and set B, that cell is only slightly gray. Since the combination of set B and set C is empty, that cell is still white.

Figure 4.8 shows AggreSet applied to a sample data set that looks like the HAVANA study data. Furthermore, AggreSet always shows a histogram for both the degree of the elements and the frequency of the sets. AggreSet supports linking and brushing on those histograms together with the set matrix. Users can interactively select or compare different attribute values. All actions are user-friendly and easy to learn. This enables users to quickly formulate more complex queries on the data set.

4.1.5  Linear diagram

Linear diagrams represent sets using straight line segments, with line overlaps corresponding to set intersections. Each column represents a combination. A example of such a linear diagram is shown in Figure 4.9.
Kim et al. [10] developed ConSet which display the combinations that are present in the data set with a linear diagram. Each set is represented by a row, as is shown in Figure 4.9. Each column represents a combination of sets. The cells of the sets that are in the combination are colored. Above each combination, ConSet displays information about the combination:

- Elements: the elements that occur in the combination;
- Pie chart: a colored pie chart to display the sets in the combination; and
- Bar chart: the frequency of the combination.

An example of ConSet applied to a small part of the sample HPV-type data, is shown in Figure 4.10a.

![ConSet tool by Kim et al. [10](a) and the UpSet tool by Lex et al. [13](b)](image)

Figure 4.10: ConSet and UpSet, applying the linear diagram to visualize Set-typed data.

Lex et al. [13] applied to data from the HAVANA study. Individual sets are columns and the rows represent combinations. Behind each combination UpSet shows information about those combinations, such as the frequency and deviance. UpSet also allows users to show information of other attributes, such as the averages of numerical attributes. Furthermore, users can select elements in the linear diagram of UpSet and display those elements in other visualizations, such as scatterplots. UpSet provides multiple options to sort the combinations in the linear diagram, such as by degree, frequency or other averages of numerical attributes and they allows users to select parts of the elements to formulate more complex queries.

### 4.2 Comparison

We applied most of the techniques on the HAVANA data set. This data set consists of 1600 elements with 27 different sets. We implemented multiple techniques to further discuss the advantages and disadvantages of those techniques with the users. For the implementation of the Euler diagram, Sankey diagram, chord diagram and force-directed layout we modified d3.js examples to display the data set. We implemented the Linear diagram, matrix, and bubble diagram ourselves. The chord diagram, bubble diagram and force-directed layout are variations on the node-link diagrams and are discussed in the following paragraphs.

**Euler diagram** The Euler diagrams are not scalable in the number of sets and in the number of combinations. Figure 4.11 shows 20 overlapping sets. The area of the disk represents the number of elements in a set. When using one circle for each set it is not possible to draw all combinations, because
CHAPTER 4. STATE OF ART

there is not enough space for those combinations at the disks. For example, the six sets that are shown without overlap all appear in at least one combination. Furthermore, it is really hard to compare the frequency of the combinations. In the figure, it is unclear if the combination of type 4 and 18 is greater or smaller than the combination of type 11 and 16.

![Figure 4.11: Visualization of a Euler diagram for a sample data set with 20 different sets.](image)

Chord diagram Figure 4.12a shows a chord diagram applied to the sample data set. The idea of the chord diagram is based on the Ecology Network. All sets are drawn on the outer circle. The length of the arcs represents the frequency of the corresponding HPV-types. Links are drawn between two HPV-types when the frequency of the combination between the two is larger than four. The width of the link represents the frequency of the combination. Even with interaction on the diagram it is hard to determine where each link is going. Other problems are: it is not possible to show combinations of three or more sets and there are not many options to integrate attributes.

![Chord diagram](image)

Force Directed Lay-Out Figure 4.12b shows a node-link diagram where each node represents a set. The size of each node represents the number of elements in each set. When an element is in two sets, a link is drawn between the two nodes of those sets. We apply a threshold on the size of the combinations, otherwise the node-link diagram would become too cluttered. Thus instead of drawing links when there is one element in two sets, we only draw the links when there are at least six elements that are in both sets. Through interaction with the node-link diagram, users can analyze individual sets as shown in Figure 4.12b.

![Node-link diagram](image)
**Radial Sets** Another variant of a node-link diagram is shown in 4.13a. The visualization technique that is used by Radial Sets is applied. The node of each set is placed on the outer circle. Each combination is shown in the middle of the circle. Links are drawn between the combinations and the corresponding sets. We used the color of the combination to represent the number of sets that constitute the combination. Furthermore, hovering is added to make it easier to investigate the sets and combinations. Figure 4.13b shows the combinations with at least six elements, while hovering over the set that corresponds to HPV-type 16.

![Radial Set Technique](image1.png)

(a) All different combinations in the HAVANA study
(b) All combinations that occur at least six times. All combinations with HPV-type 16 are highlighted

Figure 4.13: Radial set technique applied to the HAVANA study

**Matrix** Considering the tasks, we expected that the co-occurrence matrix would be the most useful matrix representation. Figure 4.14a shows the co-occurrence matrix applied to the HAVANA study data. Each cell represents the combination of two types. The color determines the number of samples in each combination. The matrix is easy to understand, it is easy to find combinations in the matrix and the matrix visualizes all combinations with two HPV-types clearly. However it is not possible to show combinations that consists of more than two HPV-types.

![Co-occurrence Matrix](image2.png)

(a) A co-occurrence matrix
(b) A linear diagram

Figure 4.14: Two visualization techniques that both display the HAVANA study data
Linear Diagram A variant of the linear diagram is shown in Figure 4.14b. When creating the linear diagram a threshold was needed, otherwise there would be too many combinations. If there are too many combinations the width of the column becomes too small to be identifiable. The advantage of this technique is that users can easily compare the frequency of each combination. Users can quickly identify which combination that occurs the most frequently. By sorting the sets on the size, it is possible to see which relatively small sets occur in relatively large combinations.

Sankey diagram We decided to look at the Parallel Sets method to visualize the development of a combination over time. This is technique is similar to a Sankey diagram. Figure 4.15 shows the persons in the study that once had both HPV-type 39 and 56. Each layer is now a different timestamp instead of a division between samples that did or did not have a certain HPV-type. For each layer we show if the samples had HPV-type 39 and 56. If not, we showed if the samples contained either 39 or 56 or none of those. Finally we draw links between the layers which connects all samples from the persons. So the person who had HPV-types 39 and 56 at the first timestamp, still had HPV-type 39 at the second timestamp. However, the diagram does not show the further history of this person as only relations between two consecutive timesteps are shown.

Discussion Together with several epidemiologists, we looked at the visualizations of the HAVANA data set. After an explanation of each technique, we concluded that the force direct layout, chord diagram and Euler diagram were not what they were looking for, mainly because it is difficult to find the combinations of two specific sets. Furthermore, the Euler diagram is not scalable enough to study combinations of two HPV-types in the HAVANA study. The chord diagram and force directed diagram were performing less than the matrix in comparing combinations and identifying the sets that belong to a combination. The Sankey diagram was more difficult to understand for most of the epidemiologists, but they thought it would be useful to be able to investigate the development of HPV-types over time.

In the second part of the brainstorm session, we compared the bubble sets, matrix and linear diagram. The bubble sets did visualize the combinations, but it is hard to identify the sets of each combination. It is possible to find those sets with interaction on the bubbles, but it is not clear from the static visualization.

The matrix was interesting because it is easy to find and compare all combinations of two different HPV-types. The linear diagram was interesting, because it allows users to directly compare the frequency of the combinations and combinations of more than two HPV-types can be studied.

While the linear diagram is able to show combinations of more than two types, the columns become too small to display all combinations. We can apply a threshold to the combinations, but this also means that not all combinations are shown. Another problem is that it is harder to identify the sets in a combination compared to the matrix. In the matrix, only the combinations with two HPV-types are shown.

In the displayed linear diagram and matrix no attributes were integrated yet. However, AggreSet and UpSet show that there are possibilities to integrate attributes in the linear diagram and matrix.
the linear diagram, UpSet showed the average of numerical attributes after each combination. Furthermore, UpSet allows the user to display the elements of combinations on scatter plots to analyze the relations between those elements and the selected attributes. In the matrix of AggreSet, users are enabled to interactively display attributes in the matrix. When hovering over the value of an attribute, it shows how many elements of each combination have the same value for that attribute.

Both the matrix and linear diagram do not offer possibilities to integrate the development of sets and combinations over time. The Sankey diagram can show the progression of selected combinations, but cannot give an overview of larger number of combinations. Based on this evaluation, we decided to use both the matrix and the linear diagram to show combinations, and a Sankey diagram to show temporal trends. In the following chapter we elaborate on this.
Chapter 5

Design

In this chapter we discuss the design of the tool. We discuss the views we created to provide the solutions to the tasks in Section 5.1, and the techniques we used to improve the usability of the tool in Section 5.2.

We present ComboVis, the name is based on the visualization of combinations. An overview of ComboVis is presented in Figure 5.1. In panel A and B, histograms of the attributes and sets are shown. Those visualizations are placed together as they all represent the number of elements with the length of a bar. In panel C and D, the frequency of the combinations is shown in a linear diagram and in a matrix. Below the set distribution, matrix and linear diagram, a Sankey diagram (panel F), and a table (panel E) are shown. Both the table and the Sankey diagram contain information about the persons in the selection instead of the samples in the selection. The table is added to ComboVis, to provide users with additional information of the persons that are currently selected.

Figure 5.1: Overview of the tool
5.1 Tasks

In this section we describe how users can use ComboVis to perform the tasks described in Section 3.2.

5.1.1 Higher level reasoning about the sets

Set distribution For the first two tasks we created a set distribution as shown in Figure 5.2. The number of samples for each HPV-type is shown. Figure 5.2a shows the answer to the first task: How many sets are in the set family. The precise number is not shown, but in most cases an estimate is good enough. Figure 5.2b shows the set distribution sorted on the number of samples of each HPV-type. This is useful to answer questions related to the second task: Analyze and compare set frequencies. For instance, to determine the HPV-type is most frequently occurring in the samples. Figure 5.2b shows $|S_i|$ for every $i \in F_S$ in a way that users can quickly compare set distributions.

Matrix We created a matrix as displayed in Figure 5.3. As discussed in the previous chapter, each cell represents the frequency of a combination of two sets. In the matrix users can perform the following tasks:

- Analyze and compare combination frequencies;
- Analyze and compare the deviance of combination frequencies;
- Identify the sets that constitute a certain combination; and
- Identify combinations that contain a specific set.

For the task analyze and compare combination frequencies, we need to show $|C_I|$ for every $I \in F_C$. We display $|C_I|$ on the color scale for every $I \in F_C \cap |I| = 2$. Figure 5.3a shows the matrix with the frequency of each combination.

For the task analyze and compare the deviance of combination frequencies, we need to show $D_I$ for every $I \in F_C$. We display $D_I$ on the color scale for every $I \in F_C \cap |I| = 2$. Figure 5.3b shows the matrix with the frequency of each combination.

User can switch between the frequency and the deviance, by a button above the matrix. In the matrix it is possible to identify sets that constitute a certain combination. Users can look at the row- and column names of the cell to identify sets that constitute a certain combination.

To identify combinations that contain a specific set users can follow the row and column of the sets to find all combinations of that set.
Linear diagram We created a linear diagram as displayed in Figure 5.4. As discussed in the previous chapter, each column represents a combination of different HPV-types. The bar above the columns represents either the frequency or the deviance of that combination. The color of the circles and the background of the column represent the degree of the combination.

The linear diagram allows users to perform the same tasks as they can on the matrix. For the task analyze and compare combination frequencies, we need to show $|C_I|$ for every $I \in F_C$. We display $|C_I|$ on the bars above each column and show every combination above a user-defined threshold in Figure 5.4a. We discuss the limitations on the number of combinations in Subsection 7.1.

For the task analyze and compare the deviance of combination frequencies, we need to show $D_I$ for every $I \in F_C$. Users can switch the height of the bars from frequency to the deviance by the button above the linear diagram. After switching the button, the linear diagram displays $D_I$ for each combination above the threshold. Figure 5.4b shows the linear diagram with the deviance of each combination. In the linear diagram it is possible to identify the sets that constitute a certain combination. Users can lookup the row names of the circles in the column to identify the sets that constitute to a certain combination. To identify combination that contain a specific set users can follow the row of the circles to find all combinations including that set.
5.1.2 Interrelation between set memberships and attributes

We show the attribute distributions for several interesting attributes, as defined by domain experts. Figure 5.5 shows the vaccination and cohort distribution of the samples in the HAVANA data set. However, to perform the tasks those attribute distributions are not sufficient.

![Vaccination distribution](image1)

(a) Vaccination distribution

![Cohort distribution](image2)

(b) Cohort distribution

Figure 5.5: Attribute distributions of the HAVANA data set.

We therefore allow users to brush the elements when they move their mouse over a bar of the attribute distribution. The bars in the set distribution are colored according to the number of elements that were brushed. For example, when hovering of the NO value for vaccination status, the set distribution shows $|E_{\text{VAC}, \text{No}} \cap S_i|$ for every $i \in F_S$. The result is shown in Figure 5.6a and allows users to compare the attribute values between multiple sets. Figure 5.6a shows that HPV-type 45 only occurs in non-vaccinated persons. Furthermore, it can be seen that many HPV-types that occur mainly in non-vaccinated persons, such as HPV-type 16, 18, 31 and 68.

![Observed distribution](image3)

(a) The occurrence of samples from non-vaccinated persons in each set

![Observed distribution](image4)

(b) The attribute distribution for each set

Figure 5.6: The set distributions of the HAVANA data set with information about the vaccination attribute.

It is also possible for users to determine the distribution of an attribute in a certain set and compare the distribution of an attribute between multiple sets. By hovering over the button with the two blue bars in Figure 5.5, the distribution for that attribute is shown in each set as depicted in Figure 5.6b. This enables users to compare the values of an attribute within each set and to compare the attribute distributions between sets. Users can see that HPV-type 59 occurs more often in samples from vaccinated persons versus samples from non-vaccinated persons. Furthermore, users can see that the distribution for HPV-type 18 and HPV-type 31 are almost similar, and that the difference between samples from vaccinated versus samples from non-vaccinated persons is bigger in HPV-type 51 than in HPV-type 68.
Hovering over an attribute displays the occurrence of the attribute value is also shown in the linear diagram and matrix in the same way as it displays the occurrence of the attribute value in the set distribution. For example, when hovering over the NO value for the vaccination status the matrix and linear diagram show \(|E_{\text{vac, no}} \cap C_I|\) for every \(I \in F_c\). The result is shown in Figure 5.7.

This information is needed for the task compare attribute values between multiple combinations. The figures in Figure 5.7 show that multiple combinations are present that only occur in samples from non-vaccinated persons, such as the combination of HPV-types 16 and 18, the combination of HPV-types 16 and 66 and the combination of HPV-types 16 and 68.

Users can also determine the distribution of an attribute in a certain combination and compare the distribution of an attribute between multiple combinations. This is done by the same action of showing the attribute distribution in each set. As a result of this action the attribute distribution is shown in both the matrix and the linear diagram. For example, the vaccination distribution is shown in the matrix and linear diagram in Figure 5.8. This enables users to compare the attribute distributions of different combinations. In the example, the combination of HPV-type 11,56 and the combination of HPV-type 31,39 have the same distribution of samples from vaccinated persons versus non-vaccinated persons. It also shows that in the combination of HPV-type 6,66 more samples were from vaccinated persons than from non-vaccinated persons.

Figure 5.7: The occurrence of samples from non-vaccinated persons in each combination

Figure 5.8: The attribute distribution for each combination
CHAPTER 5. DESIGN

5.1.3 Development of the sets and combinations over time

For the following two tasks, we use the option of splitting by attribute values to show the time attribute distribution in each set and combination:

- Analyze the changing cardinality over time for every set.
- Analyze the changing cardinality over time for every combination.

In Figure 5.9 the changing frequency over time in the sets and combinations is shown. ComboVis shows $|E_{T,j} \cap S_i|$ for every $j \in Q_T$ for every $i \in F_S$ and $|E_{T,j} \cap C_I|$ for every $j \in Q_T$ for every $I \in F_C$. Since the persons in the study are followed over time and thus getting older and subsequently more sexually active, most sets and combinations increase in frequency. However, in some sets a decrease in frequency is shown in the last year, such as for HPV-type 6, 68, and 74. One can also observe a peak in the fourth year for HPV-type 11. Like the single HPV-types, the combinations of HPV-types show similar trends. Some combination appear later in time than the single HPV-type, because combinations are less likely to be found than individual HPV-types. It is interesting to see that in some years there is a gap, such as in the combination between HPV-types 6, 51 and HPV-types 31, 51.

![Figure 5.9: Changing cardinality over time for sets and combinations](image)

ComboVis shows a Sankey diagram for the following tasks:

- Analyze how often the elements preceding and following the elements in a set, also occur in that set.
- Analyze the values of an attribute in the elements preceding and following the elements in a set.
- Analyze how often the elements preceding and following the elements in a combination, also occur in a set of the combination.
- Analyze the values of an attribute in the elements preceding and following the elements in a combination.

For the first task: **analyze how often the elements preceding and following the elements in a set, also occur in that set**, the user starts with selecting a set. The Sankey diagram in 5.10 is then shown. In the example shown in Figure 5.10, the selected type is HPV-type 56. The Sankey diagram shows all samples from the persons that had a sample with HPV-type 56 at any point in time.
CHAPTER 5. DESIGN

Figure 5.10: The Sankey diagram that showing all persons that had HPV-type 56 at any point in time.

The samples from each timestamp are divided in four groups:

- The persons that did not send in a sample;
- The persons that did not have HPV-type 56;
- The persons that had only HPV-type 56;
- The persons that had HPV-type 56 and at least one other HPV-type.

It is possible to further investigate a part of the Sankey diagram as shown in Figure 5.11. Here the persons that had HPV-type 56 and at least one other HPV-type at the first timestamp were selected.

Figure 5.11: The Sankey diagram drilled down to the persons that had HPV-type 56 plus another HPV-type at timestamp 1.

For the second task: analyze the values of an attribute in the elements preceding and following the elements in a set, users can select the set they want to analyze. Next, they can hover over the attribute value they are interested in to show how the attribute values change over time in the Sankey diagram. For example, after selecting the persons that had HPV-type 56 and at least one other HPV-type at the first timestamp and hovering over the non-vaccinated attribute value it can be seen in Figure 5.12 that only non-vaccinated persons still had HPV-type 56 at timestamp 2.

Figure 5.12: Sankey diagram, drilled down to the persons that had HPV-type 56 plus another HPV-type at timestamp 1, colored by non-vaccination status.

The Sankey diagram can also be used for the tasks on the development of the combinations: Analyze how often the elements preceding and following the elements in a combination, also occur in that combination, and show \( e_T \) and \( e_S \) for every \( e \) where there exists a \( e_j \in C_f \) for which \( e_p = e_{j,p} \) for a combination \( I \in F_C \) and an attribute \( A \). An example is shown in Figure 5.13.
The samples from each timestamp are divided into four groups:

- The persons that did not send in a sample;
- The persons that did not have HPV-types 6 and 51;
- The persons that had only HPV-types 6 and 51;
- The persons that had HPV-types 6 and 51 and at least one other HPV-type.

This way the Sankey diagram always stays consistent for each selection. The attribute hover works the same way as for the Sankey diagram of a single set, and therefore both tasks can be solved with the Sankey diagram.

![Sankey diagram example](image)

**Figure 5.13:** The persons that had both HPV-types 6 and 51 in a sample at any point of time during the HAVANA study

### 5.2 Visualization Techniques

#### 5.2.1 Brushing and linking

The idea of brushing and linking is to highlight the same data items in different views of a visualization. The set distribution, linear diagram, matrix and Sankey diagram all highlight the same data items after hovering over a set, combination or attribute value. For example, when a user hovers over HPV-type 66, the other visualizations show how many of the samples with HPV-type 66 are in each part of the tool. Combinations with HPV-type 66 logically show that all elements in that combination contain HPV-type 66. It is also possible that other combinations contain samples with HPV-type 66. For instance, the combination of HPV-types 51, 52 contains a sample that contains HPV-types 51, 52 and 66. The result is shown in the linear diagram in Figure 5.14. In the set distribution it is shown how many of the samples that contain HPV-type 6, also contain HPV-type 66.

![Brushing and linking example](image)

**Figure 5.14:** Brushing and linking after hovering over HPV-type 66
CHAPTER 5. DESIGN

Beside brushing a single set, it is also possible to brush a combination. When a user hovers over the combination of HPV-types 6 and 51 all parts of the tool that include samples with HPV-types 6, 51, are highlighted, as is shown in Figure 5.15. For example, in the set distribution it is shown how many of the samples that contain HPV-type 45, also contain HPV-types 6 and 51. In the linear diagram and matrix, the combinations in which some samples also contain HPV-types 6 and 51 are shown. For example, the linear diagram and matrix show how many of the samples that contain HPV-types 6, 31 also contain HPV-type 51.

![Figure 5.15: Brushing and linking when hovering over the combination of HPV-types 6 and 51.](image)

5.2.2 Selection

It is possible to select sets, combinations and attributes. This enables users to formulate more complex queries. For example, in Figure 5.16 the set distribution, linear diagram and matrix are shown after the selection of HPV-type 66.

![Figure 5.16: View of the set distribution, linear diagram and matrix after selection of HPV-type 66.](image)

The attribute distributions also change after each selection. This way users can see how many of the samples with HPV-type 66 were from vaccinated persons. This is applied to all attribute distributions. Besides the selection of attributes, sets and combinations it is also possible to select the combinations of a certain degree. This way users can investigate larger combinations more easily, since those are often overwhelmed by the combinations of degree two. Figure 5.17 shows all combinations with degree three.
CHAPTER 5. DESIGN

5.2.3 Details

Not all parts of the visualization show precise numbers. In the matrix the frequency of each combination is displayed by a color, which is not accurate, and in the linear diagram the exact height of a bar is might not always clear. We decided to show the exact numbers when hovering over a part of the visualization. We always show this number at the upper left corner, so that users know where to look. Another option would be to display this as a tool-tip above the hovered part, but in the matrix and Sankey diagram the tooltip would obscure other parts of the view. Figure 5.18a shows the number of samples from persons in cohort 1 in the selection. Figure 5.18b shows the number of samples that contain HPV-types 45, 51 and 68.

Figure 5.18: More detailed information about a part of the tool.

Sometimes users need more information on the expected value and the ratio between the observed and expected value. Therefore we show the observed, expected, and the boundaries of the 95% confidence interval in a separate plot in the upper part of the matrix. For example, the plot in Figure 5.19a is shown for the combination of HPV-types 16 and 39. The dot on the plot is the observed frequency of the combination of HPV-types 16 and 39. The interval represents the 95% confidence interval. If the dot (observed value) falls within in the 95% confidence interval, the combination is not occurring significantly more or less than expected. When the distribution of an attribute is shown in the combination, the plot is shown for all attribute values when hovering over a combination, as shown in Figure 5.19b. In this figure the plots of both the vaccinated and non-vaccinated persons are shown, so that users can compare the two. More information about the exact statistics used can be found in Appendix A.

Figure 5.19: More detailed statistical information about a combination.
Chapter 6

Evaluation

6.1 Structure of the evaluations

We decided to split the evaluation into two rounds. This way comments from the first round can be used to improve the tool. Furthermore, questions from users might shift the attention to other tasks which we originally thought were less important. For the evaluations, we gave no introduction of the tool. This way, we could evaluate which parts of the tool are immediately clear and thus need to be improved or require an explanation. In order to investigate if the users understood the visualizations, we organized individual evaluation sessions. Each evaluation session took about 40 minutes.

The structure of both evaluations was the same; each session consisted of two parts. First, participants were asked to execute multiple tasks. We gave tasks in increasing difficulty. We observed how the participants tried to solve those tasks. If the participants used a less efficient option to solve a task, we showed how to execute that task more efficiently and noted any comments from the participants. Furthermore, we wrote down all questions they asked to fulfill the tasks, so we could evaluate which parts of the tool were not clear enough.

After the participants solved all tasks, we asked questions about different parts of the tool. For example, what they liked/disliked about the tool, whether they were still missing something and if they would like to use the tool with their own data. The questions of both evaluations can be found in Appendix B.

6.2 Early user evaluation

The participants in the early user evaluation were four epidemiologists, two male and two female. All participants already participated in a brainstorm session in the early design phase. In the early user evaluation, we decided to only test a part of the tool. We left out the Sankey diagram in this version. Since there were some performance issues at the time of the early user evaluation, we decided to use only 800 persons from the HAVANA study. This also made sure that participants did not know the answers to the questions before using the tool, as some of them were familiar with the data in the HAVANA study.

6.2.1 Observations

In this section, we describe observations concerning questions that raised some issues.

Question 3: *How many samples have exactly two types?*

In the visualization shown in Figure 6.1, the number of samples per degree are shown. The correct answer to question 3 would be 137 as is shown after degree 2 in the degree distribution. However three out of four participants had trouble finding this number, because they did not know what the word *degree* meant in the context of the HAVANA study.
CHAPTER 6. EVALUATION

Question 5: How many samples contain types 45, 51 and 68?
Most participants had troubles answering this question. Some participants selected the samples with degree 3 in the degree distribution. However, since there can also be samples that contain those three types an other types, those samples with types 45, 51, 68 and other types are excluded after the selection of all samples with degree 3. After giving a hint that the background of the linear diagram represented the number of types in a combination, the participants found the combination in the linear diagram as shown in Figure 6.2

![Figure 6.1: Question 3](image)

![Figure 6.2: Question 5](image)

Question 13: How many samples contain type 51 or 52 or both?
The question was not completely clear to the participants. Most of the participants added the number of samples of 51 to the number of samples of 52. Table 6.1 shows 3 samples that contain type 51 or 52 or both. However, type 51 and type 52 both occur two times in Table 6.1. Thus simply adding the number of samples with type 51 to the number of samples with type 52 does not give the correct answer. Participants could select all samples that contain 51 and all samples that contain 52 after which the number of selected samples is shown in the top right of the visualization.

<table>
<thead>
<tr>
<th>sample</th>
<th>HPV-Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>3</td>
<td>51,52</td>
</tr>
</tbody>
</table>

![Table 6.1: There are three samples that contain type 51 or 52 or both in this example.](image)
CHAPTER 6. EVALUATION

Question 14: Which types occurs more frequently in samples from vaccinated persons compared to samples from non-vaccinated persons?
There is an option that shows the distribution of the Vaccination in the types and combinations. However, not all users understood that the blue button was something they could click on. The button is shown highlighted in Figure 6.3a and the result of this button is shown in 6.3b. Without this option, the question was much more complicated for the participants. After telling the participants that the button existed, the participants could answer the question without any further trouble.

![Vaccination button and visualization](image)

(a) The button is next to the question mark
(b) Vaccination distribution in the visualizations

Figure 6.3: After hovering over the button, the distributions to answer question 14 are shown in the views

Question 16: How many combinations of three types occur at least 6 times?
When clicking on the samples with three different types, they expected to see only the combinations of three types. However suppose we have the samples in Table 6.2, then also the third sample would be filtered out. The idea of the question was to count the number of combinations with degree three in the linear diagram, which most of the users understood after a more detailed description of the question and pointing at one combination of three types in the linear diagram.

<table>
<thead>
<tr>
<th>sample</th>
<th>HPV-Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>m₁</td>
<td>51,52,66</td>
</tr>
<tr>
<td>m₂</td>
<td>51,53,66</td>
</tr>
<tr>
<td>m₃</td>
<td>51,52,53,66</td>
</tr>
</tbody>
</table>

Table 6.2: Filtering samples with degree 3, would filter out m₃

Question 18: In which year is the Pearson value the highest for the combination of types 6 and 52?
We did not expect the users to answer this question without any guidance, since it takes multiple steps to answer this question. First, the user had to select the enable the distribution of the time attribute. After that they could hover over the combination 6, 52 to find the Pearson values for each timestamp in the plot that shows up as shown in Figure 6.4. A common action that participants performed was to select the combination 6,52. However, when a combination or set is selected the Pearson values also change. The Pearson value depends on the selected samples and the frequency of the sets in the selected samples. If a combination is selected, the expected and observed in the selected samples are always the same for that combination.
CHAPTER 6. EVALUATION

Unclear Table 6.3 shows how well each view and each action on each view of the tool was understood by the participants. Most participants perceived the actions on both combination views as unclear. We do not use exclusive memberships to combinations, which means that the elements in a combination of two types can also occur in another combination of two types. After an explanation of what happens when the user hovers over a combination, they understood the hovers and clicks better.

<table>
<thead>
<tr>
<th>View</th>
<th>Action</th>
<th>- -</th>
<th>-</th>
<th>0</th>
<th>+</th>
<th>++</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix</td>
<td>View</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hover</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Click</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear diagram</td>
<td>View</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hover</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Click</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set distribution</td>
<td>View</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hover</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Click</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribute distribution</td>
<td>View</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hover</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Click</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3: Understandability of the users for the different views and actions on the views.

The selection and brushing on the sets and attributes distributions were needed for some of the tasks, this might have increased the understandability of those actions. Those actions were not needed for the combinations, which might be a reason for the lower understandability of those actions. Adding a tasks in which the users need to brush or select a combination in the linear diagram or matrix, might increase the understandability of those actions. For example, Which type occurs most often together with type 6 and 52? can only be solved by brushing or selection of the combination of types 6 and 52.

The linear diagram was described as neutral because the background was not immediately clear. For one participant it was also not directly clear that the bars above each column represented the frequency of the combinations.

Furthermore, one participant described the click on the set distribution as unclear, because after hovering over a set, multiple bars show up in the linear diagram. After the selection of a set, the frequency of some of the combinations is below the user-defined threshold and thus are no longer visible in the linear diagram and matrix.

Sorting and Size The option to select a different order of the set or combinations was not used very often by the users. One user said: I would not have thought about sorting the types on frequency, so I did not look for such an option. Another reason that selection of a different order of the sets or combinations was could be that the those actions were not needed to solve the tasks. The change of the size from

---

Figure 6.4: The Pearson value for each year for the combination of HPV-types 6 and 52.
frequency to Pearson was used more often as it was needed to answer some of the questions. Not all users could find the button the first time, but after that, users did not have any trouble with the changing of size from frequency to the Pearson value.

**Missing** Most participants could not think of something that they were missing. However, one participant pointed at the distribution of the vaccinated and non-vaccinated people in one combination and said it would be nice to see how that distribution changes over time. If the difference between samples from vaccinated and non-vaccinated in a combination becomes larger over time, the vaccination may start working later; and if the difference between samples from vaccinated and non-vaccinated in a combination becomes smaller over time, the effect of the vaccination might be shorter than expected.

**Like** All participants were enthusiastic. Their answers mostly contained two or three of the following:

- A lot of data on a small screen in a understandable way
- Works much better than exploring an Excel file
- I like the interactive way to investigate parts of the data I’m interested in
- I really like the colors

**Dislike** Two participants could not think of something they disliked about the tool. One of the participants said: *Maybe that it took some time to get used to the tool.* Another one said *Some notations such as observed, Pearson and degree are not common words for me.*

**Own data** We asked if the participants would like to use the tool with their own data. All reactions were positive, however not all participants were working on the HAVANA study. Reactions from the participants that were positive included:

- *Yes, however after tomorrow I will be on vacation for a month.*
- *Yes, I would love to try out the tool with the current data.*
- *I know someone who might be interested, since she has data from a multi-variant disease in the Netherlands and the United Kingdom. She could use the same distribution as the vaccinated and non-vaccinated distribution in the HAVANA study.*

**Suggestions:** When hovering over a part of the visualization the number of samples appears in the top right. When hovering over a combination, there also appears one or multiple plots below the matrix. This was not always clear, especially when hovering over the linear diagram combinations.

Maybe another scale can be used for the matrix. At first, it was not clear which combination is larger at question 6 *Which combination of two types, has the highest Pearson value?*. However, the user noted that the linear diagram could have been used to answer the question. In the linear diagram the height of both bars can directly be compared, and it becomes clear which combination has the largest Pearson value, as is shown in Figure 6.5.

**6.2.3 Improvements**

We decided to change the following details to improve the understandability:

- Adding a legend to the linear diagram, to explain the background color of the linear diagram. The background color represents the degree of the combinations. The legend also works as a filter, such that users can, for instance, select all combinations with degree 3.
- Renaming several options; for example, degree was renamed to Nr Types/Sample.
• Improving the hover: First only the number behind Hover changed. We changed this to a more detailed hover as explained in Chapter 5. Behind each number it is described in more detail what the number represents, for example, “23 samples in the selection contain types 51 and 68.”
6.3 Final user evaluation

The participants were three epidemiologists and two lab analysts, four female and one male. Four out of five participants currently work on a HPV-study, of which two work on the HAVANA study. None of the participants participated in the early user evaluation. We used the same data set as in the early user evaluation. This way, users could not know the exact answers to the questions and we could ask the same questions as in the first session of the evaluation. Since the Sankey diagram was not included in the early user evaluation, we also included new questions for which users need the Sankey Diagram to find the answers.

6.3.1 Observations

We discuss the questions for which some participants needed hints or guidance to solve the question.

**Similar problems** For the following questions, we observed similar problems as in the early user evaluation. We tried to make the button for Question 14 stand out more, however most participants still needed a hint that it was possible to display the distributions with the button. Almost all users needed guidance to answer Question 17, as users started by selecting the combination of types 6 and 52, which changed the Pearson values.

- Question 14: *Which type occurs more frequently in samples from vaccinated persons as in samples from non-vaccinated persons?*
- Question 17: *In which year is the Pearson value the highest for the combination of types 6 and 52?*

**Improvements** The following questions gave problems in the early user evaluation. However, we noted that less participants needed hints to solve the questions and the reactions to the hints were different. For instance, after explaining that they could see the answer for question 3 in the degree histogram in the early user evaluation most reactions were *I did not know what degree meant* or *I would give degree another name*, while reactions in the second evaluation were more like *Of course, I should have known that.*

- Question 3: *How many samples have exactly two types?*
- Question 5: *How many samples contained types 45, 51 and 68?*
- Question 16: *Which combination with three types occurs the most often?*

**New problems** All participants needed some guidance through the first two questions concerning the development over time.

- Question 18: *How many persons that had type 35 in 2014, also had type 35 in 2015?*
- Question 19: *How many persons that had only type 66 in 2012, did not send in a sample in 2015?*

They needed to look at another view and the questions were about persons instead of samples. After explaining the Sankey diagram and helping the users through the first two questions, most participants could answer the third and fourth question about the development over time without any guidance.

6.3.2 Results

Table 6.4 shows the results of the understandability by the participants. The ratings for the linear diagram and matrix are higher than the ratings after the early user evaluation. The background color and ability to select all combinations of a degree might have increased the perception of the linear diagram. The participants in this evaluation also had a task for which they needed to select or brush a combination. This might have increased the understandability of the actions on the linear diagram and matrix.
Some users still found some parts of the visualization unclear. Two participants noted that they found the connection between the NR Types per samples distribution and the degree filter above the linear diagram difficult.

Two users found it unclear that after a selection, the size of the sets changed. They still looked at the total size of the sets, which was shown on the background after a selection, so that users can also compare relative sizes.

Furthermore, a small comment by most participants was: When clicking on a degree in the legend in the linear diagram it hides the combinations with that degree instead of showing all combinations with that degree. Which is exactly the opposite of what they expected, because other selections show what is selected instead of hide what is selected.

Finally, a participant found the notation of Size, Pearson, and Observed a bit vague.

Three participants could not think of something they were missing. One of the participants noted that it would be nice if there was an option to see the percentage of the samples after a selection. For instance, Figure 6.6a shows the number of samples from vaccinated and non-vaccinated persons after selecting all samples with HPV-type 66. In this case, it is still easy to compare the difference between the two values, as the difference between the number of samples from vaccinated and non-vaccinated persons is small. However, when there is an attribute with three different values that occur 10, 100 and 1,000 times in total, and those all occur 5 times after selection, it makes more sense to compare the percentages: 50, 5 and 0.5. Someone else noted that sometimes the difference between the Pearson value of samples from vaccinated persons and samples from non-vaccinated persons is small, such as in Figure 6.6b. She was wondering if the difference between the Pearson values was significant or not.

All participants were enthusiastic. Their answers mostly contained two or three of the following sentences:

- It is easy to analyze combinations of HPV-types
- I really like the colors
- A lot of data on a small screen in an understandable way
- I like the different way to look at the data we work with
- I got the feeling that I really understand the data
CHAPTER 6. EVALUATION

(a) Number of samples from vaccinated persons vaccinated and non-vaccinated persons for samples that contain HPV-type 66

(b) The expected, observed, Pearson and Fisher value for the non-vaccinated and vaccinated samples of combination of HPV-types 31,39

Figure 6.6: Two different details that were unclear in the evaluation

Dislike  Most of the participants had no remarks. One participant noted that she found some views difficult to understand. Another participant noted that she sometimes found it unclear where she could click and where she should click for certain questions.

Own data  All participants were positive about the tool and wanted to use the tool with their own data. Answers included *Immediately* and *Yes, a new round of the HAVANA study comes in two weeks*. The other participants also wanted to try the tool with their own data set. One of the participants was working on a similar data set, however, that data set did not include a timestamp. Another participant was also working on the HAVANA study, however, he used other attributes and was wondering if those also could be included.

6.3.3 Improvements

Given the limited time after the final user evaluation, we decided to only change the way the combinations are filtered. However, we provide suggestions and solutions for the problems in the next chapter.
Chapter 7

Discussion

In this Chapter, first the scalability is discussed. Next, we show that our approach is also applicable for other types of data via a case study. We conclude this Chapter with recommendations for future work.

7.1 Scalability

In this section we discuss the scalability of ComboVis. For the elements, sets and combinations we discuss the limitations and bottlenecks.

7.1.1 Sets

For the attributes and the Sankey diagram, the number of sets does not matter, since they do not display information about all sets. The linear diagram, matrix and set histogram can currently only handle a limited number of sets. When the number of sets becomes too large, the set histogram only has limited space to show which set belongs to which bar. The linear diagram has the same problem as the set histogram. When the number of sets becomes too large, there is only limited space below each combination to show which sets are in that combination. It also becomes more difficult to see exactly which circle below the combination belongs to which set.

In the matrix, it is a bit more difficult since the size of each cell becomes smaller when the number of sets increases. This makes it harder to see to which sets a cell belongs. This also means that the attribute distribution becomes smaller, which makes it harder to compare the distributions.

Currently, we suggest using ComboVis with data sets with at most 35 sets. Users can disable sets when loading a new data set, which is an easy fix when there are multiple sets that do not contribute to gaining new insights. ComboVis could also be improved to support more sets, which is described in the future work section. It is currently not possible to remove sets after loading the data set.

7.1.2 Combinations

Both the linear diagram and matrix have limitations in showing all combinations. Users can limit the number of combinations that are shown, by applying a filter that only shows all combinations that occur more often than the user-defined threshold.

The linear diagram can show all combinations as shown in Figure 7.1. However, details are less visible when the number of combinations increases. With around 100 combinations, the sets in each combination are still identifiable.

After selection of HPV-type 6, the matrix is shown as in Figure 7.2a. Some cells are still visible that are not in the row or column of HPV-type 6. For example, the cell of combination of HPV-types 31,51 is still colored. This means that the combination of HPV-types 6,31,51 occurs in the study.
CHAPTER 7. DISCUSSION

(a) 100 combinations are shown
(b) All combinations are shown

Figure 7.1: The matrix displaying the HAVANA data set with different selections.

After selection of the combination of HPV-types 6 and 51, the matrix is shown as in Figure 7.2b. Now there are still some cells visible that are not in the row or column of HPV-type 6 or HPV-type 51. For example, the cell of combination of HPV-types 18,66 is still colored. This means that the combination of HPV-types 6,18,51,66 is occurring in the study. So with interaction, it is possible to view combinations up to degree 4 in the matrix by selecting a set or a combination as shown in Figure 7.2b. However, users need to know what they are looking for and have to interact with the matrix to find those combinations. They can only find the combination of HPV-types 6,18,51,66 after selecting a combination of two of those HPV-types.

(a) HPV-type 6 is selected
(b) HPV-types 6 and 51 are selected

Figure 7.2: The matrix displaying the HAVANA data set with different selections.

7.1.3 Elements

All views, except the table, show aggregated information of the elements. The matrix, linear diagram, and Sankey diagram do not show information about individual element. If the number of elements increases, only the maximum value of the color scale of the matrix and the maximum value of the axis of the bars in the linear diagram change. In the Sankey diagram, the size of each node represents more
persons when the number of elements increases. Since those are all relative changes, the number of elements does not matter for those views.

When the number of elements increases, the histograms of the attributes and sets scale along. It does not matter if the maximum frequency of the sets is ten or ten-thousand. The frequency of the attribute values and sets are shown before the bar. If this number is 999 or smaller, only three characters are needed to show the number. However, we can convert the number to a three character notation by adding a multiplier. For example, we can convert 1600 to 1.6k, 16.000 to 16k, 900.000 to 0.9m etc. The only problem with this notation is that the exact number is not shown anymore. If someone wants to know the exact number, they can hover over the bar to see the exact number in the top left corner. So ComboVis can handle any amount of elements, however the performance of the tool does depend on the number of elements which is discussed in the next paragraph.

Computing constraints The performance of ComboVis depends on multiple factors:

- The number of elements
- The number of sets per element
- The number of visible combinations

When a user loads a new data set, all combinations have to be computed. When elements are in many sets, they are also in more combinations. When an element is in \( n \) different sets, the element is in \( 2^n \) different combinations. Since we have to compute the elements in each combination, the running time not only depends on the number of elements but also on the sparsity of the elements. It takes \( O(1) \) time to add an element to a combination, which is exactly \( \sum_{I \in F_C} C_I \). Adding all elements to the sets: \( O(|F_S| \times |M|) \) time, where \( F_S \) is the collection of set identifiers and \( M \) is the collection of elements. Adding all elements to the attribute values: \( O(|F_A| \times |M|) \) time, where \( F_A \) is the collection of all attributes.

The actions on ComboVis usually have a running time of the as follows: Compute the intersection between the brushed elements \( B \) and the visible combinations \( V_C \). For each combinations \( I \in V_C \), we have to compute \( |B \cap C_I| \). So the running time of each action is \( O(\sum_{I \in V_C} (|B| + |C_I|)) \).

Furthermore, the application is running on the client side of the browser, so the computing time also depends on the computer capacities of the user. For the HAVANA study this results in a loading time of 2s in Google Chrome on the current computers at the RIVM.

7.2 Other applications

Beside the HAVANA study, ComboVis can also be used to analyze other data sets. Other data sets have to be in the format shown in Table 7.1. In the case of the HAVANA study each HPV-type is a set, which can be replaced by other sets. In this section, we discuss other data sets which can analyzed with ComboVis.

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>ID</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Attr 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 7.1: Data input format for ComboVis
CHAPTER 7. DISCUSSION

7.2.1 Movies

We can use ComboVis to visualize a data set with movies. In this data set each element is a movie and each set is a collection of movies with the same genre. The format of this data is shown in Table 7.2.

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Action</th>
<th>Adventure</th>
<th>Children</th>
<th>...</th>
<th>Rating</th>
<th>Watches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toy Story (1995)</td>
<td>1995</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>3,2</td>
<td>701</td>
</tr>
<tr>
<td>Grumpier Old Men (1995)</td>
<td>1995</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>3,02</td>
<td>478</td>
</tr>
<tr>
<td>Waiting to Exhale (1995)</td>
<td>1995</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>2,73</td>
<td>170</td>
</tr>
<tr>
<td>Father of the Bride Part II (1995)</td>
<td>1995</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>3,01</td>
<td>296</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 7.2: The format of the data set of the movies

**Observations** When looking for certain relationships between genres, users know better what to expect as shown in Figure 7.3. For movies, some combinations make sense and some do not. For example, there are more children, fantasy, adventure movies than expected or there are more action, sci-fi, adventure movies than expected. This is also the case for combinations of movies that occur less often than expected. For example some of the combinations that occur less often than expected are: comedy and thriller or comedy and horror. There are many relationships between movie genres, but not many interesting questions arise about those relations as most of them only confirm our intuition.

![Figure 7.3: The deviation from the expected values for the combinations of genres](image)

**Limitations**

- Attributes: The attributes in the movie data set were different attributes as in the HAVANA data set. In the HAVANA set, we mostly dealt with categorical attributes with two or three different values; in the movie data set most attributes are numerical. Better support for numerical attributes would allow us to answer more interesting questions about the data. For example, whether the average rating of certain combinations of genres is higher or lower than the average or if certain genres were more popular in the past.

- No development over time: The data set did contain timestamps, namely the release year of the movies. However, the Sankey diagram is about multiple elements from the same source and each movie is only released once.

- Not independent: In the HAVANA study we assume that there are no relations between the different HPV-types. In the movie data sets we cannot assume that there are no relations between genres, since it is clear that fantasy and adventure occur more often than expected and children and horror less often than expected. In ComboVis it is not possible to integrate this knowledge, since it always uses the same computation for the expected value.
7.2.2 Actors

In the movies case study, the development over time of genres could not be investigated. We can investigate whether actors act mainly in movies with the same genres or in movies with different genres. For the 500 most popular actors, we looked at the genres of all the movies they acted in for each year for the last five years. The format of this data set is shown in Table 7.3. Furthermore, we added the gender of the actor and the number of movies in each year as attributes.

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>Gender</th>
<th>NrMovies</th>
<th>Action</th>
<th>Adventure</th>
<th>Animation</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>MarkHamill</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>2011</td>
<td>HarrisonFord</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>2011</td>
<td>CarrieFisher</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>2011</td>
<td>EllenDeGeneres</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>2011</td>
<td>TomHanks</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 7.3: The format of the data set of the actors

Observations There are many actors that act in many movies in one year. This means that the average degree of an element is a lot higher than in the HAVANA study and in the movies data set. The attribute distribution of the gender is shown in Figure 7.4. Since there are more male than female actors in our data set, the attribute distribution shows more male than female actors in each combination. However, in the combination of romance and drama this difference is much smaller as in other combinations.

Limitations

- Performance: Since the data is really clustered with a lot of genres for an actor in a year, the initial loading time is much longer. The loading time for the data set with actors is between three and five minutes, while the loading time of the movies data set is below three seconds. For example, there exists an actor with 13 genres in one year. This means that this element is already in $13^2 = 169$ combinations.

- Sankey diagram: Since those actors act in movies with different genres, the Sankey diagram loses two of the four options, namely only those two genres and no sample. The no sample does not occur because we can look up all movies from those actors.
7.3 Future work

**Selections** Multiple options to select samples in the data are not supported yet. For example, exclude all samples in which HPV-type 16 was found. It was not a priority to support complex queries in the HAVANA study. In other use cases, more complex queries might be needed to investigate the data set.

Sometimes users might want to continue working from where they left off or they might want to show something to colleagues. In those cases, it might be useful to save queries and configurations.

Finally, in AggreSet users have the ability to compare different selections. This way more complex selections can be compared against each other, such as all samples with HPV-Type 16 from vaccinated persons versus all samples from the second year from persons in cohort 2.

**Support numerical attributes** Currently, only categorical attributes are supported. Numerical attributes can be used, when they are binned into groups. However, there are other ways to integrate numerical attributes with less loss of information. It is for example possible to display the values in an area chart instead of a histogram. Instead of hovering over a bar, users could brush an interval between two values on the area chart.

More interesting information to display might be the average or median of the attribute. For example, is the average age in certain HPV-types or combinations higher than in others? This could be a replacement for the attribute distribution of categorical attributes.

**Improve scalability** We already discussed the limitations in the scalability of ComboVis. In this paragraph we suggest some solutions for those limitations.

*Elements:* The problem with the elements is mainly the performance. There are multiple solutions:

1. Server-side web application: In a server-side web application, the server can compute the requests faster than most browsers can. However, a disadvantage of this is that there needs to be a safe way to transfer confidential data to the server.

2. Stand-alone application: Stand-alone applications are not limited by the overhead of the browsers and can use the CPU of the users more efficiently. However, the advantage of a web application is that it does not need an installation procedure and therefore no extra steps are needed for the users to use the web application.

*Sets:* We suggest the following improvements to make it easier to handle data sets with more than 35 sets:

1. Aggregation: Combining sets in multiple groups can be a solution to reduce the number of sets.

2. Filtering: Let the user easily swap between the sets they want to investigate. This way users can select the sets that they are interested in. This is already applied in UpSet.

3. Scrollable: Let the user scroll through the list of sets, for instance to find the sets they are interested in. This also means that not only the set histogram has to be scrollable, but also the matrix and linear diagram. This is already applied in AggreSet.

*Combinations:* We suggest the following improvements to increase the number of interesting combinations:

1. Scrollable: Make the linear diagram scrollable instead of decreasing the width of the columns.

2. Filtering: Apply different or more filters to show only the interesting combinations. Most of the times, not all combinations are interesting so filtering the combinations might be a good solution. However, each user might be interested in different combinations thus it should be easy to express which combinations are interesting and which are not. Now we only show combinations that occur more than X times. Another filter might be on all combinations which occur significantly more often or less often as expected.
**Options for histograms**  The histograms now always show the number of elements in each value or set relative to the maximum number of elements in one of the bars of the histogram. After selection, the new bars are not scaled to the new maximum. This means that the same height always represents the same number of elements. After selections, bars sometimes become really small. Then it is hard to compare 10/2000 versus 16/2000, but it would have been easy to compare 10/16 versus 16/16. So, multiple options are possible to improve the usability of the histogram:

1. Scaling to the selected maximum.
2. Show percentage instead frequency

**Sankey diagram**  In the case study with the movies, we saw that there were multiple limitations to the Sankey diagram. In the domain of set visualization, not much research has been done to show the development of sets and combinations over time. However, in the area of visualization of events over time more research has been done. Integrating one of those techniques instead of the Sankey diagram could solve some of the problems, such as binding events to multiple timestamps.
Chapter 8

Conclusions

In this final chapter, we look back at the research questions and the original goal of this Thesis.

**Can we use visualizations to display deviating combinations?** For example: Can we find all combinations of variants that occur significantly more than expected? Yes, we can use ComboVis to quickly find deviating combinations. Since we can easily limit the combinations such that ComboVis can display display deviating combinations in the HAVANA study, it is easy to find and compare those combinations. In the HAVANA study, the user-defined threshold could filter out most combinations which were not interesting enough. When a combination occurs less frequently than the user-defined threshold, it is unlikely that the combination still occurs more than expected. For less frequent combinations, the lower bound of the 95% confidence is often below zero, which means that those combinations do not occur less often than expected.

**Can we use interactive visualization to analyze combinations in the context of other attributes?** For example: Can we find all combinations of variants that occur more often in non-vaccinated persons than in vaccinated persons? Yes, we can easily compare non-vaccinated to vaccinated persons in combinations and types. This was the most important attribute in the data set, but it is not hard to integrate other attributes. However, when there are too many values for a categorical attribute the distribution becomes unclear. While there is currently no support for numerical attributes, it is possible to integrate support for those attributes in ComboVis.

**Can we use interactive visualization to display time-varying infectious disease data?** For example: Can we visualize which variant people had preceding or following a specific combination of multiple variants? ComboVis provides a Sankey diagram to display longitudinal infectious disease data. However, this works because all persons are measured each year. This means all samples are collected at the same time, which is usually the case in longitudinal infectious disease studies. In other data sets this is usually not the case, for example, in the case study of the movies we had to transform the data to investigate changes over time. Those transformation are causing a loss of information. Furthermore, it is still difficult to extract all information from the Sankey diagram and users found this diagram difficult to understand.

The focus of this thesis was on the interactive visualization of longitudinal multi-variant infectious disease studies. The goal is to help researchers understand the collected data more efficiently. We think ComboVis is an effective and efficient way to extract the information about such studies and thus can help researchers understand the collected data more efficiently. However, more practical tests are needed to see if ComboVis contributes to research by helping understanding the data more quickly or providing researchers with new insights.
Bibliography


Appendix A

Statistics

A.1 Deviance

When we want to examine whether a combination of two HPV-types occurs more or less frequently as expected in the study, we can investigate the difference between the observed and the expected frequency. However, the absolute difference between 20 and 25.9 is the same as the absolute difference between 250 and 255.9, while the relative difference is smaller in the second example. This means that we need another value to standardize the difference between the expected and observed between multiple combinations. We use the Pearson Standardized Residuals for Testing Independence.

Table A.1: Frequency matrix for Type 1 and Type 2

<table>
<thead>
<tr>
<th></th>
<th>Type 2</th>
<th>Not Type 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Not Type 1</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>n</td>
</tr>
</tbody>
</table>

The frequencies are usually defined in a frequency matrix, such as in table A.1. With the values in this table, we can define the following variables to compute the Pearson Standardized Residuals for Testing Independence for the frequency matrix in Table A.1:

Observed frequency of type 1 and type 2:

\[ o = a \]

Expected frequency of type 1 and type 2:

\[ e = \frac{(a + b)(a + c)}{a + b + c + d} \]

Probability of type 1:

\[ p_1 = \frac{a + c}{n} \]

Probability of type 2:

\[ p_2 = \frac{a + b}{n} \]

The Pearson Standardized Residuals for Testing Independence:

\[ p = \frac{o - e}{\sqrt{e(1 - p_1)(1 - p_2)}} \]

The Pearson Standardized Residuals for Testing Independence claims with 95% confidence that two types occur way more or less than expected when the Pearson value is above 2 or below -2.
Example  Given the data shown in Table A.2, the Pearson Standardized Residual and other quantities are:

Observed frequency of type 1 and type 2:

\[ o = 20 \]

Expected frequency of type 1 and type 2:

\[ e = \frac{130 \times 265}{1330} = 25.9 \]

Probability of type 1:

\[ p_1 = \frac{265}{1330} \]

Probability of type 2:

\[ p_2 = \frac{130}{1330} \]

The Pearson Standardized Residuals for Testing Independence:

\[ p = \frac{20 - 25.9}{\sqrt{25.9(1 - 0.2)(1 - 0.1)}} = -1.37 \]

Since the Pearson value is between -2 and 2, we can say with 95% confidence, that the combination does not occur more or less frequently than expected. When the observed frequency of the combination would be 50 and the other numbers would not change, the Pearson value would be:

\[ \frac{50 - 25.9}{\sqrt{25.9(1 - 0.2)(1 - 0.1)}} = 5.58 \]

Since 5.58 is greater than 2, type 1 and type 2 occur significantly more frequently than expected.

Data notation  For a combination \( I \), we can define the following definitions, where \( M \) is the collection of elements in the data set.

The expected value of a combination \( I \):

\[ E_I = |M| \prod_{i \in I} \frac{|S_i|}{|M|} \]

The probability in the divider of the Pearson value for a combination \( I \):

\[ P_I = \prod_{i \in I} (1 - \frac{|S_i|}{|M|}) \]

The Pearson value for a combination \( I \):

\[ D_I = \frac{|C_I| - E_I}{\sqrt{E_I P_I}} \]
A.2 Fisher’s test

The Pearson Standardized Residuals for Testing Independence does not work that well for smaller frequency matrices. The frequency matrices become too small for the Pearson Standardized Residuals for Testing Independence when the expected value of the combinations is lower than 5. Fisher’s exact test is a statistical significance test used in the analysis of frequency matrices, such as table A.1. As we expect combinations with an expected frequency lower than 5, we additionally perform the Fisher’s exact test to determine the significance of the deviation. The result is significant, when the outcome is below 0.05. The significance of the Fisher’s test is given by:

\[ P_{\text{Fisher}} = \frac{(a+b)(c+d)}{a(a+c)(n)} \]

When we apply it to our example from table A.2, we find:

\[ P_{\text{Fisher}} = \frac{265 \cdot 1165}{20 \cdot 1330} = 0.2033 \]

Since the Fisher’s test returns a p-value greater than 0.05, we can again say that the combination does not occur significantly less frequently than expected.
Appendix B

Evaluation Questions

B.1 Early user evaluation

B.1.1 Tasks

1. In which year most samples were sent in?
2. In how many samples was type 54 found?
3. How many samples have exactly two types?
4. How many samples contained both type 6 and 66?
5. How many samples contained types 45, 51 and 68?
6. Which combination of two types, has the highest Pearson value?
7. What is the Pearson value for the combination of types 31 and 33?
8. Which combination of two types occurs most frequently in year 2015?
9. How many samples contain only type 51?
10. Which types do not occur in samples from vaccinated persons?
11. Which type occurs the most frequently in samples from persons from cohort 2 in 2012?
12. Which combination of two types has the highest Pearson value from the samples in vaccinated persons?
13. How many samples contain type 51 or 52 or both?
14. Which type occurs more frequently in samples from vaccinated persons than in samples from non-vaccinated persons?
15. Which combination occurs more frequently in samples from vaccinated persons than in samples from non-vaccinated persons?
16. How many combination of three types, occur at least 6 times?
17. Which combination has the highest Pearson value in the samples from year 2013 and 2014?
18. In which year is the Pearson value the highest for the combination of types 6 and 52?
Table B.1: Questions for the understandability of the views

**B.1.2 Questionnaire**

1. Did you find the sorting useful?
   - Not very useful
   - Not useful
   - Neutral
   - Useful
   - Very useful

2. Did you find the switching between the Observed frequency and Pearson value useful?
   - Very unclear
   - Unclear
   - Neutral
   - Clear
   - Very clear

3. Did you find the text easy to read?
   - Very unclear
   - Unclear
   - Neutral
   - Clear
   - Very clear

4. If no, which were unclear?

5. Did you find the tool initiative?
   - Very unclear
   - Unclear
   - Neutral
   - Clear
   - Very clear

6. If no, what was unclear?

7. Is there something you miss?

8. What did you like about the tool?

9. What did you dislike about the tool?

10. Would you like to use the tool with your own data?
B.2 Final user evaluation

B.2.1 Tasks

1. In which year most samples were sent in?
2. In how many samples was type 54 found?
3. How many samples have exactly two types?
4. How many samples contained both type 6 and 66?
5. How many samples contained types 45, 51 and 68?
6. Which combination of two types, has the highest Pearson value?
7. What is the Pearson value for the combination of types 31 and 33?
8. Which combination of two types occurs most frequently in year 2015?
9. How many samples contain only type 51?
10. Which types do not occur in samples from vaccinated persons?
11. Which type occurs the most frequently in samples from persons from cohort 2 in 2012?
12. Which combination of two types has the highest Pearson value from the samples in vaccinated persons?
13. Which type occurs the most frequently with types 6 and 51?
14. Which type occurs more frequently in samples from vaccinated persons than in samples from non-vaccinated persons?
15. Which combination occurs more frequently in samples from vaccinated persons than in samples from non-vaccinated persons?
16. Which combination with three occurs in the most samples?
17. Which combination has the highest Pearson value in the samples from year 2013 and 2014?
18. In which year is the Pearson value the highest for the combination of types 6 and 52?
19. How many persons that had type 35 in 2014, also had type 35 in 2015?
20. How many persons that had only type 66 in 2012, did not send in a sample in 2015?
21. How many persons that had type 52 and at least one more type in 2012, still had type 52 and at least one more type in the next year?
22. How many vaccinated persons that had type 52 and at least one more type in 2012, still had type 52 and at least one more type in the next year?

B.2.2 Questionnaire

<table>
<thead>
<tr>
<th>Part</th>
<th>Matrix</th>
<th>Linear diagram</th>
<th>Set distribution</th>
<th>Attribute distribution</th>
<th>Persons over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visualization</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
</tr>
<tr>
<td>Hover</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
</tr>
<tr>
<td>Click</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
<td>- - 0 + ++</td>
</tr>
</tbody>
</table>

Table B.2: Questions for the understandability of the views
APPENDIX B. EVALUATION QUESTIONS

1. Did you find the text easy to read?
   Very unclear  Unclear  Neutral  Clear  Very clear
2. If no, which were unclear?
3. Did you find the tool initiative?
   Very unclear  Unclear  Neutral  Clear  Very clear
4. If no, what was unclear?
5. Is there something you miss?
6. What did you like about the tool?
7. What did you dislike about the tool?
8. Would you like to use the tool with your own data?
Appendix C

Implementation

ComboVis is a web application written in JavaScript, CSS and HTML. For the implementation of the visualizations we used the library d3.js. The source code of the project can be found on: https://github.com/Barttje/ComboVis/tree/master/Project/DATA. In Figure C.1 an overview is given of the implementation.

**Data loader**  The data loader loads a CSV file and processes the columns. The data loader inspects the format of each column to determine whether the column represents a set or attribute. The result is shown to the user, who can change the type of each column. For example, to change sets to attributes or to remove attributes that are not interesting.

**Setup**  In the setup all information is initialized. First all data is stored in another format. Each combination, set and attribute value are pointing to their corresponding elements. This way all visualizations are represented by a collection of elements and it allows us to quickly compare the elements that are brushed. After all data is in the correct format, the interface and other information for the visualizations are determined. For example the colors, threshold, selections, height and width. Finally the interaction on each of the visualizations is initialized. Each interaction only recomputes the relevant information. For example, after selecting a sort method, the new sort order of the sets or combinations is determined. After recomputing, the interaction adjusts the visualizations where needed.

![Figure C.1: Overview of the implementation of the tool](image)