Otto-von-Guericke University, Magdeburg

Faculty of Computer Science

Project Report

Visual Analytics Support for Analysis of Cohort Study Data: Requirements and Concepts

Author:

Lena Cibulski

August 10, 2016

Advisor:

Prof. Dr.-Ing. habil. Bernhard Preim

Department of Simulation and Graphics
Otto-von-Guericke University, Magdeburg

1 Introduction

Epidemiology studies health-related conditions in a population to derive disease-specific risk factors. Cohort studies, where groups of individuals are observed in terms of various aspects, are conducted to investigate the risk of developing a target disease, when being exposed to certain factors. Such factors can be numerous, for example behavioral factors (e.g. smoking), socio-demographic factors (e.g. marital status, education), or genetic factors [FFF12].

For data generation, subjects are asked to participate in interviews, medical examinations, and image acquisition. Consequently, the resulting cohort study data is large and heterogeneous, which makes an analysis challenging. So far, analyzing cohort study data is conventionally performed using strongly hypothesis-driven approaches. For this purpose, assumed correlations between medical conditions, life-style related features, and socio-demographic aspects with respect to a target disease are statistically assessed. The underlying hypotheses originate from observations that clinicians make in their daily routine. Research with respect to cohort studies is an interdisciplinary field: physicians, statisticians, and computer scientists share their knowledge to identify associations between risk factors and diseases [KLG+16].

Given the heterogeneity and high dimensionality of cohort study data, a hypothesis-driven statistical analysis might miss relevant correlations. Visual analytics has proven to be of use for exploration of large and high-dimensional data sets [Kei01]. In particular, if no concrete starting hypothesis is available, such techniques can be useful for hypothesis generation based on data exploration.

The aim of this project is to develop a concept for a comprehensive visual analytics system, which integrates subspace clustering, decision trees and regression together with standard techniques like histogram and scatter plot for interactive visual analysis of cohort study data. The system is intended to be expandable by further techniques, such as association rules, which could be integrated subsequently. We aim at supporting epidemiologists in their workflow and at allowing them to gain new insights into their data by providing various techniques. This also enables hypothesis generation based on their findings.

We plan to develop a system that does not only have exemplary character, but that supports epidemiologists in their daily practice. For this purpose, the concept is evaluated in close collaboration with domain experts from the University Medicine of Greifswald to ensure practical relevance.

1.1 The Study of Health in Pomerania

In this project, we focus on data that results from the Study of Health in Pomerania (SHIP), which attempts to describe health conditions with respect to a wide range of diseases [VAS+11]. The study is realized under the guidance of the Institute for Community Medicine at the University of Greifswald. It started in 1997 with a population sample of 6265 subjects living in West Pomerania (SHIP-0). Follow-up examinations were conducted between 2002 and 2006 (SHIP-1) and between 2008 and 2012 (SHIP-2). A third follow-up has just started in 2016. Parallel to SHIP-2, a second study was started with baseline examinations between 2008 and 2011 (SHIP-TREND). Examinations include interviews, laboratory data, sleep monitoring, ultrasound and MRI image acquisition as well as medical, dental, and dermatological examinations.

The SHIP significantly helps to determine the prevalences of risk factors and diseases and has already contributed to findings on increased levels of obesity and high blood pressure as well as hepatic steatosis.

1.2 Previous Work

Cohort study data provided by the SHIP has already been subject to various analysis approaches that resulted in several findings concerning disease prevalences, risk factors, and population-representative reference values. Visual analytics approaches have also been established based on the SHIP data – at Otto-von-Guericke University primarily by Paul Klemm, who proposed visual analysis techniques for both explorative and confirmative analysis of population study data in his doctoral thesis [Kle16]. In the following, we will briefly describe a selection of the techniques.

Paul Klemm and colleagues present overview visualizations based on non-image variables, such as socio-demographic factors, but also introduce novel approaches, where both image and non-image variables are linked to identify associations in the data. For overview visualizations they propose the so-called decision tree quality plot [KGL+15] as well as a 3D regression heat map [KLG+16].

A decision tree quality plot contains individual decision trees as data points in a scatter plot, with tree size and classification error on the axes. The decision trees are created with the C4.5 algorithm for each non-image variable using image-derived variables as split attributes. The tree size was chosen as a measure of tree complexity, as a classification with a few precise rules is aimed at. However, there are other metrics, e.g. the tree depth, that might yield different results, depending on the application. In this way, they analyze the predictive power of the spine shape for non-image features, e.g. back-pain indicators.

A regression-based approach enables statistical assessment of relations in cohort study data. The results of the statistical analysis are visualized using a three-dimensional regression heat map [KLG+16], which uses saturated colors for strongly correlating variables. In this way, the heat map serves as an overview visualization highlighting potentially interesting associations between variables and a target disease.

For an explorative analysis of data containing medical images, information from image segmentation is linked to non-image variables to unveil associations. For example, shape-based clusters of participants are created, which can then be related to non-image variables to identify correlations of the shape to other socio-economic or medical variables [KLR+13].

The analysis of organ shapes and epidemiological features is advanced by providing a web-based framework that combines information visualization techniques, epidemiological data representations, statistical measures, and shape-based techniques [KOJL+14]. Multivariate visualizations establish a connection between image-derived and non-image features by augmenting standard techniques from information visualization with renderings of mean shapes.

A recent survey on image-centric cohort studies and strategies for evaluating the resulting heterogeneous data is also given by Preim et al. [PKH+16].

2 Initial Thoughts and Concept

The key question guiding this project is the following: how can we use visual analytics techniques to support epidemiologists in their daily work?

To answer this question, we have to get an idea of the current state concerning workflows and requirements in the field of epidemiology. As epidemiologists themselves do not have much time, the situation analysis is in a first step carried out based on literature and discussions with people that have been working in related domains for quite some time. In this way, we intend to extract the common workflow, techniques that are currently used, analysis tasks, and general requirements that have to be considered when designing a system for epidemiologists.

Based on aspects derived from the situation analysis, we develop a concept for a visual analytics system in terms of a prototype that incorporates an initial layout and potential functionality. The implemented prototype serves as material for later discussion with the domain experts, where concrete ideas and functionality can be evaluated. From there, further requirements and issues to be considered can be raised. In this way, the initial concept can contribute to an efficient exchange of information.

2.1 Situation Analysis

Epidemiological Workflow Deriving risk factors is traditionally realized by statistically assessing the correlation of selected epidemiological features with the target disease. As suggested by Klemm et al. [KOJL+14], it can be divided into different steps as follows:

1. Physicians derive hypotheses from observations in daily routine.
2. Epidemiologists compile a list of epidemiological features that model the hypothesis and identify confounding features to be included.
3. Epidemiologists define a cohort that fits the guiding hypothesis.
4. Statisticians assess the association of selected features with respect to the investigated disease.

If the initial hypothesis could not be validated or follow-up questions arise during the statistical analysis, the described steps are performed iteratively to explore alternatives [ZGP14]. Depending on the investigated disease, a hypothesis may also focus on a specific group of participants, e.g. men aged over 50 years. In this context, validating the hypothesis might require a comparison of the considered sub-population to either the overall population or other groups of participants with certain characteristics, e.g. complement sub-populations.

Reproducibility is a key requirement, both when conducting a population study and when analyzing the resulting data [Kle16]. The first is related to minimizing intra- and inter-observer variability, while the latter means that all conclusions drawn from data analysis have to be reproducible and statistically valid.

For their purposes, epidemiologists mostly use basic visualizations, which help them understand the fundamental characteristics of their data [Kle16]. Histograms are used to access the distribution of numerical data, for example age or weight. Bar charts for categorical variables fulfill the same purpose. For visualizing descriptive statistics of a population, namely minimum, lower quartile, median, upper quartile, and maximum as well as outliers, with respect to a specific variable they use box plots [BBK13]. The bivariate relationship of two numerical variables is graphically represented using a scatter plot. Pairs of values are displayed as data points, which results in a point cloud that may provide insight into the variables' dependency structure. Often, a regression line is drawn into the scatter plot to depict the relationship of both variables. Kaplan-Meier plots are used to depict the probability of a participant not to be affected by an event, such as death [BBK13]. In this context, such plots are often used for visualization of survival rates. They are a special case of line charts, which are commonly used to visualize the relationship between an arbitrary variable and time. On a very detailed level, tables are used to clearly convey exact values for variables or derived measures, such as correlation measures.

To summarize, we can say that statistical analysis is essential in the epidemiological workflow to validate hypotheses and therefore create medical knowledge. Reproducibility of results also plays an important role, as conclusions that do not meet the requirements will not be accepted by the epidemiological community [PKH+16]. So far, advanced visualizations are not used; results are commonly presented using statistical standard diagrams or tables.

Analysis Tasks Aiming to study the causes of diseases, epidemiologists face different tasks when analyzing cohort study data. To properly integrate visual analytics techniques in the epidemiologists' workflow and to enable a thorough exploration of cohort study data, these analysis tasks need to be considered. As can be obtained from literature, there are two main research aims: (1) determination of hypotheses based on data exploration and (2) hypothesis validation, i.e. is a statistical power given? Based on our literature review, we identify more concrete tasks for analysis of epidemiological data and classify them.

The following tasks address the main research questions that epidemiologists aim to answer by analyzing cohort study data: (A1) characterization, (A2) association-causation, (A3) sensitivity analysis, (A4) temporal changes, and (A5) cohort comparison.

During an analysis session, epidemiologists explore several alternatives; if assumptions proved to be wrong or to obtain comprehensive insights into the given data. For this purpose, they need to independently adjust investigated data subsets to their needs. Corresponding analysis tasks are: (B1) variable selection, (B2) cohort definition, and (B3) cohort modification.

Finally, there are tasks concerned with interactive visual analysis features on a high level of abstraction: (C1) assess data quality, (C2) gain an overview of the data set, (C3) view details for a selection, (C4) analyze large data sets, (C5) flexibly choose visualizations, and (C6) re-view performed actions.

A complete list with more detailed descriptions for each of the analysis tasks can be found in Appendix A.

2.2 Derived Requirements

Based on extensive literature research and the described situation analysis, we identify basic requirements that need to be considered when thinking about a concept for an interactive visual analysis system. Experiences concerning the application of different requirements engineering techniques to the field of epidemiology, such as interviews, user observation, and domain knowledge workshops, can be found in the work of Thew et al. [TSP$^+$09]. They focus on a scenario-based design and perform iterative cycles of requirements analysis, design exploration, and user feedback to gain knowledge of the domain and to develop a system that supports the epidemiologists' research questions.

Hypothesis Generation When performing their analyses, epidemiologists are mostly focused on hypotheses proposed by clinicians. Hypothesis generation in the form of finding associations without knowing where to start is rarely performed. It requires quite an effort, as the entirety of variables and subjects has to be taken into account. Techniques for hypothesis generation therefore need to be highly scalable to cope with a large number of dimensions.

We believe that visual analytics techniques are of benefit for such applications. Visual representations simplify getting an overview of the characteristics of the data, e.g. the variables' distributions. *Seeing* how certain variables are connected, i.e. how one of them changes as another one is manipulated, may reveal starting points for a hypothesis-based analysis as commonly carried out.

Designed systems therefore need to allow users to view their data from different perspectives and enable constant switching of the focus between different data representations and analysis modes (i.e. explorative and hypothesis-based). This especially holds for cases, where a large number of variables and combinations are worth considering.

Definition of Sub-populations Investigated hypotheses might be about sub-populations that exhibit certain characteristics, e.g. similar smoking behavior. For explorative analysis, sub-populations and their comparison play an even bigger role. Epidemiologists therefore need to be able to flexibly define sub-populations of participant with respect to certain criteria. This could either happen in an interactive way, e.g. using visual interfaces, or (semi-)automatically with respect to optionally given constraints, i.e. an algorithm identifies interesting participants from a large population.

Because the generation of medical knowledge and its acceptance strongly rely on the statistical relevance of detected associations, it is important to show the number of participants that are contained in one sub-population, as significance is not given if too few participants were considered for the validation of a hypothesis [FFF12]. Cohorts might also need to be refined during an analysis session, for example if participants can be excluded based on certain findings. Refinement also includes the expansion of a sub-population. Epidemiologists need to be supported in this task to enable a smooth analysis process.

To estimate the importance of a sub-population for a specific association, it must be compared to other sub-populations or to the overall population with respect to a certain focus. This task can be simplified by providing the possibility to simultaneously view the desired characteristics of both (or multiple) sub-populations, which enables a direct comparison.

Data Security Data security is certainly an issue when developing a system to be used for epidemiological purposes. Whenever working with epidemiological data, the confidentiality of participant information must be maintained. Following existing regulations on data privacy and data security, individuals must give their informed consent to a permanent storage of certain aspects of their personal data. Another solution to respect data privacy is an anonymization of the data. The issue of data security is about ensuring that no natural person can be identified by a profile that was created using a combination of identifiers from different sources and other information. Accordingly, applying for access to epidemiological data, which are part of an individual's personal data, is a long and elaborate process.

As a consequence, the developed system has to follow strict guidelines and has to meet certain requirements in order to maintain data privacy. Unauthorized access to the participant's data needs to be inhibited.

Communication The development of the visual analytics system is intended to be performed in iterative cycles of discussions with domain experts and adjustments based on the feedback. As we collaborate with domain experts in Greifswald, while working in Magdeburg ourselves, personal discussions, where domain experts and developers view and talk about the software together, are only possible in a limited number due to the travel expenses. As a consequence, we have to find a way to distribute system updates and changes to the domain experts with the least effort possible.

2.3 Initial Concept

Based on the situation analysis (Section 2.1), we identified three central components of an epidemiological analysis: (1) definition of sub-populations, (2) statistical validation of hypotheses, and (3) data exploration.

To support all of the components, our idea is to provide subspace clustering, decision trees, regression, and standard information visualizations for a thorough analysis of epidemiological data. We intend to support hypothesis generation based on findings that are made during analysis. Subspace clustering has not yet been established for the field of epidemiology, but could be beneficial for automatically defining sub-populations with certain characteristics for further analysis. Decision trees are already used in the health care domain, primarily for decision-making concerning individual patient treatment, but have only rarely been applied to epidemiological data. They can help to assess the predictive power of variables with respect to a target. In contrast to decision trees, regression approaches are common practice in epidemiology. Nevertheless, overview visualizations and visual representations of statistical results can be beneficial for drawing conclusions about associations in the data. Information visualizations, such as histogram or scatter plot, show the data from different perspectives and in this way support data exploration.

When looking at the components, one might ask whether it would be suitable to provide one single tool for each of the described techniques. If epidemiologists concentrate on one approach at a time anyway, for example data exploration or regression analysis, then providing a complex system might be overwhelming for them. On the other hand, single tools do not support direct transfer of analysis results to other techniques to be investigated from an additional perspective. We believe that a simultaneous consideration of all the mentioned techniques holds great potential for the analysis of epidemiological data. Therefore, we aim at combining them into one complex system that makes use of a linking and brushing approach to connect different visual data representations.

As personal meetings involve certain travel expenses and time investment, we believe that modern web technologies should be used to provide fast communication between development and expert input. Data security is closely connected to this issue, as hosting the system and the data (at least temporarily) online might make it open to unauthorized access from the outside. HTML5, CSS3 and JavaScript can be used for basic UI elements. We intend to use the Data-Driven Documents (D3.js) [BOH11] library for visualizations. It allows to simply attach data to individual visual representations and provides powerful transformation and mapping tools. For the computation of regression measures, we want to use R[1], a software environment for statistical computing, which is already in use for a concurrent regression project[2] that is considered to be integrated into our system. These two approaches can be brought together by Shiny[3], a web application framework for R that can be connected to the D3.js library.

Explorative vs. Guided Analysis An important decision that needs to be made prior to the the design process is whether to concentrate on supporting an explorative analysis or a guided analysis.

An explorative analysis aims at providing insight into the given data, by using different visualization techniques. Presenting data in a visual form and allowing users to interact with it helps to draw conclusions, as the data set is explored [Kei01]. Different representations can be considered simultaneously to deepen the understanding. Exploration is the prerequisite for hypothesis generation. It is well-suited for cases, where little is known about the data and no starting hypothesis is available. An explorative analysis could be realized in the form of coordinated multiple views, which allow a simultaneous analysis of different data representations and support exploration by a linking and brushing approach, where data subsets can be selected and are displayed accordingly.

Guiding an analysis means to support the epidemiologists in deciding what to do next, for example by offering only certain techniques for the next step that suit to the action that was performed before. Steps can also be skipped according to the users' preferences and priorities. As users can only decide from a pool of suggested techniques to use in a following step, the flexibility of a guided analysis is limited. Hypothesis generation is not supported, because the data cannot be freely explored.

A guided analysis is straightforward and does not require much training effort. It also supports one of the key requirements in epidemiology: reproducibility. The sequential structure of a guided analysis makes it relatively easy to follow the steps that were performed and in this way create a history of the analysis

[1] Open Source: r-project.org

[2] Contact persons: Mehdi Arian Alborzi and Daniel Schneider

[3] Open Source: shiny.rstudio.com

session. The user can step back and forth in the analysis process to re-view the sequence of actions she performed. Analysis states could even be saved and loaded, for example to apply the same analysis procedure to a different data set to see how findings can be reproduced. Saving analysis states can also be used for communicating results to colleagues, as they can easily understand what was done in the analysis.

Possible realizations of a guided analysis are the string of pearls metaphor and the film strip metaphor. For the first one, analysis steps can be viewed as pearls on a string, where the string stands for the guiding direction. Individual steps can be highlighted, for example by enlarging the corresponding pearls. If multiple possible steps are highlighted, one can also skip steps in between. The film strip metaphor is implemented in the work of van den Elzen and van Wijk [vdEvW13]. They propose a visual exploration system, where multiple alternatives from the current state are offered as small multiples to enable exploration of the next possible steps. The alternative chosen by the user can then be analyzed using large singles. Intermediate steps in the analysis process are preserved and can be revisited.

Supporting Data Exploration Following the aim of hypothesis generation, we decided to realize a flexible approach in the form of a coordinated multiple views system. Jonathan Roberts provides a survey of coordinated and multiple views in exploratory visualization, where he summarizes the state-of-the-art of related topics, such as data processing, view generation, exploration techniques, and infrastructure [Rob07]. The system is intended to support epidemiologists in exploring their data to identify and validate associations between variables for derivation of medical knowledge. Linking and brushing methods allow to select (= brush) data points in one view and concurrently highlight the same elements in any other linked view. In this way, the selected subset of data points can be viewed for detailed analysis, compared to the entire data set or other subsets, or outliers can be detected between multiform views.

As we want to integrate both visualization as well as data mining and statistics techniques, we generally need two domains: (1) a visualization domain, where training data or computation results can be visualized, and (2) an active data mining domain, where parameters for decision tree induction, regression, or subspace clustering can be determined and data mining is performed. The focus between both domains can be iteratively switched, i.e. data mining results are visualized and computation parameters are adjusted based on findings observed in the visualizations.

To retain the possibility of belatedly extending the system by additional functionality, we focus on a plug-in based approach. Each technique is realized as an individual view with a graphical user interface and a canvas for visualizations. Such a view can be completely tuned to its corresponding technique. All views are then arranged together with a menu bar and a sidebar in a global layout.

Different layouts allow for a highly flexible analysis of data mining and statistical techniques next to standard visualizations like scatter plots and histogram, which provide additional information about the investigated data. The system can easily be extended by simply creating a view for an additional technique, which is then automatically integrated with the existing system.

The main components mentioned in the beginning of this subsection are transferred to three analysis branches for the system: (1) cohort definition (subspace clustering), (2) validation (statistical analysis), and (3) exploration (standard information visualization). We call them *modes*. Each of the views is assigned to one of these modes. Adding a new view works in the following way: either the new technique belongs to one of the analysis modes, then it is added to the list of views related to the respective mode. Or the view takes on a completely new direction, then another mode can simply be added, to which the view is assigned. The analysis modes and individual views can be switched at any time during analysis or views from different modes can be considered simultaneously.

As an example: if an epidemiologist already has a hypothesis and only wants to statistically validate it, she can simply use the validation mode to perform his analysis. This procedure corresponds to the common workflow she is familiar with. If she now identifies an interesting relation in the data, she can additionally open another view from the same or a different mode and, based on his findings, perform an exploration there. In this way, a variety of workflows can be supported, from the common one that epidemiologists currently follow to a workflow that combines a variety of different techniques.

The Prototype Using Axure RP Pro[4], we implement a prototype that incorporates the layout and functionality previously described. Figure 1 shows the prototype's user interface with a menu bar and four panels. The menu bar (Figure 1, top) contains basic actions such as load and save as well as view- and data-specific operations. Saving and loading of analysis states is provided to enable an export of data and results for presentation or communication. In the menu bar, the layout can be changed from one single canvas to either two or four canvas. The double canvas can be seen in Figure 2. The multiple canvas layout supports simultaneous analysis of modes or views, while the single canvas supports straightforward analysis using one specific mode. The menu bar also contains tools for brushing, e.g. creating or reversing a brush.

Defined sub-populations, independent from how they have been defined, are listed in the cohort panel (Figure 1a). Initially, this list contains the cohort consisting of all participants in the underlying data set. Another sub-population contained in this list can correspond to one of the clusters yielded by the subspace clustering algorithm, which contains participants with similar characteristics. Users can also define a sub-population by brushing a certain number of participants according to some criteria and then exporting the current brush as a

[4] www.axure.com

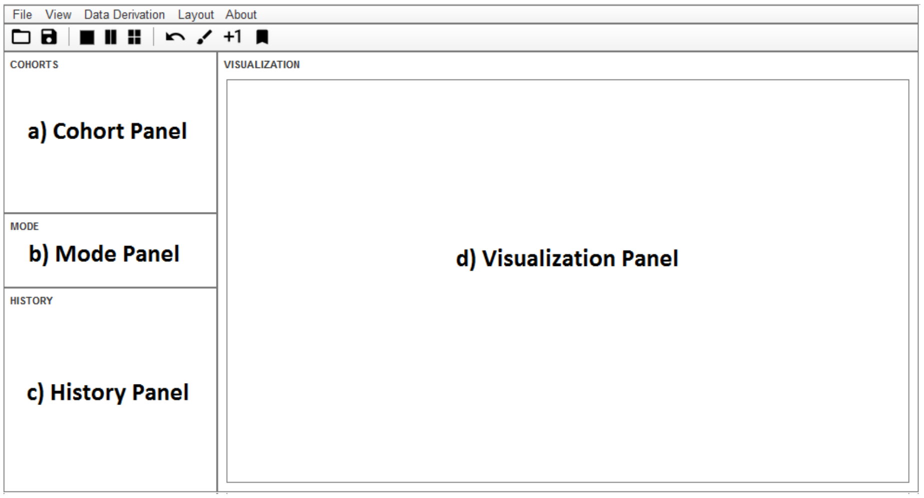

Figure 1: *The system consists of a menu bar (top) and four panels. The side bar consists of (a) a list of defined cohorts for re-use, (b) a panel for choosing the analysis mode, and (c) a history panel to preserve performed actions. A selected cohort can be displayed in the (d) visualization panel.*

cohort, therefore adding it to the cohort list. To apply an operation to a cohort, e.g. visualize it, the corresponding list item can be clicked, which results in the contained participants being brushed for further investigation. Sub-populations can be compared by defining multiple brushes – one for each cohort – that can be viewed simultaneously.

The mode panel (Figure 1b) serves for choosing a mode to be used for one visualization canvas in the visualization panel (Figure 1d). In Figure 2, an exemplary configuration can be seen, where content is added to the panels shown in Figure 1. Tool tips reveal that the blue rectangle in Figure 2 corresponds to the validation mode, the green rectangle to exploration, and the orange rectangle is associated with subspace clustering. The user can simply select a desired mode for a certain visualization canvas by dragging the mode from the mode panel into the respective canvas area. If there is only one view associated with the chosen mode, the corresponding view is displayed. However, especially for the exploration mode there are various views provided for analysis.

There are examples, where the visualization type is dynamically chosen based on the data type of variable(s) selected to be visualized, e.g. in the web framework for analysis of lumbar back pain [KOJL+14]. In contrast to that, we display a drop-list, from which a view can be selected to display the current set of participants. Modes in a canvas can also be replaced by simply dropping another mode into the same area.

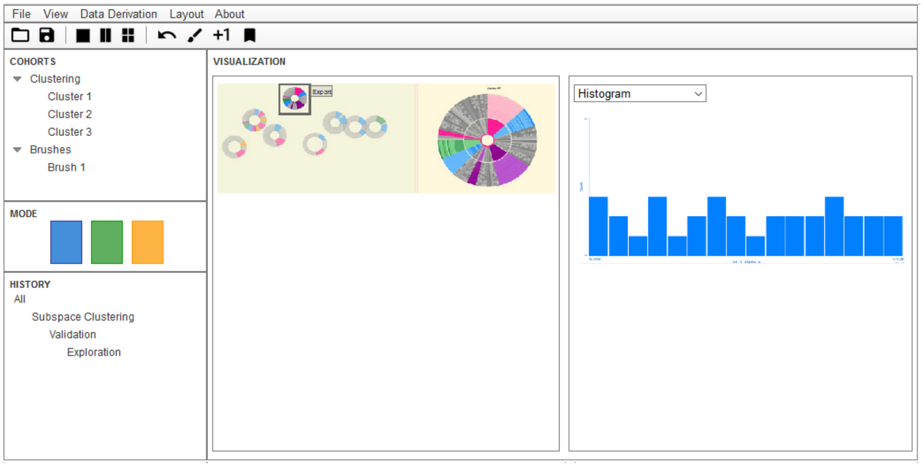

Figure 2: *Exemplary configuration of the proposed system. Subspace clustering (left) and exploration mode (right) are considered simultaneously. A histogram was chosen from the drop-list to visualize a variable's distribution (right).*

Performed actions can be viewed again by taking a look at the history panel (Figure 1c). Every time a cohort is defined or an algorithm is executed, for example, an item is added to the tree widget. This item can also capture parameters that were used for the executed algorithm, which is very important for reproducibility as described in 2.3. The entire tree widget, as can be seen in Figure 2, shows how the analysis process evolved.

A data view, which shows the variables' exact values for the currently brushed participants in a data table, can be opened at any time during analysis for detailed investigation of interesting participants.

With the provided functionality and a large number of supported visualization and data mining techniques, the presented prototype would highly contribute to the analysis of population study data and would enable the definition of sub-populations as well as their comparison. Epidemiologists can either follow their well-known workflows by using one analysis mode and the single canvas or they can perform a thorough explorative analysis using the possibility of flexibly choosing multiple views at a time and switching between modes and views.

3 User Study – Meeting with Domain Experts from the University Medicine Greifswald

After having prototyped an initial visual analysis system, it is essential to deepen our understanding of the functionality that is needed and applicable for epidemiological analysis. To get an idea of the system's practical relevance and to get a better impression of the epidemiologists' work, we arranged a visit to domain experts. From 7 to 8 July 2016 we traveled to Greifswald to meet Henry Völzke (HV) and Till Ittermann (TI) from the Institute for Community Medicine at the University Medicine Greifswald. HV has worked as a specialist for internal medicine for eight years before he turned towards the domain of epidemiology, where he is the leader of the SHIP in Germany. TI has studied statistics and performs traditional epidemiology, where he is mainly concerned with association analysis. Prior to the meeting, we compiled a list of questions concerning the current state in epidemiology, their current analysis process, and the evaluation of proposed functionality, which we sent to both of them in advance. The complete list of questions can be found in Apppendix B.

On the first day of the visit, we brought each other up to date concerning the work that was done on both sides since the last collaboration. This included (1) an overview of what has been done in the field of visual analytics of cohort study data, (2) a short presentation of the current state of subspace clustering and (3) a presentation of the prototype presented in this report. In return, HV and TI informed us about the recent development regarding population studies. Afterwards, we discussed the importance of central approaches in epidemiology, such as identifying correlations and replication of findings, and how visualization and data mining could be integrated to support the epidemiologist's work. In this context, HV and TI also gave their feedback regarding the presented prototype. It will be described in detail in Section 3.1. Upon our request, they also told us about their common workflow, which primarily consists of three steps: (1) generation of sub-populations, (2) restriction of sub-populations, and (3) replication of sub-populations. These steps will be discussed in Section 3.2.

On the second day, TI performed an exemplary analysis session in our presence. He first showed us an example of thyroid dysfunction and then explained the purposes of individual analysis steps. Furthermore, he also told us about confounding and (non-linear) regression modeling. A more detailed description of the analysis session can be found in Section 3.4.

3.1 Feedback on the Prototype

Unfortunately, we did not get very much feedback on the presented prototype. This could also be interpreted as a sign, in such a way that the system was not interesting enough for epidemiological work to be discussed in much detail. Both of them did not seem to see the benefit of simultaneously considering different data representations for epidemiological analysis. TI immediately asked if a text editor could be added to specify parameters in a script, rather then defining them via a visual interface. He likes scripts, because – in addition to being used for execution of programs – they can serve as log files at the same time, storing the input parameters that led to certain results. In this context, the issue of reproducibility also comes in again. It is essential that analysis results are reproducible, in the sense that the analysis yields the same results when being performed with the same values again. On the other hand, it would be beneficial to be able to re-run the same analysis on another data set or sub-population, to see how the results differ or can be reproduced. With a complex system like the proposed one, it is rather difficult to keep track of the individual steps that were taken to obtain a result, so that reproducibility is limited. Efforts can be made to preserve information about the steps followed in the course of an analysis.

In contrast to the overall idea, decision trees and subspace clustering as individual techniques were rather approved. Decision trees were rated as interesting for finding combinations of variables that predict a target disease. In this case, the variables do not necessarily have to be causally correlated, but can be predictors in certain combinations. Analyzing sub-populations is very important for population studies. Subspace clustering is interesting for generating sub-populations for further analysis. Here, visualization is primarily needed for displaying the data mining results. The approach presented by Shiva Alemzadeh (which should have been integrated into the proposed system) was lively discussed and HV and TI suggested some adjustments concerning groups of variables and the color scheme. TI also remembered Uli Niemann's Interactive Medical Miner [NSVK14] to be related to the subspace clustering and suggested to somehow combine both approaches. As clustering and data mining are still an emerging field in epidemiology and not very well accepted by epidemiologists so far, a fixed procedure is to be preferred when working with such techniques.

Our overall impression is that they do not need a complex visual analytics system to support their workflow; instead, for their purposes, they are perfectly satisfied with their statistical methods. For supporting hypothesis generation, which was the overall aim of the proposed system, they rather prefer subspace clustering as a single tool to then process the results with their own methods.

3.2 Workflow

HV and TI provided a more detailed description of how they approach a statistical assessment of interesting associations between (combinations of) variables and a chosen target. This analysis is guided by hypotheses that were derived by physicians in their daily routine. The workflow that is discussed in this section corresponds to step 4 of the epidemiological workflow mentioned in Section 2.1.

Prior to the actual steps, there is an optional centralization and standardization of the involved variables with respect to the value range. Centralization aims at shifting variable values in such a way that zero divides the range of values in half. This is achieved by subtracting the mean value of all variable values. In addition, standardization scales the data to be in the range from minus one to one. For some applications, it might be reasonable to dichotomize the variables.

3.2.1 Generation of Sub-Populations

As a first step of statistical analysis, a number of sub-populations that are potentially relevant for an investigated outcome is generated. Relevant for the outcome means that a sub-population satisfies a certain minimum support of at least 5% or 10% of the population to ensure statistical significance. However, if the absolute number of subjects is too small, significance is not given. In general, an overview of the relative sizes of generated sub-populations is essential. Epidemiologists are in particular interested in sub-populations that differ strongly from the overall mean with respect to a specific association. As an example, one could identify a sub-population with participants that exhibit a stronger association between Thyroid Stimulating Hormone (TSH) and Blood Pressure (BP) than the overall population. Epidemiologists should be able to assess this deviation from the general mean, for example in the form of a deviation plot or integrated into a visualization of the sub-populations. As the sub-populations are in any case restricted in the next step, one can be liberal regarding the generation and selection of sub-populations in this step to not miss interesting associations.

3.2.2 Restriction of Sub-Populations

The aim of this step is to identify and select the most relevant sub-populations for further analysis, depending on some interestingness measure. Selected sub-populations should deviate among each other as much as possible to cover large portions of the observation space. When using a clustering algorithm, e.g. subspace clustering, for automatic identification of sub-populations, this can be realized by merging clusters that are very similar, i.e. where large parts of the subspaces (dimensions that were used to produce the clusters) overlap. In general, clusters relying on a low number of dimensions should be preferred to avoid overfitting. Which sub-populations can be used for further processing also depends on which observed variables are also available in other population studies. This is an important issue for an external replication of the sub-populations as described in the following step.

3.2.3 Replication of Sub-Populations

The last step is about finding a previously selected sub-population in another population; preferably in an independent population, i.e. from a different study. If we consider the Thyroid dysfunction example again, we would search the other population for participants with high deviation from the overall mean regarding the association between TSH and BP. This is called replication and gives us an impression of the generalizability of the underlying association (e.g. between TSH and BP). A question that was left open is what it exactly means for the association to find a sub-population again in a different population.

Replication of sub-populations is only possible, if data from multiple population studies are available. In our case, SHIP-TREND could be used, which would rather be an internal replication. The replication results would be even more convincing, if subgroups identified in SHIP could be found again in populations from the Rotterdam study or another external study.

Now, how do we know that we have found a sub-population again? In general, there are two approaches to replication: (1) compute regression measures (e.g. the absolute risks) for the same sub-population, i.e. a group of participants with very similar characteristics, in different populations and see how the measures differ; (2) apply the algorithm, that was used to find the first sub-population, to other populations without changing parameters and assess the similarity between the sub-populations. As the first approach highly relies on a great overlap in the observed attributes of the different population studies, the second approach seems to be more feasible to us. In general, one must find a way to deal with shifted variable ranges from one population study to another.

For a comparison of sub-populations, we discussed several measures to evaluate the similarity of two sub-populations:

- Size – sub-populations are similar if they contain a similar portion of the overall cohort.
- Deviation from the global mean – if the sub-populations deviate in the same way from the overall mean of a cohort with respect to a specific association, they can be assumed to be similar.
- Involved variables – sub-populations are considered comparable if they are characterized by (almost) the same variables, e.g. participants in both sub-populations exhibit a high level of TSH and normal BP.
- Similar distribution regarding involved variables – if the distributions of involved variables are similar with respect to, for example, interquartile range, sub-populations can be considered to be similar.

As a consequence of the last measure, sub-populations may be visualized using box plots for the contributing variables to assess the similarity of distributions.

In addition to replication in other populations, it would be interesting if a sub-population can be found at a different point in time for the same population. For SHIP, this is possible, because it is a longitudinal study as described in Section 1.1. This would also avoid the issue of data security, that arises as soon as data from various population studies needs to be compared. However, to cope with this issue when performing external replication, one could hand over identified sub-populations to epidemiologists working on other studies and ask them to see what they can find in their data that relates to this sub-population.

3.2.4 Publication of Results

The process of publication starts with an initial literature review, which is later developed into the scientific background of a paper. The scientific context and already known correlations play an important role in this phase. Afterwards, the concrete study and methods used for analysis have to be described, followed by an explanation of what was investigated and what was found. The analysis results are then discussed in teams of two to three people, which usually consist of the first and last author of the paper and additionally a statistician. A meaningful part of this discussion is to establish a connection from the analysis results to the scientific background described in the introduction of the paper. If needed, the introduction is refined according to the obtained findings. Prior to the submission, the final draft is once again discussed together with the remaining authors.

3.3 Validation of Findings

In general, there are two main objectives in the context of epidemiological analysis: (1) hypothesis-based analysis of assumed associations between risk factors and a target by identifying sub-populations with increased risk, and (2) identification of unknown risk factors, i.e. explorative analysis. To both analysis types applies: an independent validation of the findings is essential and needs to be carried out carefully. The replication concept presented in Section 3.2.3 is one way of validating analysis results.

The importance of validation is evidenced by considering the selection bias that longitudinal studies often suffer from. Such an information bias is caused when invited subjects do not take part in the examinations or drop out after some time. This is an even more serious problem, as we can assume that the non-responders significantly differ from the responders and that certain diseases, e.g. depression, might be under-represented in a study, because affected people tend to not respond or to even give wrong answers. As a consequence, every finding has to be treated with caution.

Assumptions and findings are generally put in the context of scientific literature. There is no guarantee for the causal correctness of identified correlations, but identified correlations can be substantiated if similar issues are reported in literature or the findings could be replicated in an independent cohort. Once they identified associations in the data, epidemiologists also search for known clinical endpoints in literature that are related to these findings. Such endpoints, recent discussions in scientific literature, and insights that were already gained using small groups of subjects can also be used to choose a direction for further analysis. In this context, explorative analysis opens up a new field, as it does not rely on a starting hypothesis and thus does not necessarily originate from existing discussions.

If the scientific literature does not provide enough material that relates to the insights gained from analysis, additional studies must be consulted, which leads back to the replication concept (Section 3.2.3). For explorative analysis using data mining and visual analysis techniques, causal relations need to be identified by carefully checking the results with respect to confounding bias. It also needs to be considered if identified associations play a role in clinical practice at all. Validation then might be carried out by means of a subsequent randomized control trial, which is a prospective study based on newly acquired data, where neither the examiner nor the proband has an influence on the risk exposure.

3.4 Exemplary Analysis Session

On the second day of our visit in Greifswald, Till Ittermann told us about the basic procedures in epidemiology and performed an exemplary analysis in our presence, while explaining some steps in detail. He uses STATA[5] for statistical analysis, a software package that provides a lot of functionality for statistical data analysis and graphics.

Confounding For populations studies like the SHIP, epidemiologists often apply a case-control design, where the results may be confounded. To obtain reliable insights, the list of considered epidemiological features needs to be adjusted to confounders. Directed acyclic graphs are used to display causal interference, where the assumed risk factor(s) are on the left, the outcome on the right, and possible confounding variables like age, sex, and smoking status are in between. The graph is then used to determine actual confounders that need to be adjusted to. If there are multiple possible sets of variables to adjust to, the variables that are measured most accurately are chosen.

Regression In epidemiology, linear correlations are rather exceptions. Many correlations that were assumed to be linear turned out to be U- or J-shaped, for example risk related to blood pressure, because low blood pressure is also a risk. However, approximately linear correlations are possible, as patients with extreme characteristics may not be contained in a selection.

Various linear and non-linear regression models were discussed at our visit. *Linear regression* predicts the mean of a variable Y given the value of a variable X. If the variables are not normally distributed, a linear regression model might not work out. In this case, *median regression* could be used instead. *Poisson regression* is commonly used to model the relative risk. For rare events, logistic and Poisson regression are rather similar, while they differ strongly when considering more frequent events.

[5] www.stata.com

When a non-linear regression model is needed, different approaches are worth considering: categorization of the exposure variable, adding a quadratic term, fractional polynomials, or restricted cubic splines. From these approaches, TI prefers *fractional polynomials*. This method tries different transformations of the exposure variable A, e.g. $log(A)$, A^2, $\sqrt{A}$, or $\frac{1}{A}$, and chooses the transformation that minimizes the deviance of the regression model. This regression can also be realized using STATA, which performs the fractional modelling for all available variables automatically. The results of non-linear regression can be presented in the form of a data table that depicts a comparison with a reference point. As an example, the odds ratios for a Body Mass Index (BMI) of {20, 22.5, 25, 27.5, 30} can be compared to a reference BMI value of 25 for being a smoker. Clinicians like such data tables for presentation.

Visualizations that were used Throughout the analysis session, TI used various basic visualization techniques to plot intermediate results. *Bar charts* are generally used to graphically represent the distribution of categorical data. In the showcase analysis, they were used to compare the prevalences of a disease depending on age and gender between two studies, SHIP-0 and SHIP-TREND.

Box plots are considered when descriptive statistics of a continuous variable are of interest. The box itself depicts the lower and upper quartile of the variable's distribution, while a line inside the box represents the mean. Minimum and maximum values are usually denoted by lines that go upward and downward from the box. Outliers might also be included, e.g. depicted as circles.

Another standard technique commonly considered are *scatter plots*, which depict bivariate relationships between numerical variables. In our example, a scatter plot was used to show the relationship between age and TSH. Regression lines may also be fitted to the cloud of data points in a scatter plot to illustrate the relationship. TI showed a scatter plot where, instead of a regression line, the reference intervals for age and TSH were plotted in the form of fit-curves for lower and upper percentiles (e.g. 2.5^{th} and 97.5^{th} percentile).

As became clear during the analysis session, interaction analysis is important. In statistics, interactions may arise when the relationship of three or more variables is observed. Let us consider the relationship of three variables. If there is an interaction between two of the variables, their simultaneous influence on the third variable is not additive. This means that the relationship between one of the interacting variables and the third variable depends on the value of the other interacting variable. In an *interaction plot*, the relationship between an interacting variable and the third variable is depicted using lines. Depending on the value of the other interacting variable, these lines can look different. Uncertainty regions can be added to such a line chart, where the uncertainty may be represented by percentiles, for example.

Data tables in their original form were used a lot to show and compare concrete values, e.g. for regression measures.

3.5 Summary

The meetings with Henry Völzke and Till Ittermann in Greifswald were highly informative and we gained a lot of knowledge about epidemiological procedures. The most important insight to be obtained is that visual analytics cannot aim at supporting the epidemiologists in their overall workflow. For the hypothesis-based statistical analysis and validation of identified associations, they acquired a sophisticated workflow involving visualization and analysis tools that they are familiar with. As a conclusion, a coordinated multiple views system for interactive visual exploration of their data, as we originally had in mind, would not be of use for them.

Nevertheless, the definition of sub-populations is one of the most important data mining tasks and would be helpful to find out more about groups of participants with certain characteristics. Thus, we should rather focus on supporting hypothesis generation by providing data mining and visual analytics techniques for generation and visualization of potentially interesting sub-populations. In this context, a replication of sub-populations in independent studies is highly important and should be supported by the provided techniques. In contrast to large sections of the epidemiology community, HV and TI are quite open towards new approaches from the visual analytics and data mining domain, although there might be some work to do in order to adapt such techniques to the reproducibility and verification requirements of the epidemiology domain. Besides subspace clustering for generating sub-populations, HV and TI also consier decision trees to be interesting for identifying (combinations of) variables that predict a certain outcome (e.g. a target disease).

Independently from the concretely realized technique, logs that store performed actions and corresponding parameters are highly important. They allow for reproducibility and support the communication of the course of events as well as (intermediate) analysis results.

Visualizations, in particular overview visualizations, are primarily needed for data mining results, e.g. individual sub-populations with their characteristics and deviation from the overall mean. For presentation of analysis results, the epidemiologists in Greifswald primarily use data tables or basic statistical visualizations. Here, more advanced visualizations might help to establish a deeper understanding and interpretation of the presented issues. Narrative techniques could be provided for people who are not an expert in statistics, thus being suitable for dissemination of important results to the broad masses, for example. This also refers to prevention, where people could visually explore what-if scenarios (e.g. how is my risk reduced if I quit smoking?) all by themselves to get a better feeling for the conveyed information. Other applications of visualizations in the epidemiology domain might include quality control and teaching.

4 Consequences for the Design of a Visual Analytics Support

After having discussed the benefits and restrictions of guided and explorative analysis approaches in Section 2.3, we can say that a guided analysis is better suited for the field of epidemiology. In this application domain, reproducibility and controlled validation are more important than flexibility.

In their paper about requirements analysis and usability engineering for epidemiology, Thew and colleagues state that "[...] providing a fully configurable system would be a wasted effort if users can't find time to tailor it" [TSP+09]. Our meetings with Henry Völzke and Till Ittermann revealed similar insights. Originally, we had a fully flexible visual analysis system in mind, which incorporates various techniques and allows for easy extension due to a plug-in layout. We imagined it to offer great potential for a thorough support of hypothesis generation based on data exploration, which is very little performed up to now. However, it turned out that such an approach is not adequate for the epidemiologists' purposes and the benefits of coordinated multiple views for their work cannot be revealed in a short time. This might also be due to the fact that epidemiologists do not have the time to overcome a long learning phase for a completely new and complex system [TSP+09].

When discussing the proposed prototype and overall epidemiological procedures with the domain experts, we obtained further requirements that are worth considering when designing a visual analytics support. Systems designed for epidemiological applications should always include some kind of history log, where performed analysis operations and corresponding parameters are preserved for later re-use or communication. Developers should also think about the role of a text editor accompanying the visual interface. Our two domain experts were concerned about being able to look at the course of performed analysis operations in a textual form. They even prefer specifying parameters in a text editor rather than using visual components. One could think about a combination of both text editor and visual interface, where visual inputs can be converted into textual form for later re-use and communication. Furthermore, the converted inputs can also be used to update parameters or specify additional ones in a textual form. This would enable epidemiologists to choose between visual and textual input – depending on what is preferred and most suitable for the given task. Such a combination could also include a log at the same time.

In general, systems need to provide ways of reproducing results and re-visiting intermediate analysis states. This includes the possibility to flexibly exchange the underlying data set for replication as described in Section 3.2.3. We should also not forget about the issue of protecting personal and medical data. Analysis tools must be designed in a way that prohibits unauthorized access to the data.

We also talked about the role of visualizations in epidemiology. As could be observed during the exemplary analysis session (Section 3.4), basic visualizations such as histogram, scatter plot, or interaction plot are considered for statistical analysis, which we should probably leave as such. The use of more advanced visualizations is primarily beneficial for hypothesis generation, e.g. the definition of potentially interesting sub-populations. Overview visualizations and visualizating data mining results can provide valuable insights into relationships that might not be identified otherwise. A more detailed explanation of visualizations for epidemiological analysis can be found in Paragraph 3.4.

As an overall conclusion, we should focus on providing visual analytics and data mining to support the epidemiologists in individual, concrete issues, rather than aiming at solving their complete analysis workflow.

References

[BBK13] Ruth Bonita, Robert Beaglehole, and Tord Kjellström. *Einführung in die Epidemiologie.* Huber, 2013.

[BOH11] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[FFF12] Robert H. Fletcher, Suzanne W. Fletcher, and Grant S. Fletcher. *Clinical Epidemiology: the Essentials.* Lippincott Williams & Wilkins, 2012.

[Kei01] Daniel A. Keim. Visual Exploration of Large Data Sets. *Communications of the ACM*, 44(8):38–44, 2001.

[KGL$^+$15] Paul Klemm, Sylvia Glaßer, Kai Lawonn, Marko Rak, Henry Völzke, Katrin Hegenscheid, and Bernhard Preim. Interactive Visual Analysis of Lumbar Back Pain. 2015.

[Kle16] Paul Klemm. *Interactive Visual Analysis of Population Study Data.* PhD thesis, Otto-von-Guericke University Magdeburg, 2016.

[KLG$^+$16] Paul Klemm, Kai Lawonn, Sylvia Glaßer, Uli Niemann, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. 3D Regression Heat Map Analysis of Population Study Data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):81–90, 2016.

[KLR$^+$13] Paul Klemm, Kai Lawonn, Marko Rak, Bernhard Preim, Klaus D. Tönnies, Katrin Hegenscheid, Henry Völzke, and Steffen Oeltze. Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data. In *VMV*, pages 121–128, 2013.

[KOJL$^+$14] Paul Klemm, Steffen Oeltze-Jafra, Kai Lawonn, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1673–1682, 2014.

[NSVK14] Uli Niemann, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Interactive Medical Miner: Interactively Exploring Subpopulations in Epidemiological Datasets. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 460–463. Springer, 2014.

[PKH$^+$16] Bernhard Preim, Paul Klemm, Helwig Hauser, Katrin Hegenscheid, Steffen Oeltze, Klaus Toennies, and Henry Völzke. Visual Analytics of Image-Centric Cohort Studies in Epidemiology. In *Visualization in Medicine and Life Sciences III.* Springer, 2016.

[Rob07] Jonathan C. Roberts. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In *Fifth International Confer-*

ence on Coordinated and Multiple Views in Exploratory Visualization, pages 61–71. IEEE, 2007.

[TSP$^+$09] Sarah Thew, Alistair Sutcliffe, Rob Procter, Oscar De Bruijn, John McNaught, Colin C. Venters, and Iain Buchan. Requirements Engineering for e-Science: Experiences in Epidemiology. *IEEE Software*, 26(1):80, 2009.

[VAS$^+$11] Henry Völzke, Dietrich Alte, Carsten Oliver Schmidt, Dörte Radke, Roberto Lorbeer, Nele Friedrich, Nicole Aumann, Katharina Lau, Michael Piontek, Gabriele Born, et al. Cohort Profile: the Study of Health in Pomerania. *International Journal of Epidemiology*, 40(2):294–307, 2011.

[vdEvW13] Stef van den Elzen and Jarke J van Wijk. Small Multiples, Large Singles: A New Approach for Visual Data Exploration. In *Computer Graphics Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.

[ZGP14] Zhiyuan Zhang, David Gotz, and Adam Perer. Iterative Cohort Analysis and Exploration. *Information Visualization*, page 1473871614526077, 2014.

Appendix A Analysis Tasks

The following tasks address the main research questions that epidemiologists aim to answer by analyzing cohort study data.

- **A1: Characterization** – identify characteristics of a variable's distribution in a population. Determine prevalence and incidence.
- **A2: Association-causation** – identify which variables correlate with a target variable and how strongly. Identify risk factors that are associated with a specific disease. Determine risk groups.
- **A3: Sensitivity analysis** – how do changes in risk factors affect the outbreak of a disease? How do varying parameter settings change the statistical results? How do changes in cohorts affect test scores?
- **A4: Temporal changes** – how do characteristics evolve over time?
- **A5: Cohort comparison** – compare one cohort to another with respect to a target disease or over time and space.

During an analysis session, epidemiologists explore several alternatives; if assumptions proved to be wrong or to obtain comprehensive insights into the given data. For this purpose, they need to independently adjust investigated data subsets to their needs.

- **B1: Variable selection** – identify and select suitable dimensions to perform a statistical analysis.
- **B2: Cohort definition** – determine relevant subpopulations for analysis.
- **B3: Cohort modification** – refine or expand cohorts based on recent findings, e.g. via filtering.

Finally, there are tasks concerned with interactive visual analysis features on a high level of abstraction.

- **C1: Assess data quality** – are data suitable for the investigated problem? Are there missing data? Was data acquisition inaccurate?
- **C2: Overview** – familiarize with the given data set. Get a first notion of potential hypotheses and analysis workflows.
- **C3: Details-on-demand** – view details, such as confidence or p-value, for a selected association.
- **C4: Scalability** – analyze thousands (up to 20000) of subjects.
- **C5: Flexibility** – constantly change the perspective, e.g. between different techniques or overview and in-depth analysis.
- **C6: History** – keep track of performed analysis operations.

Appendix B User Study – Questions

B.1 Current Analysis Process

1) In what context does an analysis take place? What happens before and after?
2) How does an explorative analysis proceed?
3) Is the analysis rather straightforward? Which steps are processed?
4) What are you looking for, if no starting hypothesis is available?
5) Do you search for concrete variables to work with at the beginning?
6) Which components do you use in a typical analysis workflow?
7) What kind of visualizations do you currently use?
8) Are there problems at any point of the analysis process?
9) What happens to missing data? Do you treat them in a special way?
10) Do you use data mining approaches in any way?
11) How do you select the participants to be investigated? Via filtering? Or do you always work on the entire data set?
12) What do you use medical image data for?
13) When you use regression, how many independent variables are chosen to form combinations for a regression model?
14) In which range is the computation time for correlation measures of a complete data set located?
15) Do you sometimes think: ”To better understand this issue, it would really help to see the distribution of this variable at the same time.”?

B.2 Other Questions Concerning the Current State

1) How much time is averagely available for working with analysis software?
2) Which circumstances have to be in place for a successful epidemiological analysis?
3) How do you handle the statistical relevance of identified associations? What happens if the relevance is given/not given?
4) How is the quality of the underlying data assessed?
5) How do you recognize erroneously caused associations? Is it possible at all to recognize them?
6) How are analysis results and intermediate states communicated?

7) Scalability: how many participants are approximately contained in a cohort or a data set to be investigated?

8) Are there color schemes that you prefer because you are used to them?

B.3 Proposed Functionality

1) Is it beneficial for you to explore the data or would you like to get suggestions for potential regions of interest from the start?

2) If no starting hypothesis is available, to what extent is an initial visualization useful? What kind of visualization would be useful?

3) Would you benefit from a direct comparison of two or more cohorts regarding one or more variables?

4) Do you need regression models beyond linear or quadratic regression?

5) Would it be beneficial to consider other views or techniques parallel to the regression analysis?

6) To what extent to you trust data mining techniques that propose certain risk factors, for example? What requirements must be satisfied to trust such algorithms?

7) Are you interested in the changes of an outcome as risk factors or statistical parameters are varied?