Visual Analytics for Epidemiological Cohort Studies

Bernhard Preim[1], Shiva Alemzadeh[1,4], Till Ittermann[2], Paul Klemm[1], Uli Niemann[3], and Myra Spiliopoulou[3]

[1]University of Magdeburg, Department of Simulation and Graphics
[2]University of Greifswald, Department for Community Medicine
[3]University of Magdeburg, Institute of Technical and Business Information Systems
[4]VR Vis, Center for Virtual Reality and Visualization, Vienna

1. Introduction

Medical visualization research typically aims at clinical medicine, e.g. improved support for diagnosis and treatment planning. Our work, in contrast, aims at the public health sector, where *prevention* of diseases is the major task. An important branch of public health is epidemiology, an interdisciplinary area that involves medical and statistical expertise which aims at reliable statements w.r.t. the frequency of diseases and other health indicators as well as risk factors for developing a disease.

To create new epidemiological knowledge, large and comprehensive data are acquired in cross-sectional or cohort studies to cover life style aspects, sociodemographic factors, blood, urine and other samples as well as questions related to medical problems in the past. Recent studies include medical image data, such as MRI and ultrasound, or genetic samples. Examples for ongoing cohort studies are the Rotterdam study, the SHiP (Study of Health in Pommerania), the UK Biobank and the German National cohort (see [PKH*16] for an overview of these studies). The overall amount of information per participant comprises a few thousand variables. Therefore, next to traditional hypothesis-driven research, epidemiology can benefit from the advances in data mining and visual analytics for the bulk analysis of correlated high-dimensional data and for the identification of vulnerable subpopulations.

In a long-term cooperation between the Faculty of Computer Science in Magdeburg and the Institute for Community Medicine of the University of Greifswald we aim at an improved analysis of large cohort study data using the SHiP as an example. SHiP encompasses two cohorts: SHiP (aka SHiP-core) started with 4.308 participants in 1997 (SHiP-0) with follow-up investigations every 5 years and is currently at its 4th wave SHiP-3 (1.700 participants, 2014-2016). The second cohort SHiP-Trend started wih 4.420 participants in parallel with wave SHiP-2 and with the same protocol. These population-based studies allow the investigation of diseases, such as diabetes and back pain and disorders, such as fatty liver, that are known to be precursors of diseases. We aimed at visual analytics solutions to support the major tasks that we derived in a number of workshop-style discussions.

- Integrated analysis of shape-related variables derived from medical image data and abstract data, e.g. related to the health problems and drugs,
- Identification of strong correlations between life style-related variables and a disorder, such as fatty liver or back pain,
- Identification of subpopulations that differ strongly from the entire population w.r.t. their risk for a health problem, and
- Analysis of quality problems, in particular missing values.

In the following we describe the solutions developed for supporting these tasks. All these solutions are web-based prototypes, realized with D3, that enable the cooperative analysis of the data.

2. Integrated Analysis of Shape-Related Variables and Abstract Data

We used the SHiP data to analyze the shape of the lumbar spinal canal to investigate whether it is associated with lower back pain — a hypothesis of epidemiologists [KLR*13]. After the lumbar spine was extracted and transformed to a 3D surface model, its centerline was generated as a representative for the shape of the lumbar spine model. Afterwards, these centerlines were grouped into clusters with an agglomerative hierarchical clustering that was previously used to cluster streamlines from bloodflow data. To visualize these clusters, a ribbon-based visualization was designed (Fig. 1).

Later, we extended this framework to a web-based exploration tool [KOL*14]. Based on an analytical workflow, our framework enables the generation of hypotheses and its subsequent statistical analysis. All variables in the cohort are listed, the expert can drag and drop certain variables onto the main canvas, which leads to a representation of the mean lumbar spine model of the patients in the selected group. Additional refinement or selection of new variables results in a visualization showing correlations. Brushing and linking options support the analysis. We employed a subset of the SHiP (2.240 participants) with 134 variables from which 21 are metric and 113 categorical. In addition, we derived 9 variables from the spinal canal centerline. No participants were excluded (to avoid selection bias), but the number of variables was strongly restricted for data protection reasons.

Results. Our comprehensive analysis of various shape variables, e.g. mean shape and mean torsion, with respect to back pain did not yield any significant result for the entire population, for males

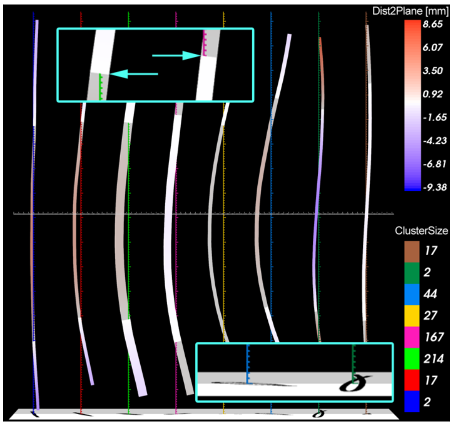

Figure 1: *Clustered centerlines of the spinal canal are displayed. Cluster size encoded by width of the ribbon and color encodes distance to the midsaggital plane. Selection of a cluster leads to the display of related image data (From: [KLR*13]).*

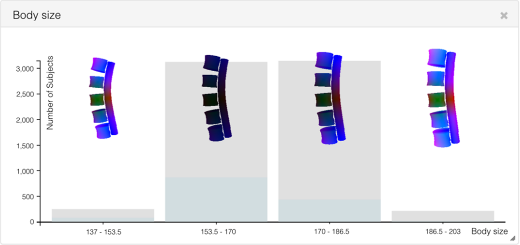

Figure 2: *A shape-variance visualization is enhanced with information visualization techniques. The bars indicate the sizes of subgroups defined by body size. The two background colors discriminate participants with and without back pain within the last three months (From: [KLR*13]).*

or females and for large age groups. A more fine-grained analysis that involves individual vertebrae and the angle between them and the spinal canal *may* lead to an association. Instead, we found a number of correlations with other data: Increased body fat, body weight, blood pressure, alcohol consumption, the presence of attentiveness disorder and a large amount of sleep are associated with back pain. For the epidemiologists, the strength of some of these correlations was surprising and may serve as starting point for testing new hypotheses. Many aspects have to be considered to derive valid conclusions [KOL*14]. As an example, it is often necessary to categorize scalar values. Equally-sized bins is a simple strategy but not appropriate when outliers are present. Quantiles of the distribution are the better choice for binning.

3. Identification of Strong Correlations Between Variables and a Disorder

Klemm et al. [KLG*16] presented the 3D regression heat map to analyze correlations between variables. The idea of this approach is to let the experts input simple regression formulas, e.g., $Cancer \sim X+Y$ to explore the correlations. This calculates all combinations of pairwise variables for a correlation of cancer by using the R^2 metric. For the visualization a heat map was employed. In case the experts type $Z \sim X+Y$, a regression cube was generated showing for every slice a 2D heat map of correlations. This approach is computationally intensive. Even with parallel computing for a dataset with 100 variables and 2000 participants 14 hours are needed to compute $Z \sim X+Y$. Nonetheless, the approach was considered helpful to find out non-obvious correlations.

Results. This system was employed to analyze various subsets of the SHiP data. One example was a dataset compiled to investigate factors that contribute to an increased breast density—a known risk factor for breast cancer. Thus, the epidemiologist steered the 3D regression heat map generation with the formula a $ParenchymaPercentage \sim X+Y+Z$. Strong correlations are emphasized with saturated colors and perceived as hotspots. Strong correlations were observed for glandular tissue density and parenchyma segmentation metrics. Also, strong correlations were observed with diabetes which confirmed previous knowledge. A surprising finding was the strong correlation with kidney disorder (correlation around 0.9). Further analysis revealed that only 8 participants exhibit this disorder. Thus, the sample size is too small for a valid conclusion.

4. Identification of Subpopulations with Increased Risk

Subspace clustering is frequently used for analyzing high-dimensional data where global clustering is not promising due to the curse of dimensionality. For public health data subspace clustering is beneficial, since persons are likely to be similar in some attributes but not in all. Subspace clustering is typically a two-stage process: clusterable subspaces are analyzed in the first stage and a clustering method is applied to these clusterable subspaces. We developed a visual exploration workflow for subspace clustering results for the SHiP data [AHN*17]. As an overview, subspace clusters are displayed in a 2D view where multi-dimensional scaling was applied to map the similarity between the clusters to the spatial proximity (Fig. 3). Similarity for subspace clusters relates to the overlap between the dimensions and the overlap of instances of subspaces. For selected subspace clusters, details are presented in additional views (Fig. 4). The colors used in these displays represent different categories of the data that were carefully discussed with the epidemiologists. They suggested the following categories: laboratory values, medication, physical status and habits. To explore a selected subspace in detail, scatterplot matrices are also available. For supporting an overview, also scaled bar charts were employed. They reveal for a health risk, such as high blood pressure, the portion in the different clusters. Thus, it becomes obvious when a cluster (representing a subpopulation) exhibits a strongly increased risk compared to the global mean (see Fig. 4). These and other design decisions were based on discussions with epidemiologists w.r.t. good overview visualizations.

Subspace clusters may have arbitrary shapes. For epidemiologists, however, hyperrectangular clusters are beneficial, such that a subpopulation may be described by rules, e.g. "While in the study population only 18 % exhibit goiter, in the subpopulation described by Body mass index > $30.5kg/m^2$, Thyroid Stimulating Hormone > 1.5mU/l goiter occurs in 52 %". We therefore support the transformation from arbitrary shapes to hyperrectangular clusters such that

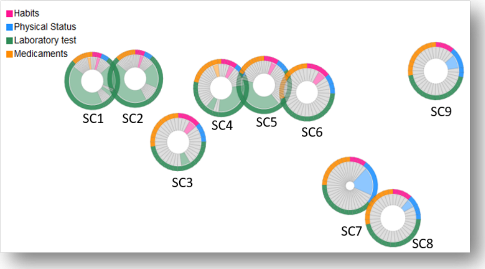

Figure 3: *The results of subspace clustering applied to cohort study data are shown in an overview based on a similarity metric that evaluates the overlap of instances and dimensions. Each subspace cluster is shown as a donut where donuts with a larger hole represent clusters with few members only. Gray values represent dimensions that do not contribute to this subspace cluster. The detail view (right) reveals information on the participants. Darker colors represent greater values and black represents missing values (From: [AHN*17]).*

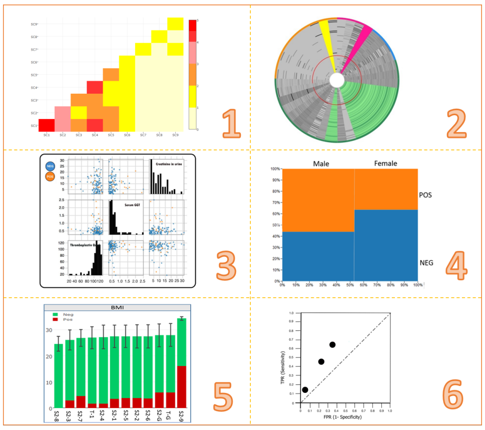

Figure 4: *A number of views are combined for an in-depth analysis of the disorder fatty liver among SHiP participants. A matrix view (1) shows the dimension overlap between any pairs of subspace clusters. A donut chart displays the currently selected cluster (2). A scatterplot matrix reveals the distribution of pairwise variables where orange represents participants with an outcome, such as fatty liver (3). A mosaic plot (4) juxtaposes the relative frequency of participants with a positive outcome between male and female participants. The barchart view (5) shows the portion of participants with a positive outcome (red). One cluster stands out with a particularly high number of participants. Error bars represent the statistical uncertainty (From: [AHN*17]).*

users draw in rectangles in scatterplots showing the dependencies between two variables [AHN*17]. Since epidemiologists aim at a verification of findings from any data mining technique, the authors also supported the validation of the subpopulations in an independent cohort, e.g. subpopulations determined in the SHiP data are validated by means of the SHiP trend cohort.

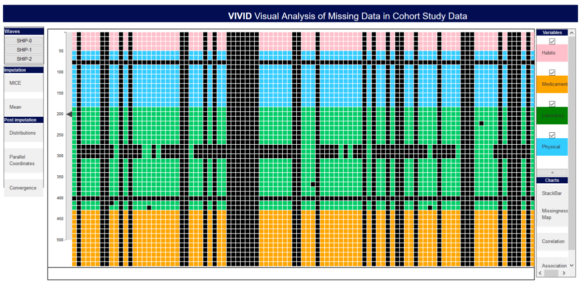

Figure 5: *Overview of missing data from an epidemiologic study: Rows represent variables and columns show participants. Missing values are indicated in black. Completely black columns represent drop-out participants (From: [ANI*17]).*

Results. We focused on fatty liver as a widespread disorder. The information of the liver status was extracted from radiologists. Participants with a liver fat concentration of > 10% are considered as positive for fatty liver. Our analysis revealed a subpopulation with a 50% portion with a fatty liver. This subpopulation comprises 6.7% of the overall population, it is older than the entire population (average age 59 years), has a high body mass index (34,2 on average) and also a high number of diabetes patients (18%). For replication, it was analyzed whether these findings also apply to the SHiP-Trend data. Here, the corresponding subpopulation is interactively constructed and compared with the original one, with respect to the distribution of laboratory values. Indeed, the SHiP-Trend subpopulation exhibits very similar values for variables, such as creatinine and serum GGT. Also the relative size of these subpopulations is similar. Thus, the identified SHiP subpopulation with strongly increased risk for fatty liver is likely to be valid.

5. Visual Analytics of Missingness

Missingness may occur in one cycle of a longitudinal study, i.e., the values for one participant are not complete, and between cycles where participants do not show up in a later stage. *Missingness maps* [HK*11] give an overview that may serve to identify patterns (see Fig. 5). There are various strategies to cope with missingness.

- *Complete case anaylsis*, where only complete datasets are analyzed,
- *Single imputation*, where missing values are replaced with a median or average value, and
- *Multiple imputation*, where dependencies between variables are considered and multiple replacements are computed.

The first two strategies are straightforward to realize but not appropriate for typical PH data. If only complete cases are considered, the number of cases often shrinks drastically. The resulting subset is no longer representative, if the missingness is not completely at random, i.e. the likelihood of a missing value for one variable depends on the value of other variables. Single imputation would preserve all datasets and thus the representative character, however the median or average is often not a good guess for the missing value. If all missing values are replaced with the (same) average value, the distribution changes such that the variability is reduced.

Multiple imputation is appropriate when the missingness is not completely at random. Multiple imputation is based on a regression analysis: for a variable v_1 that is affected by missingness, the (linear) correlation to all other variables is computed. Only if the correlation value for a variable v_i is high (e.g. one of the N highest values or above a threshold), it is used to *predict* the missing value. Thus, the computation of the *relevant* predictors is accelerated. Imputation is performed several times, leading also to an estimate of the uncertainty involved. This strategy is known as *multiple imputation with chained equations* (MICE) and is available in R that is frequently used in epidemiology.

The number of iterations (default value is 5 in MICE) strongly influences the computational effort. We showed the converging behavior of the imputation and confirm that five iterations is a good choice [ANI*17]. We used MICE for the analysis of the SHiP data w.r.t. hepatic steatosis [ANI*17]. Visual analytics plays an essential role in the handling of missing data. Because in case of missing at random the distributions of missing and observed values are usually different, it is necessary to check the distributions of observed values and imputed values to validate the imputation. An extended version of the workshop paper was invited by the Computer Graphics Forum and is "Probably accepted" [AVPea19].

6. Related Work

So far, visual analytics solutions for public health were focused on more urgent problems, primarily on the detection of outbreaks of infectious disease and the response management, e.g. [MRHea10]. These applications have different requirements, in particular the spatio-temporal distribution of cases is essential and thus (dynamic) map displays are used. Powerful cohort construction for patient data were developed by Krause et al. [KPS16]. Our work differs, since epidemiological data has not such a strong temporal component, e.g. the flexible temporal query interface is not needed. We refined subspace clustering visualization techniques developed by Tatu et al. [TMF*12] and Assent et al. [AKMS07].

7. Summary and Outlook

We developed a number of visual analytics solutions to leverage the potential of large cohort study data in epidemiology. Our design considers also aspects of data quality and trust in the generated results, e.g. a validation component for subpopulations. While the solution that we fully realized is restricted to linear regression, we also investigated the automatic search for quadratic regression between pairs of variables with a low linear regression. The longitudinal character of the SHiP and other cohort data is not adequately analyzed so far. It would be interesting to analyze and display the development of subpopulations. As a first step in this direction, in a student project data from different time points can be loaded and are analyzed with an exhaustive search for association rules. The resulting subpopulations can be displayed over time (see Fig. 6).

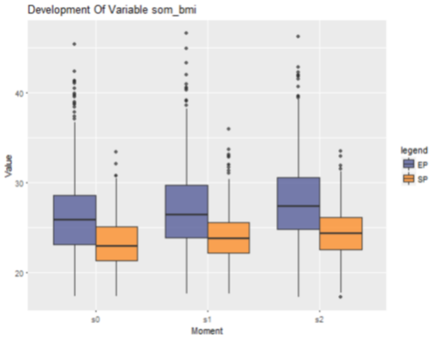

Figure 6: *A subpopulation has an increased body mass index over time. However, this is typical since the entire (ageing) population had the same development (From: [May18]).*

References

[AHN*17] Alemzadeh S., Hielscher T., Niemann U., Cibulski L., Ittermann T., Völzke H., Spiliopoulou M., Preim B.: Subpopulation Discovery and Validation in Epidemiological Data. In *Proc. of EuroVis Workshop on Visual Analytics* (2017), pp. 43–47. 2, 3

[AKMS07] Assent I., Krieger R., Müller E., Seidl T.: VISA: visual subspace clustering analysis. *SIGKDD Explorations 9*, 2 (2007), 5–12. 4

[ANI*17] Alemzadeh S., Niemann U., Ittermann T., Völzke H., Schneider D., Spiliopoulou M., Preim B.: Visual Analytics of Missing Data in Epidemiological Cohort Studies. In *Proc. of EG Workshop on VCBM* (2017), pp. 43–51. 3, 4

[AVPea19] Alemzadeh S., Völzke H., Preim B., et al.: Visual Analysis of Missing Values in Longitudinal Cohort Study Data. *Computer Graphics Forum* (2019). 4

[HK*11] Honaker J., King G., et al.: Amelia ii: A program for missing data. *Journal of statistical software 45*, 7 (2011), 1–47. 3

[KLG*16] Klemm P., Lawonn K., Glasser S., Niemann U., Hegenscheid K., Völzke H., Preim B.: 3D Regression Heat Map Analysis of Population Study Data. *IEEE Trans. Vis. Comput. Graph. 22*, 1 (2016), 81–90. 2

[KLR*13] Klemm P., Lawonn K., Rak M., Preim B., Tönnies K. D., Hegenscheid K., Völzke H., Oeltze S.: Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data. In *Proc. of Vision, Modeling, and Visualization* (2013), pp. 121–128. 1, 2

[KOL*14] Klemm P., Oeltze-Jafra S., Lawonn K., Hegenscheid K., Völzke H., Preim B.: Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 1673–1682. 1, 2

[KPS16] Krause J., Perer A., Stavropoulos H.: Supporting Iterative Cohort Construction with Visual Temporal Queries. *IEEE Trans. Vis. Comput. Graph. 22*, 1 (2016), 91–100. 4

[May18] Mayer B.: *Visual Analytics of Participant Evolution in Longitudinal Cohort Study Data* . Tech. rep., University of Magdeburg, Faculty for Computer Science, 03 2018. 4

[MRHea10] Maciejewski R., Rudolph S., Hafen R., et al.: A Visual Analytics Approach to Understanding Spatiotemporal Hotspots. *IEEE Trans. Vis. Comput. Graph. 16*, 2 (2010), 205–220. 4

[PKH*16] Preim B., Klemm P., Hauser H., Hegenscheid K., Oeltze S., Toennies K., Völzke H.: Visual analytics of image-centric cohort studies in epidemiology. In *Visualization in Medicine and Life Sciences III*. Springer, 2016, pp. 221–248. 1

[TMF*12] Tatu A., Maass F., Färber I., Bertini E., Schreck T., Seidl T., Keim D. A.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proc. of IEEE Visual Analytics Science and Technology, VAST* (2012), pp. 63–72. 4