Interactive Visual Analysis of Image-Centric Cohort Study Data

–blind–

Abstract—Epidemiological population studies impose information about a set of subjects (a *cohort*) to characterize disease-specific risk factors. Cohort studies comprise heterogenous variables (*features*) describing the medical condition as well as demographic and lifestyle factors. The data is analyzed using a priori defined hypotheses to find statistically significant correlations between variables (*interactions*). Modern cohort studies also incorporate medical image data. The statistically driven epidemiological workflow only allows to determine *interactions* between image-derived metrics such as distances extracted from landmarks of the segmentation model.

We propose an Interactive Visual Analysis approach that enables epidemiologists to examine both image-based as well as non-image data, e.g., sociodemographic variables and attributes derived from the image data. This is achieved by combining brushing and linking enabled coordinated information visualization views and interactive 3D shape renderings with epidemiological data representations such as pivot tables and key figures as association measures. The presented concepts are applied expert-guided to gather and evaluate hypotheses about the aging process of the lumbar spine. It shows to be a more flexible comparison between image and non-image data. The new framework allows for hypothesis validation and hypothesis generation by incorporating human pattern recognition as well as data mining methods. Using all reliable information from the image segmentation linked to non-image variables aims to unveil *interactions* by applying an iterative analysis approach.

Index Terms—Interactive Visual Analysis, Epidemiology, Spine

1 Introduction

Epidemiology aims to characterize health and disease by determining risk factors. Clinical problems, such as the selection of diagnostic tools and efficient treatment, are tackled using results of epidemiological research. Also, the introduction of preventive measures in medicine and beyond, are based on epidemiological research, where, for example, subgroups with increased risk are identified [10]. On the other hand, observations made by clinicians in the daily routine are translated into hypothesis for epidemiological research. These are used to determine environmental and lifestyle factors as well as medical attributes which may influence a condition of interest. The data variables necessary are gathered using structured interviews and clinical examinations. Statistical methods like regression analysis aim to check the attribute list for plausibility.

Longitudinal population-based studies, such as the Study of Health in Pomerania [37], aim to gather as much information as possible about a defined sample of people (a *cohort*). The sample is drawn randomized to avoid selection [10]. When for example one investigates risk factors for prostate cancer in male subjects, the outcome is strongly dependent on the age. Therefore results need to be age-adjusted to be comparable. Confounding variables, are often not obvious at all and characterizing them is already an epidemiological result.

Modern cohort studies often include medical image data. Since it is unethical to expose people to radiation, non-harming imaging such as Magnetic Resonance Imaging (MRI) or Ultrasound Imaging is used. To quantify these data it is necessary to label each voxel regarding structure affiliation (*segmentation*). Manual segmentation carried out by radiological experts is possible but very costly and prone to inter- and intra observer variability. Segmentation algorithms allow for (semi)-automated analysis of the data but require sophisticated methods due to high inter-subject variability caused by the subject diversity. Analyzing spatial data with respect to other epidemiological factors requires techniques which reach beyond standard statistical methods.

We propose an Interactive Visual Analysis approach [33] to provide a way to analyze both image- and non-image data. Visual queries and direct feedback of Visual Analytics systems allow for a fast exploration of the data space. Intended as an extension to the well established epidemiological tools it provides a way to rapidly validate hypothesis as well as trigger *hypothesis generation* using Data Mining methods such as clustering. In order to characterize the healthy aging process we aim to determine changes for subjects which indicate unusual pathological changes.

Our contributions are:

- Applying the Interactive Visual Analysis technique set to the epidemiological problem domain by characterizing requirements of this context.
- Provide an overview over the workflow for analyzing cohort study data to gain insight into the large diverse subject spaces.
- Provide visualization techniques which combine both information visualization and 3D rendering of organ shapes as well as combining them with well known epidemiological graphics and key figures.
- Implement the presented methods as a web framework based on WebGL, D3JS and NodeJS.

2 Medical and Technical Background

In this section we want to give insight into the epidemiological workflow when analyzing cohort study data to identify the problems we address in this paper.

2.1 Epidemiological Workflow

Many different experts work at epidemiological studies, ranging from specialized doctors to medical computer scientists with focus on biometrics and statisticians. Epidemiologists follow a workflow mainly driven by statistic tools to validate hypothesis about disease specific risk factors. Following Thew and colleagues, the workflow can be characterized as follows [32].

- Hypothesis most commonly base on observations made by clinicians in their daily routine.

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

- A set of attributes depicting conditions affected by the hypothesis is compiled accordingly.
- Confounding variables need to be adjusted so that they do not affect the effect size of a attribute.
- Statistical methods such as regression analysis are applied to measure the effect size of attributes to the outcome of interest.

The workflow is shown in Figure 4 (a).

Reproducibility of results is an key requirement in epidemiology. Longitudinal studies require the acquired attributes to be comparable to evaluate them. If the data acquisition process changes, a information bias is introduced to the data, hampering inference in acquisition cycles. To determine, whether a subject is prone to be affected by a certain disease, *relative risks* are expressed through the evaluation of p-values which indicate statistical significance. Statistical correlations are prone to *confounding*, meaning that two factors are dependent and therefore should be normalized with respect to each other. Statistics tools such as SPSS and STATA play a major role for analyzing epidemiological data. Graphic data representation is largely used to present results rather than gaining insight.

Grouping subjects using epidemiological factors is essential in order to allow per-group risk determination. Grouping is carried out hypothesis driven. Age for example is also divided into groups (e.g. in 20 year-steps) when investigating its influence on a condition. These groups depend strongly on the condition of interest and therefore there is no defined standard on how to categorize these values.

2.2 Epidemiological Data

Epidemiological data is highly heterogenous and incomplete. Information about medical history and examinations, genetic conditions, geographical data, questionnaire results and image data yields a complex data space for each subject. Often data are derived from acquired variables to either group or threshold values or get information derived from reviewed data such as breast density data for women. This underlines also the problem of missing data since for ethical, legal or medical reasons some data variables can not be gathered for each subject. Follow-up examinations or -questions for conditions also produce variables only available for a small amount of subjects.

Indicators for medical conditions as well as questions about a subjects' lifestyle are also often *dichotomous*–they have two manifestations (often *Yes* or *No*). This allows for the calculation of *odds ratios* which describe the relation of two *dichotomous* variables, allowing for direct comparison of their influences. Dichotomous data can also be derived by combining aggregating data variables to yield only two manifestations (e.g. subjects younger or older than 50).

Image acquisition. Imaging techniques emitting hazardous amounts of radiation for the subject are not suited for ethical reasons. MRI data is more expensive to obtain than CT data but does not affect the subjects health and is therefore the main method for collecting cohort study imaging data. The image quality is a tradeoff between accuracy and affordability [26]. This often yields image resolutions inferior to those of clinical day-to-day practice, which makes their analysis more challenging.

Image analysis. Decisions have to be made on how image data are *compared* and *quantified*. Segmentation masks labeling the voxel of a shape of interest would be ideal since many different key figures can be derived from them. Since these masks require sophisticated algorithms custom tailored to the data sets the epidemiologists are forced to measure the data by hand, which is a very tedious work with respect to the number of necessary landmarks and number of subjects. Information derived by landmarks are also not nearly as expressive and versatile as segmentation masks. They are also prone to a high inter-observer variability and hard to reproduce. This gains even more momentum when analyzing multiple time steps! Morphometric information from landmarks comprises thickness, diameter or length of a structure as well as grey-value distribution in an area (used for determine type of tissue).

2.3 The Study of Health in Pomerania (SHIP)

Starting 1997 with a cohort consisting of 4.308 subjects this cohort study located in northern Germany aims to characterize health and disease in the widest range possible [37]. Data is collected without focus on a specific disease. This allows the data set to be queried regarding many different diseases and conditions. Subjects were examined in a 5-year time span, continuously adding new parameters including MRI scans in the last iteration of 2012. The MRI protocol features a rich number of different sequences. Also for women, breast MRI scans were acquired. A second cohort `SHIP-Trend` was established in 2008 to acquire data about a younger population. The protocols for analyzing the subjects between `SHIP` and `SHIP-Trend` remained the same, making them comparable. The overall examination time for each person attending the study is two days.

3 Prior and Related Work

Einfuehren von Helwigs Terminologie?

Designing a visualization which communicates all aspects of the data equally is challenging. Given the number of features of epidemiological data sets and their different manifestations, it is often a good solution to combine the strength of different visualization techniques in a unified system [3, 22]. The Principal Component Analysis (PCA) and similar techniques are able to reduce the dimension by extracting most expressive components, but make the influence of each variable hard to determine.

The work of Turkay and colleagues is closest to ours albeit our focus on processing medical image data and variables with categorical manifestations [35]. Investigating Data on an norwegian aging study their methods aim to amplify a hypothesis generation process. Statistical measures of continuous variables such as mean, standard deviation, skewness, or inter-quartile range are used to create *dimension plots*. These transform dimensions into data points and make them comparable with respect to the derived descriptive measures. This not only allows for comparing all continuous variables in a single plot but make their distribution change comprehensible. This requires a good descriptive measure which captures the kind of change the user is interested in or which reflects unexpected data behavior. The technique was applied to variables generated by segmenting the brain into 45 parts and measure the voxel number, volume and properties of the intensity values. The method is strongly dependent on the descriptive measures of the epidemiological factors. Hypothesis based on observations of changes in these plots may impose over-fitting to the data because the measure highlights only subsets of statistical changes. Our approach sticks more to the information extracted from the segmented image data and derive variable interaction with non-image epidemiological factors.

Gresh and colleagues proposed `WEAVE`, one of the first systems which analyzed medical image and non-image data using linked views [14]. Blaas and colleagues presented a similar system which analyzed medical image data and variables derived from them using views from the feature- and physical space [2]. This approach already incorporated Data Mining approaches such as dividing the data space using a k-nearest-neighbor technique and Principal Component Analysis. Steenwijk and colleagues employ a relational database to organize the data to visualize subject data using linked views such as parallel coordinates, scatterplots and time plots [31]. Zhang and colleagues provide a web-based system for analyzing subject groups with linked views and batch-processing capabilities for categorizing new subject entries into the data set [39]. Their understanding of a cohort differs from the understanding of the term in an epidemiolgical context.

Commercial systems such as `Tableau` or `Spotfire` provide a rich user interface that allows to apply Visual Analytics techniques without the need of writing any code. With little effort, linked views can be created using these tools, but the data processing possibilities such as derivation of new variables or the volume rendering capabilities are very limited. These systems share limitations in extensibility to a specific problem domain.

Klemm and colleagues used lumbar spine variabilities based on an semi-automatic shape-detection algorithm of 490 participants of the SHIP-2 [20]. Hierarchical agglomerative clustering divided the population into shape-related groups. As proof of concept a relation between size of the segmented shape and measured size of the subjects was measured. This work focuses on incorporating these derived data as new features of the overall data set, making it possible to include it into the hypothesis validation and generation process. When applying clustering techniques on the non-image data it was found that k-Prototypes and DBSCAN is appropriate in the epidemiological context but is strongly dependent on the chosen variables and distance measure [19].

Generalized Pairs Plots (GPLOMS) are an information visualization technique that allows for heterogenous variables to be pairwise compared using appropriate plots in a plot matrix grouped by type [18]. This technique is also useful to gain an overview over numerous variables and their distributions. It uses histograms, bar charts, scatterplots and heat maps to visualize the different variable combinations with regard to their type. Brushing, linking and filtering are feasible with *GPLOMS*, but have limitations such as making only one category brushable at a time. A similar approach was taken by Dai and colleagues for risk factor exploration as they also incorporate choropleth maps of epidemiological factors (e.g. mortality rates in a region) with parallel coordinates, bar charts and scatterplots with integrated regression lines [8]. From their findings regarding the interaction of cancer-related socio-demographic factors are drawn in a *Concept Map* where related factors are connected via graph-edges.

Chui and colleagues visualized interactions in time-dependant epidemiological data using time-series plots highlighting risk factors differences in age and gender [7].

Comparing tissue between many subjects in an epidemiological context requires methods which allow for shape variance visualizations. Caban and colleagues investigated the suitability of variance visualizations of shape distribution models and concluded in their user study that users favor spherical glyph representations over deformation grids and likelihood volumes [5]. The distribution of shapes in a space derived from a PCA is plotted by Busking and Colleagues in a 2D-projected plane of the space [4]. We incorporate the idea of combining 3D-Shape rendering with information visualization techniques. Differences between structures are highlighted using color mapping of the difference to the mean shape, but is rather hard to recognize due to small renderings of each subject in the shape-space. Via mesh morphing interpolated views can be created by the user in a separate view as well as comparisons in a contour view. The data needs to abstracted or summarized in order to work in this context. In order to detect local deformation changes, Hermann and colleagues investigating shape related difference by letting the user specify a deformation of interest and showing corresponding changes in the shape using covariance tensors [17]. This method allowed for rapid hypothesis validation and was able to reproduce textbook knowledge. By plotting p-values in ventricle surfaces, Chou and colleagues were able to map disease-associated values directly on a 3D tissue representations [6]. This requires a geographic colocation of associated features.

The strength of the Interactive Visual Analysis approach described in the next section is its versatility with respect to the application field [22]. Oeltze and colleagues combined a linked view representation of results from a statistical analysis with feature localizations of the human blood flow with the goal of its evaluation [24]. While we take similar steps when analyzing the data such as employing statistical tests, our data is mostly independent from the medical image data and is not describing it–except the variables derived the segmentation model itself.

4 Interactive Visual Analysis in Cohort Study Data

- Acceptance plays a big role - you have to pick up the user by presenting information they are used to in a new way

As described in subsection 2.1, the epidemiolgical workflow is a

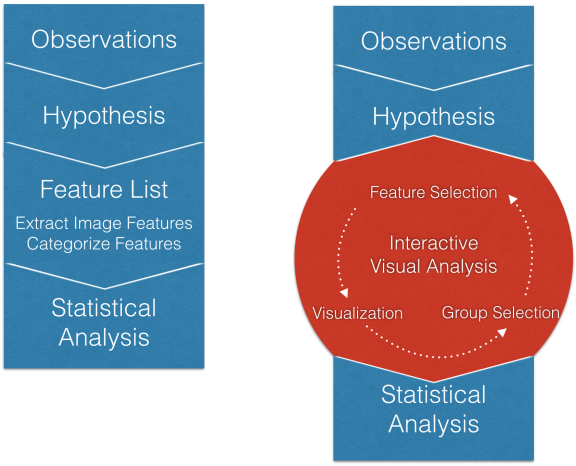

Fig. 1. ToDo Abbildung noch nicht final. Visual Analytics systems are able to complement parts of the epidemiological workflow, not replace it. The appropriate combination of statistical- and interactive driven analysis shows promising potential to unveil the information in the data. (a) shows the standard epidemiological workflow, (b) the IVA supported one. The iterative red highlighted part is called the *IVA Loop*.

sequence of steps taken by domain experts and needs to be reproducible and comprise statistical integrity. Figure 4 (a) describes this workflow as consecutive series of steps. The workflow we propose by introducing the *IVA* principle into the epidemiological application domain does not aim to replace the existing workflow but to complement its weaknesses. In the current state the workflow treats the data like a black box. A list of features describing the hypothesis is compiled and analyzed using statistical tests. The resulting value decides whether the data supports the hypothesis or not. It would be possible that there are actually features of the data set which support the hypothesis by discriminating the population in the expected way, but with this approach they are not highlighted in any way. This becomes even more important when the workflow is adapted to the analysis of the medical image data. Domain experts would have to annotate landmarks tediously to derive metrics such as distances. Not only does this leave out the majority of the information in the medical image data by abstracting it to single values, it is easily possible that information left out would heavily influence the result. Considering more complex parts of the data would make those results more trustworthy and also could identify possible anatomical confounders–an epidemiolgical research result in itself. Statistical tests check for validity of the number but not for their completeness or plausibility!

IVA tries to illuminate the black box by making the domain experts part of the feature list selection process. Figure 4 (b) highlights the iterative process as part of the epidemiolgical workflow. Note that it also aims to project back into the hypothesis formulation step to amplify hypothesis generation This has to be handled with care since overfitting of expectations to the available data is an imminent danger as described by Turkay and colleagues [35].

4.1 Image Centric Cohort Study Data in Interactive Visual Analytics Context

In the *IVA* context, data is divided into two major view types. The human body exposing shape information for the *physical view* [25]. This information space is usually displayed via volume rendering techniques [24]. These variables are also referred to as *independent variables*. *Dependent variables* in the epidemiological context can be divided twofold:

- Variables derived from the image data. These measures abstracts shape information as quantification to allow for compari-

son. These variables describe image data and can also be used to brush in the image space.

- Epidemiological socio-demographic or medical attribute data. These values belong to every subject which is represented in the image space, but does not describe shape information. This is the data epidemiologists usually want to correlate with image data.

4.2 Data Preprocessing

To include heterogenous epidemiological data in an *IVA*-framework it is necessary to process it to obtain standardized views to the available features. Due to the different acquisition modalities there have to be different techniques incorporated. Data obtained using questionnaires or medical tests are often stored using statistical packages such as `SPSS` or `Stata` which have a proprietary data format with limited export capabilities. The best solution for us was to simply export the data in the respective tool to a character separated text file and then convert it to data types which are easier manageable such as JSON or XML using our own classes. In order to verify that the conversion worked as expected and the data is valid, it is good practice to use data wrangling tools such as `OpenRefine` to validate the data (find missing data, clean up bad formatting). Exporting the data dictionary, which stores information about each manifestation of a feature is also an important step to get a detailed description of data variables and the meaning and unit of measurement of their values. The reasons for missing data have a wide range from ethical to medical and personal issues. Therefore, these are also included as error codes which have to marked as such in the data dictionary.

Processing the image data associated to each subject consists for the most part of information extraction about a structure of interest. This is either done manually by experts setting landmarks (sometimes supported by algorithms connecting the land marks such as graph cuts [13]) or by a (semi)-automatic detection, registration and segmentation. Algorithms, applied to the data, do not only have to deal with a large inter-subject variability of the structure of interest but also need to be reproducible [26]. Model-based approaches have shown in principle to be effective for segmentation [11, 12] and detection [27]. If a segmentation yields only binary masks separating the structures, algorithms such as Growing and Adaptive Shapes can be applied creating a surface grid where each point is comparable throughout the population [9]. Intensity-based comparisons can be achieved using rigid image registration, but model based results however are preferable [21]. Comparison based on grey values is usually carried out to to measuring the quantity of fat, water, and–application specific–iron content (liver) or distribution of grey and white brain tissue.

Morphometric variables are derived to allow for statistical comparison of the tissue which incorporate mostly position, volume and relative distances and alignment to other structures.

4.3 IVA Patterns

The explorative procedures when analyzing data using *IVA* can be divided into three different patterns, handling interaction between domain and range variables.

4.3.1 Local Investigation

This pattern projects information from image space to the range perspective. As opposing to other *IVA* application domains, this step is more complicated in the epidemiological context. Shape information can not be brushed by incorporating ROI-selections but rather has to employ techniques that specify local deformation changes [17] or subjects that belong to a shape class. Methods available for *feature selection* strongly depend on the type of registration that was applied to extract the tissue of interest. Model-based segmentations or masks yield data structures capable of calculating mean shapes and distances between individuals or subject groups. Feature selection is also possible by applying clustering algorithms in order to get shape-groups [20]. These algorithms can be used to investigate interactions between shape-groups and other non-image based variables. Another application is the outlier analysis. Outliers can indicate segmentation errors or an outstanding group of individuals who may share a pathology.

4.3.2 Feature Localization

As described before, the vast majority of data points are considered to be dependent with respect to the image domain in the *IVA* context. Selecting subjects based on image derived data can be seen as additional possibility of shape-related grouping. The epidemiologist is primarily interested in the shape of subjects within a range of a set of variables that describe the current hypothesis. Epidemiologists are used to categorize data into groups that fit their hypothesis formulation. Continuous variables such as age are for example often divided in categories like young, aged and elderly. Categorization is strongly dependent on the hypothesis and therefore requires suitable brushing techniques as described in Section 5.3.

4.3.3 Multivariate Analysis

Introduced in the information visualization, the multivariate analysis incorporates brushing and linking of of views displaying non-image parameter. The need of statistic measures which describe how variables correlate with each other given the selected groups is special for the application domain. These associations also summarized using Pivot Tables which are popular in epidemiology and which are described in the following section.

5 Interaction- and Visualization Techniques

We employ the fourth and highest level which includes next to brushing and linking, advanced brushing, extraction of attributes the development of visualizations custom tailored to the data sets. *IVA* lev-

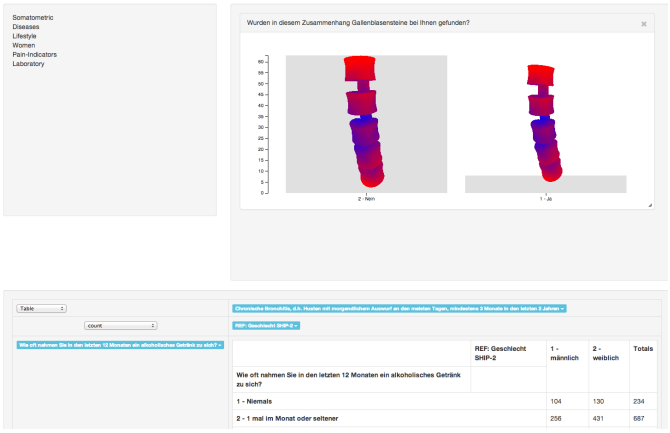

Fig. 2. *Image is not final* Screenshot from the front-end which is divided as follows: (a) Sidebar which contains all features as well as the groups defined in the analysis process; (b) Canvas area where features can be added via drag and drop and the visualization chosen automatically according to the data type; (c) Interactive Pivot Table which shows exact numbers for each displayed variable combination. The data displayed is used to analyze the lumbar spine.

els define different levels interaction. The employed techniques needs to respect the epidemiological requirements. Hypothesis generation bears the chance of over-fitting the data to expectations. To avoid this, a timeline needs to be introduced which keeps track how many variable variations the user evaluated before coming to an conclusion. Since epidemiologists are used to process groups based on table representations we decided to introduce an interactive solution in form of a Pivot Table.

5.1 Structure and Workflow

We divide the workspace into three major parts as seen in Figure 2. TODO auf Fig 5.1 zurckkommen The sidebar holds all available vari-

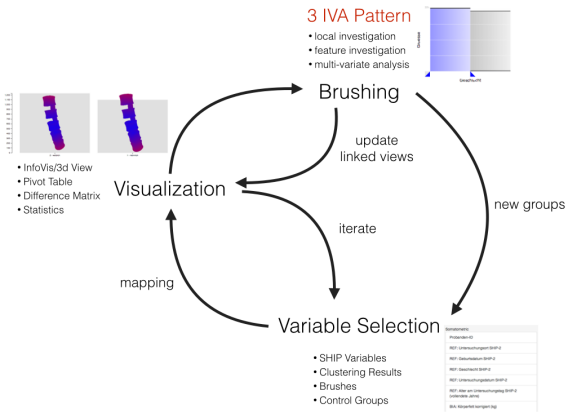

Fig. 3. Detailed Version of the *IVA Loop*. Usually starting with a selection of Variable of interest (user-driven or via data mining techniques), the data is mapped using a visualization technique appropriate for the selected data types. The data is visualized in the independent and dependent domain space, which can then be brushed, yielding new groups which can be investigated using further variables. Note that adjacent steps are directly connected via feedback loops allowing for iterative refinement and give as much freedom to the user as possible.

ables as well as groups either derived by user input or automatic clustering.

From the sidebar, elements of interest can be put into the canvas area via drag and drop. Doing so will automatically create an information visualization suitable for the current data type. 3D shape information about the investigated structure is displayed for each variable manifestation. By dropping variables on existing visualizations the system creates a visualization that allows for comparison (e.g. mosaic plots for ordinal variables). Elements can be brushed in the information visualizations and are linked to the other representations in the canvas. Shape based clustering can either be applied to all subjects or subgroups.

All elements in the canvas view are also represented in the interactive Pivot Table which gives detailed information how the subjects are distributed given the displayed variables. Details to the different views are presented in the following sub-sections.

ToDo
- wie werden Confounder gefunden?
- welche statistischen Kennzahlen werden eingebunden?

5.2 Sidebar

An overview of all variables is presented in a sidebar where they are categorized to different types such as somatometric, disease- or lifestyle related, pain indicators and laboratory data. It also contains subject groups either defined by user brushing or by automated clustering. Groups are treated exactly like other variables since they work exactly the same way which is dividing the subject space into different labeled categories. Bar charts show the distribution of manifestations of each variable in the sidebar.

5.3 Adaptive Feature Visualization

Inspired be the previously discussed *GPloms* [18] the visualization type is chosen dynamically based on the variable types and number of visualizations which need to be displayed. If possible the medical image data is directly included into the plot as well by including mean shapes for each manifestation (Figure 2 (b)). The 3D-plots can be navigated using standard mouse inputs and the camera is synchronized so that a direct comparison is given. If a feature is dropped on an existing plot, the visualization changes dynamically to properly make them comparable. Each plot can be brushed using widgets. It is able to duplicate brushes in order to create new groups which are evenly spread out. A use case for this is when a continuous feature has to be divided into even groups.

ToDo
- Dichotomous data
- Time-Line
- Statistical Analysis (Odds ratios)

TODO 2
- 3D Plots are small multiples - reference Tufte!

5.4 Pivot Tables

Pivot tables are frequently used to present the data in epidemiological paper. As seen in Figure 2 (c) it is a good way to display how many subjects are in each group. Pivot table quickly get confusing and cluttered when they are divided into to many subgroups. We tackled this problem by making the order and number of displayed variables adaptable. This also applies on the designation of Row or Column Variables. The mean shape for each resulting sub-group is also displayed for each subject.

5.5 Automated Feature Suggestion

TODO REWRITE USING CRAMERS V

As discussed previously, highlighting potential interesting values in the data set is one major benefit of the *IVA* powered approach. Turkay and colleagues used the approach to calculate various key figures based in the distribution functions of each feature derived from the image data [35]. Since the majority of our data are categorical features, we have to employ different solutions. A solution for this problem are odds ratios, which are a standard statistical tool for stating relations between features. Odds ratios can only be calculated for 2×2 contingency tables which usually represent the presence of a condition in a population divided by a characteristic (e.g. male/female subjects with presence or absence of back pain). To calculate odds ratios for variables with more than two manifestations, we calculate local odds ratios for each possible combination, yielding a matrix for each feature combination [28]. When the subjects are divided into groups and the calculation is carried out again for all feature combination, the difference in sum of the odds weighted with the number of manifestations per feature can be used to indicate if the feature combination yields a difference. These difference are then highlighted in a separate tab of the side bar "Interactions". Interesting interactions then can be assessed creating a linked view using the standard drag and drop workflow.

In the following sections we will discuss details on the implementation which relies on modern Web-Technologies.

ToDo
- This can be improved–summing up the values possibly not the cleverest solution–calculation of variance etc. possible
- Matrix Visualization?

5.6 Implementation

In order to provide a fast communication loop between method development and expert input, we decided to base all implementations on modern web technologies which benefits from various advantages:

- No additional software needs to be installed, most people use decent state-of-the-art web browser, even on mobile devices.
- The Client-Server structure allows it to employ heavy computation on a server machine and transfer results to the client
- Since image data for several thousand subjects claims hundreds of Gigabytes disk space it can remain safely on the server and elements can be transferred on demand. High confidentiality standards of the data can be met by restricting access via a account system

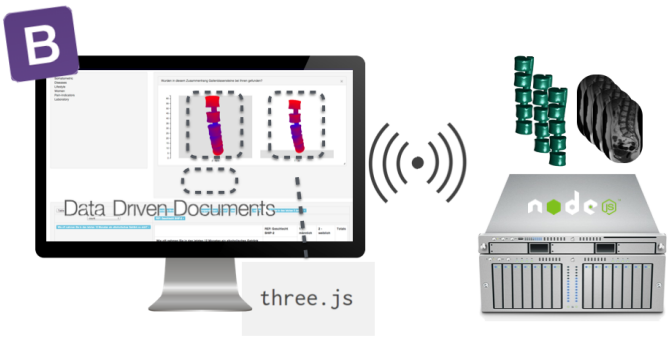

Fig. 4. *Image is not final.* The front-end solution (left) uses state of the art web technologies such as HTML5/CSS3, WebGL and SVG to display the data. The NodeJS based back-end (right) stores all image- and non-image data and transfers it to connected clients. All computation heavy operations like calculation of mean-shapes or -distances as well as statistical processing is done by the server to keep hardware requirements of client systems low. Client-Server communication is accomplished via the WebSocket protocol.

- Recent developments in WebGL applications running in browsers with near native performance push the development into to the web which results in many open source libraries which are well documented, rich in examples and driven by active communities.

These advantages do not come without drawbacks. Many methods which specialized libraries like the Visualization Toolkit (VTK) or R for statistics have build in need to be written from scratch in order to fit in the context.

The back-end is written using Node, which based on the Google V8 Javascript runtime environment. Due to its event-driven non-blocking I/O model it is fast and does not freeze in case of heavy workload like mesh calculation.

Non-image data for all subjects including the data dictionary is stored on the Server in a JSON file. Image data is available as raw DICOM files as well as segmented Meshes which can be used to compare subjects. On client connection the requested files are transferred. The server processes calculation heavy statistical tasks such as calculation of Odds Ratios or Chi Square tests for all variable combinations in order to keep the computation time on the client as low as possible.

The front-end is created using Twitter Bootstrap as foundation for the layout and basic UI elements using HTML5, CSS3 and Javascript. Information Visualizations such as Scatterplots and bar charts are created using the popular Data Driven Documents library which works well for attaching data to visible elements like vector graphics. WebGL rendering is done using the Threejs which allows GPU Accelerated data rendering. Communication between Client and Server runs through the WebSockets protocol. Since our clustering algorithms are written in MatLab we hat to access them using the Node Server. We accomplished this by converting them to parameterized standalone console applications which are spawned by node on client request and then reads the result from the console standard-out and returns it in a proper format to the client. All parameter steered console applications can be incorporated in this context.

6 Application

We applied the presented set of techniques to a data set which is compiled to analyze lower back pain. It is one of the most common reasons for an adult to see a physicians in the western civilization [36]. Epidemiological analysis of lumbar back pain such the work of Harreby and colleagues [15] is largely focused on non-image information. If at all, only a few shape related features are included in comparable studies, for example by Lang and colleagues [23]. To our knowledge, this is the first approach on analyzing shape related information of the whole lumbar spine with other epidemiological features. Determining risk factors in this area can lead to [10] (evtl. weglassen, Dopplung bei Epidemiological Background?):

- a better understanding of effects of preventive measures such as occupational health and safety regulations
- prognostic features for diagnosis and treatment of lumbar back pain
- determination of particularly effected risk groups

Characterizing the healthy aging process of the spine is a large stretch goal that allows to determine age-normalized probabilities for spine-related diseases by incorporating individual risk factors.

Data confidentiality and ethical reasons prohibit us from accessing the complete data space of the SHIP feature space. Our clinical partners compiled a feature list which is a tradeoff between complexity and limitations of the responsible ethics committee. (vllt. etwas hart formuliert).

6.1 The Lumbar Spine Dataset

We divide the data set in image and non-image data. There are 136 features describing diagnosed diseases, lifestyle factors, women specific factors, pain indicators, laboratory values, somatometric variables and are ordered accordingly. The image data was acquired on a 1.5 Tesla scanner (Magnetom Avanto; Siemens Medical Solutions, Erlangen, Germany) by four trained technicians in a standardized way. The spine protocol consisted of a sagittal T1-weighted turbo-spin-echo sequence (676 / 12 [repetition time msec / echo time msec]; 150° flip angle; 500 mm field of view; $1.1 \times 1.1 \times 4.0$ *mm* voxels) and a sagittal T2-weighted turbo-spin-echo sequence (3760 / 106 [repetition time msec / echo time msec]; 180° flip angle; 500 mm field of view; $1.1 \times 1.1 \times 4.0$ *mm* voxels) [16].

6.2 Data Preprocessing

Transformation operations on the data to prepare it for the presented prototype are denoted as data preprocessing.

Image-Data. The lumbar spine was detected in the image data using a hierarchical finite element method according to Rak and colleagues [27]. This semi-automatic method requires the user to initialize the Tetrahedron-based finite element models (FEM) with a click on the L3-vertebra. Two user defined landmarks on the top and bottom of the L3-vertebra are used to obtain an initial height estimation of the model. It uses a weighted sum of T1- and T2-weighted MRI images to detect the lumbar spine shape. The registered models capture resilient information about shape of the lumbar spine canal as well as the position of the L1-L5 vertebrae [20]. Due to TODO ANTWORT VON MARKO HIER, 983 models are obtained. For clustering purposes we extracted the centerline of the spine canal of the lumbar spine canal which captures information about lordosis and scoliosis which are medical terms for spine curvature [20].

Non-Image Data. To ensure fast and easy data access outside of statistical processors like `SPSS` or `STATA`, the data was exported to the JSON file format which can easily be parsed by modern programming languages. Each feature is stored as object which contains:

- the data as array of values–categorical values and error codes are stored using IDs
- the data type (continuous, nominal, ordinal, dichotomous)
- a detailed description of the variable
- the data dictionary which translates value- or error IDs to the actual values

Continuous variables are discretized to allow for *Cramérs' V* contingency coefficient assessment. In epidemiology, continuous data is usually categorized into ordinal groups of equal size. Since the number of categories often strongly depends on the hypothesis, the discretization steps can be adapted dynamically. To allow for hypothesis generation we set the number of groups to five if not specified otherwise.

6.3 Shape Visualization and Clustering

The tetrahedron-based model detection model described in Section 6.2 consists of corresponding grid points for each structure instance. This allows for calculation of shape-distance and similarity. This information is used to calculate mean-shapes as described in Section 5.

Shape distance is mapped onto color. For dichotomous variables, the color codes distances between mean shapes of the two groups, for variables with more than two manifestation it encodes the distance to the global mean shape of all subjects (Figure 5).

Shape based clustering is carried out via Agglomerative Hierarchical clustering of the spine canal ,centerlines which are described in Section 6.2 [20]. Since it is not possible to determine the number of clusters in a given group, it is automatically computed using a *knee/elbow* point which is described as tradeoff between number of cluster and a cluster evaluation metric [29]. For details, see [20]. The method has proven to produce comprehensible results on a preliminary data set and was able to reproduce textbook knowledge [20].

6.4 Exploratory Analysis of the Lumbar Spine Dataset

Expert guided analysis assessed the suitability of our approach for supporting both hypothesis-free analysis as well as hypothesis generation.

6.4.1 Hypothesis free Analysis

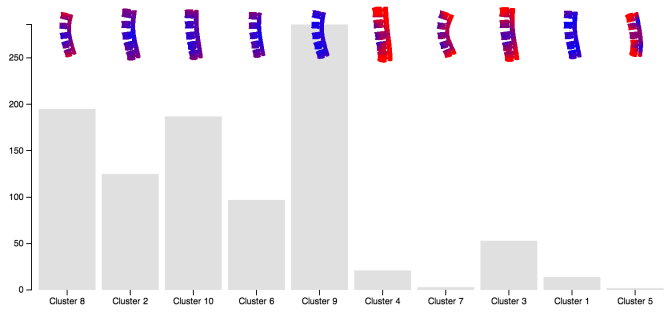

Fig. 5. *Image is not final.* (a) Clustering result of all subjects. The bar chart height indicates the number of subjects in the cluster. The difference to the mean-shape is color-coded where blue represents no difference and red a large difference.

Analyzing the data set without a prior hypothesis requires a starting point which gives a overview over the data first [30]. Shape based hypothesis free exploration starts with a shape-grouping step achieved by image-based clustering. The results for this step can be seen in Figure 5 (a).

Cluster 9 represents the subjects with average shape. Other shapes differ with respect to size, such as Cluster 2, 8, 10 and 3 where the last one also is more straight, which is usual for subjects with larger body height. Cluster 4, 7 and 5 contain outliers, characterized by their unusual shape and small number. Cluster 8 was of special interest because of its large distance to the mean shape while still exposing the second highest subject count. Looking at the *Cramérs' V* contingency values of the group reveals interactions of this group with employment status, body size, age, thyroid nodules and blood-fat value. TODO Feedback einholen von Frau Hegenscheid, TODO Implement Color coding only in a certain direction!
TODO Selection of the variables and display them using the pivot table!

While this approach does not assume any hypothesis, the data is met with a selection bias when compiling the list of related features by the domain experts. It is arguable if this first filtering step, which is purely based on expert experience, is really a disadvantage, since it rules out many parameters which may interact with the presented features but the value of this knowledge would be small [38].

6.4.2 Hypothesis based Analysis

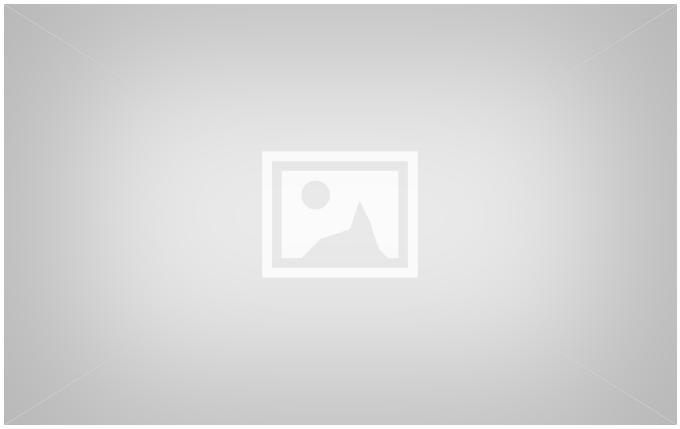

Fig. 6. *Image is not final.* (a) "Did you experience back pain in the past three month" Yes No; (b) Clustering of Yes

If the user has already a hypothesis about a relation between a non-image feature regarding shape the workflow slightly differs from the hypothesis free analysis. The starting point of the analysis is the selection of a feature of interest by dragging it into the canvas area and view the subjects distribution as well as their shape differences. In our use case, the epidemiologist was interested in the questionnaire answer "Did you experience back pain in the past three month". The mean shapes of the resulting visualization as seen in Figure 6 (a) show no difference between the two groups. Either there are no differences or the variance information was lost in the mean-shape calculation. Since the focus is on subjects which suffer from back pain, the clustering result of these subjects are then drawn into the canvas area, yielding six cluster as seen in Figure 6 (b). Cluster 5 stood out for having a so-called hyperlordosis, a strong curvature of the lumbar spine which is a indicator for back pain.

Cramérs' V contingency values highlighted relationships of this cluster with joint degeneration, meat eating habits, preoccupation, back pain, neck or shoulder pain and waist circumference. Since the prior selection only yields subjects that report back pain, the pain indicators specify the pain localization for the subjects.

It is well known that overweight is a indicator for back pain. While the Body Mass Index (BMI) is a key figure for assessing height and weight of a subject, it does not tell us anything on how the weight is distributed in the body. Our clinical partner were interested in the fact that this group presented a correlation with waist circumference. Our finding follows the recent trends that indicate that BMI is not a good measure for assessing body-shape since healthy weight is dependent on many other measures [1]. It indicates that waist circumference rather than the BMI interacts with unusual shaped spines for subjects with lumbar back pain. The influence of the parameter is now in the focus of further analysis.

6.4.3 Follow-Up Tasks and Concluding Domain-Expert Feedback

Defining a causal relationship solely based on observed correlations of two features is *cum hoc ergo propter hoc*–correlation does not automatically imply causation [34]. The observed correlations need to be carefully checked for confounder and medical soundness! Statisticians validate causal inferences of the drawn conclusions.

Features that potentially interact with a disease related condition need to be validated. To increase the probability of the observation not to be random, our clinical partners cross check for interactions in another cohort, the SHIP-TREND.

Our clinical partners pointed out the usability of the presented methods to guide their attention to features which where not in the focus of their attention and expectance. The explorative nature of the methods work well for gathering interactions which may act as confounder, as outcome of a disease or as an actual cause or risk factor. This distinction is hard to make and requires a lot of clinical experience. The combination of multiple views with shape information help to connect many different information sources to make the large information spaces cognitive feasible.

On the feature level, our clinical partners were interested in the MRI scans for the subjects in the outlier clusters because they are highly likely to exhibit pathologies. We plan to include a DICOM-viewer to meet this wish.

7 Summary and Conclusion

- Future Work: Matrix-View of differences - more intuitive
- UI more flexible by hiding UI-panes - Dynamic Discretization to better fit the data distributions - More Shape based filtering - reorder by size, curvature, etc. ... - Color code different Aspects - Only difference in x or y direction - Calculation of contingency based on the contingency differences - how does the current selection differ compared to the whole data set?

Discretization of continuos variables into bins of equal size may bias the underlying data by not considering the variable distribution function [10]. To reduce the number of false positive findings, the data space can be randomly be cut in half the hypothesis then can be cross-validated for statistical soundness. This requires a large number of subjects, especially if the investigated features are rare and are only presented by a few subjects. As suggested by Fletcher and Fletcher, we plan to validate the hypothesis using the SHIP-TREND cohort [10].

References

[1] R. S. Ahima and M. A. Lazar. The health risk of obesity–better metrics imperative. *Science*, 341(6148):856–858, 2013.

[2] J. Blaas, C. Botha, and F. Post. Interactive visualization of multi-field medical data using linked physical and feature-space views. *Proceedings of EuroVis' 07*, pages 123–130, 2007.

[3] A. Buja, J. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In *Proceedings of IEEE Visualization*, pages 156–163, 419, 1991.

[4] S. Busking, C. Botha, and F. Post. Dynamic Multi-View Exploration of Shape Spaces. *Computer Graphics Forum*, 29(3):973–982, 2010.

[5] J. J. Caban, P. Rheingans, and T. Yoo. An Evaluation of Visualization Techniques to Illustrate Statistical Deformation Models. *Computer Graphics Forum*, 30(3):821–830, 2011.

[6] Y.-Y. Chou, N. Leporé, C. Avedissian, S. K. Madsen, N. Parikshak, X. Hua, L. M. Shaw, J. Q. Trojanowski, M. W. Weiner, A. W. Toga, P. M. Thompson, and Alzheimer's Disease Neuroimaging Initiative. Mapping correlations between ventricular expansion and CSF amyloid and tau biomarkers in 240 subjects with Alzheimer's disease, mild cognitive impairment and elderly controls. *NeuroImage*, 46(2):394–410, June 2009.

[7] K. K. Chui, J. B. Wenger, S. A. Cohen, and E. N. Naumova. Visual analytics for epidemiologists: understanding the interactions between age, time, and disease with multi-panel graphs. *PloS one*, 6(2), 2011.

[8] X. Dai and M. Gahegan. Visualization based approach for exploration of health data and risk factors. In *Proc. of the International Conference on GeoComputation. University of Michigan, USA*, volume 31, 2005.

[9] L. Ferrarini, H. Olofsen, W. M. Palm, M. A. Van Buchem, J. H. Reiber, and F. Admiraal-Behloul. Games: growing and adaptive meshes for fully automatic shape modeling and analysis. *Medical image analysis*, 11(3):302–314, 2007.

[10] R. H. Fletcher, S. W. Fletcher, and G. S. Fletcher. *Clinical epidemiology: the essentials*. Lippincott Williams & Wilkins, 2012.

[11] O. Gloger, J. Kühn, A. Stanski, H. Völzke, and R. Puls. A fully automatic three-step liver segmentation method on lda-based probability maps for multiple contrast mr images. *Magnetic Resonance Imaging*, 28(6):882–897, 2010.

[12] O. Gloger, K. D. Tönnies, V. Liebscher, B. Kugelmann, R. Laqua, and H. Völzke. Prior shape level set segmentation on multistep generated probability maps of mr datasets for fully automatic kidney parenchyma volumetry. *IEEE Transactions on Medical Imaging*, 31(2):312–325, 2012.

[13] D. Greig, B. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–279, 1989.

[14] D. L. Gresh, B. E. Rogowitz, R. L. Winslow, D. F. Scollan, and C. K. Yung. WEAVE: a system for visually linking 3-D and statistical visualizations applied to cardiac simulation and measurement data. In *Proc. of IEEE Visualization*, pages 489–492, 2000.

[15] M. Harreby, J. Kjer, G. Hesselsøe, and K. Neergaard. Epidemiological aspects and risk factors for low back pain in 38-year-old men and women: a 25-year prospective cohort study of 640 school children. *European Spine Journal*, 5(5):312–318, 1996.

[16] K. Hegenscheid, R. Seipel, C. O. Schmidt, H. Völzke, J.-P. Kühn, R. Biffar, H. K. Kroemer, N. Hosten, and R. Puls. Potentially relevant incidental findings on research whole-body MRI in the general adult population: frequencies and management. *European Radiology*, 23(3):816–826, 2013.

[17] M. Hermann, A. C. Schunke, T. Schultz, and R. Klein. A visual analytics approach to study anatomic covariation. In *IEEE PacificVis 2014*, Mar. 2014.

[18] J.-F. Im, M. J. McGuffin, and R. Leung. Gplom: The generalized plot matrix for visualizing multidimensional multivariate data. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2606–2614, 2013.

[19] P. Klemm, L. Frauenstein, D. Perlich, K. Hegenscheid, H. Völzke, and B. Preim. Clustering Socio-demographic and Medical Attribute Data in Cohort Studies. In *Bildverarbeitung für die Medizin (BVM)*, pages 180–185, 2014.

[20] P. Klemm, K. Lawonn, M. Rak, B. Preim, K. Tönnies, K. Hegenscheid, H. Völzke, and S. Oeltze. Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data. In J. F. Michael Bronstein and K. Hormann, editors, *VMV 2013 - Vision, Modeling, Visualization*, pages 121–128, Lugano, 11.-13. September 2013.

[21] P. Klemm, S. Oeltze, K. Hegenscheid, H. Völzke, K. Toennies, and B. Preim. Visualization and exploration of shape variance for the analysis of cohort study data. In *Vision, Modeling & Visualization*, pages 221–222. The Eurographics Association, 2012.

[22] Z. Konyha, K. Matkovic, and H. Hauser. Interactive visual analysis in engineering: A survey, Apr. 2009.

[23] M. Lang-Tapia, V. España-Romero, J. Anelo, and M. J. Castillo. Differences on spinal curvature in standing position by gender, age and weight status using a noninvasive method. *Journal of applied biomechanics*, 27(2), 2011.

[24] S. Oeltze, H. Doleisch, H. Hauser, P. Muigg, and B. Preim. Interactive Visual Analysis of Perfusion Data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 13(6):1392–1399, 2007.

[25] S. Oeltze, H. Hauser, and J. Kehrer. Interactive visual analysis of scientific data, 2013. Half Day Tutorial at IEEE VIS, Seattle, WA, U.S.

[26] B. Preim, P. Klemm, H. Hauser, K. Hegenscheid, S. Oeltze, K. Toennies, and H. Völzke. *Visualization in Medicine and Life Sciences III*, chapter Visual Analytics of Image-Centric Cohort Studies in Epidemiology. Springer, 2014.

[27] M. Rak, K. Engel, and K. Toennies. Closed-form hierarchical finite element models for part-based object detection. In *VMV 2013 - Vision, Modeling, Visualization*, pages 137–144, Lugano, 11.-13. September 2013.

[28] T. Rudas. *Odds ratios in the analysis of contingency tables*. Number 119. Sage, 1998.

[29] S. Salvador and P. Chan. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In *Proc. of Tools with Artificial Intelligence. ICTAI*, pages 576 – 584, 2004.

[30] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc of Visual Languages*, pages 336–343. IEEE, 1996.

[31] M. Steenwijk, J. Milles, M. van Buchem, J. H. C. Reiber, and C. Botha. Integrated Visual Analysis for Heterogeneous Datasets in Cohort Studies. *Proc. of IEEE VisWeek Workshop on Visual Analytics in Health Care*, 2010.

[32] S. Thew, A. Sutcliffe, R. Procter, O. de Bruijn, J. McNaught, C. C. Ven-

ters, and I. Buchan. Requirements Engineering for e-Science: Experiences in Epidemiology. *Software, IEEE*, 26(1):80–87, 2009.

[33] J. J. Thomas and K. A. Cook. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press, 2005.

[34] E. Tufte. The cognitive style of powerpoint: pitching out corrupts within cheshire, 2003.

[35] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser. Hypothesis generation by interactive visual exploration of heterogeneous medical data. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 1–12. Springer, 2013.

[36] M. van Tulder, B. Koes, and C. Bombardier. Low back pain. *Best Practice & Research Clinical Rheumatology*, 16(5):761 – 775, 2002.

[37] H. Völzke, D. Alte, C. Schmidt, et al. Cohort Profile: The Study of Health in Pomerania. *International Journal of Epidemiology*, 40(2):294–307, Mar. 2011.

[38] H. Wiley. Hypothesis-free? no such thing. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA (US), 2008.

[39] Z. Zhang, D. Gotz, and A. Perer. Interactive visual patient cohort analysis. In *Proc. of IEEE VisWeek Workshop on Visual Analytics in Health Care*, 2012.