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 (*associations*). Modern cohort studies also incorporate medical image data. The statistically driven epidemiological workflow only allows to determine *associations* between image-derived metrics such as distances extracted from landmarks of the segmentation model.

We propose an Interactive Visual Analysis (*IVA*) 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 *associations* by applying an iterative analysis approach.

Index Terms—Interactive Visual Analysis, Epidemiology, Spine

1 Introduction

Epidemiology aims at characterizing 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 [11]. 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. Methods like regression analysis check the attribute list for statistical soundness.

Longitudinal population-based studies, such as the Study of Health in Pomerania (SHIP) [41], gather as much information as possible about a defined sample of people (a *cohort*). Cohort studies often include medical image data. These data need to be concurrently analyzed with the non-image data. This requires their segmentation and the derivation of features such as kidney volume, curvature of the spine or tissue texture of the liver. Semi-automatic techniques are more promising but challenging since the used methods as MRI and ultrasound are subject to inhomogeneity and noise. Analyzing spatial data with respect to other epidemiological factors requires techniques which reach beyond standard statistical methods.

Compiling a list of features for tests of statistical resilience based on experience-driven hypotheses leaves out other features in the data which potentially interact with a disease. This also applies to the chosen landmarks which are used to quantify medical image data information. The standard workflow lacks methods which highlight features of interest the epidemiologists did not consider.

We propose an Interactive Visual Analysis (*IVA*) approach [36] to provide a way to analyze 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. Easy publishing of developed methods driven by modern web technologies intends to trigger a fast feedback loop between us and the epidemiologists. We applied our approach to a data set compiled to analyze diseases related to the lumbar spine and aim to determine features, which indicate pathological changes.

Our contributions are:

- an Interactive Visual Analysis workflow for analyzing image-based epidemiological data including both hypothesis-driven and for hypothesis-generation based on a characterization of the standard epidemiological workflow
- visualization techniques, which combine both information visualization and 3D rendering of organ shapes as well as combining them with epidemiological graphics and key figures.
- highlighting interesting subject groups and feature associations using shape-based clustering and statistical contingency measures
- implementing the presented methods as a web framework based on `WebGL`, `D3.js` and `NodeJS`.

2 Epidemiological Background

In this section we describe the epidemiological workflow and associated requirements.

2.1 Epidemiological Workflow

The diversity of epidemiology is reflected in the different experts who work at cohort 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 hypotheses about disease specific risk factors. Following Thew and colleagues, the workflow, shown in Figure 1 (a), can be characterized as follows [35]:

1. A Hypothesis is derived from 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.

2. A set of attributes depicting conditions affected by the hypothesis is compiled accordingly.

3. Confounding features are adjusted so that they do not influence the effect size of a attribute.

4. Statistical methods such as regression analysis assess the association of selected features with the investigated condition.

Reproducibility of results is an epidemiological key requirement. 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.

In longitudinal cohort design, grouping subjects using epidemiological features is essential in order to allow per-group risk determination. Grouping depends on the underlying hypothesis. Age for example is 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 standard on how to categorize these.

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 the association of two features are influenced by a third feature which needs to isolated. A famous example for a confounder is the correlation observable between shoe size and mortality where it can be observed that people with larger shoes have a smaller life expectation than people with small shoe size. Shoe size actually is associated with gender where women have a smaller feet than men but have a longer life expectation.

Statistics tools such as `SPSS`[1] play a major role for analyzing epidemiological data. Graphic data representation is largely used to present results rather than gaining insight.

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 yield a complex data space for each subject. For ethical, legal or medical reasons some features can not be gathered for each subject. The most obvious example is women-specific questions about menstrual status or number of born children. Follow-up examinations or -questions for conditions like medications taken after a diagnosed disease also yield features 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 (*Yes* or *No*). Dichotomous data can also be derived by aggregating features to yield only two manifestations (e.g. subjects younger or older than 50 years). Medical examinations mostly comprise categorical (e.g. levels of back pain) and continuous values (e.g. age or body size).

Image acquisition. Imaging techniques emitting hazardous amounts of radiation for the subject are not suited for ethical reasons. MRI the main method for collecting cohort study imaging data. The image quality is a tradeoff between accuracy and affordability [30]. This often yields image resolutions inferior to those of clinical day-to-day practice, which makes their analysis more challenging. The equipment used to gather medical image data is kept, if possible, on the initial software and hardware version during a longitudinal study to ensure comparability in and between acquisition cycles.

Image analysis. Decisions have to be made on how image data are *compared* and *quantified*. Segmentation masks labeling the voxels of an anatomical structure would be ideal since many different key figures, e.g. volume, largest diameter or aspect ratio, can be derived from them. Since reliable and efficient segmentation techniques for these data are not available in general, 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 such as top and bottom point are also not nearly as expressive and versatile as segmentation masks describing the whole vertebra. 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 determining the type of tisse).

[1]Product of IBM; `www.ibm.com/software/de/analytics/spss/`

2.3 The Study of Health in Pomerania (SHIP)

After the pioneering Rotterdam study (started in 1990) several MR imaging studies initiatives have evolved. They slightly differ in clinical focus, acquired data and epidemiological research questions. Starting in 1997 with a cohort consisting of 4.308 subjects, the SHIP, located in northern Germany, aims to characterize health and disease in the widest range possible [41]. Data is collected without focus on a group of diseases. 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 [17]. The MRI protocol features a rich number of sequences. A second cohort `SHIP-Trend` was established in 2008. 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

Designing a visualization which conveys all data aspects equally is challenging. Given the number of features of epidemiological data sets and their different manifestations, the strength of different visualization techniques need to be combined [4, 24]. 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.

Their focus on hypothesis generation based on parallel assessment of multiple data features makes the work of Turkay and colleagues is closest to ours albeit our emphasis on processing medical image data and variables with categorical manifestations [39]. Their methods aim to amplify a hypothesis generation process for analyzing data of a Norwegian aging study. Statistical measures of continuous variables such as mean, standard deviation, skewness, or inter-quartile range are used to create *dimension plots* that make them comparable with respect to the derived descriptive measures such as voxel number of a segmented structure. The method is strongly dependent on the descriptive measures of the epidemiological factors.

Hypotheses based on observations of changes in these plots may impose *overfitting* 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 associations with non-image epidemiological factors.

Visualizing Image and Non-Image Data. Gresh and colleagues proposed `WEAVE`, one of the first systems, which concurrently analyzed medical image and non-image data using linked views [15]. 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]. They incorporated data mining methods such as dividing the data space using a k-nearest-neighbor technique and the PCA. 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 [34]. 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 [43]. Their understanding of a cohort differs from the understanding of the term in an epidemiological context by denoting every subject group divided by parameter as individual cohort.

Visualizing Heterogenous Non-Image Data. Generalized Pairs Plots (GPLOMS) are an information visualization technique comparing heterogenous variables pairwise using a plot-matrix grouped by type [20]. They are useful to gain an overview over numerous variables and their distributions. Histograms, bar charts, scatterplots and heat maps are used to visualize variable combinations with regard to their type. Dai and colleagues explored risk factors by incorporating choropleth maps of epidemiological features (e.g. mortality rates in a region) with parallel coordinates, bar charts and scatterplots with integrated regression lines [9]. Their findings yielded a *Concept Map* which linked cancer-related associations via graph edges. Chui and colleagues visualized associations in time-dependent epidemiological data using time-series plots highlighting risk factors differences in age and gender [7].

Commercial Data Visualization. Commercial systems such as Tableau[2] or Spotfire[3] 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, but the data processing possibilities such as derivation of new variables or the 3D rendering capabilities are very limited. These systems share limitations in extensibility to a specific problem domain.

Visualizing Shape Variance. 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 [6]. 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 [5]. Via mesh morphing interpolated views can be created by the user in a separate view as well as comparisons in a contour view. Distance to the mean shape is color-coded. We incorporate the idea of combining 3D-Shape rendering with information visualization techniques. Hermann and colleagues identify local deformation changes by investigating shape related difference [19]. The user specifies a deformation of interest and showing corresponding changes in the shape using covariance tensors. This method allowed for rapid hypotheses validation and was able to reproduce textbook knowledge.

SHIP-Data analysis. Klemm and colleagues visualized lumbar spine variabilities based on an semi-automatic shape-detection algorithm of 490 participants of the SHIP-2 cohort [22]. 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 shown. 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 [21]. Niemann and colleagues presented a interactive data mining tool for assessment of risk factors on hepatic steatosis, the fatty liver disease [26]. Association rules created by data mining methods can be analyzed interactively with their presented tool and highlight potentially overlooked features.

Interactive Visual Analysis The strength of the *IVA* approach described in the next section is its versatility with respect to the application field [24]. Oeltze and colleagues combined a linked view representation of results from a statistical analysis with feature localizations of the tissue perfusion with the goal of its evaluation [27].

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.

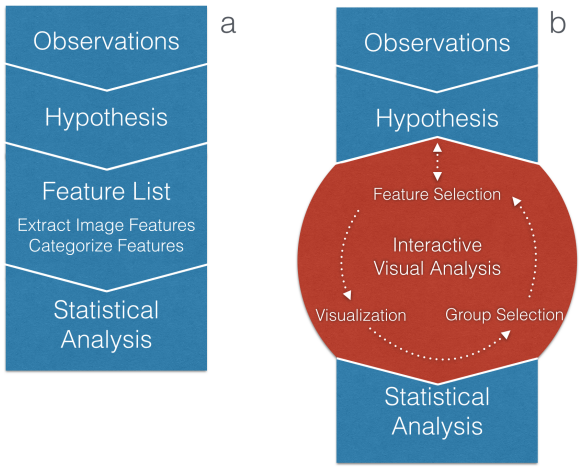

Fig. 1. *IVA* tools are able to complement parts of the epidemiological workflow, not replace it. The appropriate combination of statistical- and interactive driven analysis shows promising potential to unveil the information in the data. (a) shows the standard epidemiological workflow, (b) the IVA supported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

4 Image Centric Cohort Study Data in Interactive Visual Analysis Context

Subsection 2.1 described the epidemiological workflow as a sequence of steps taken by domain experts which need to comprise reproducibility and statistical integrity. Figure 1 (a) shows this workflow as consecutive series of steps. Introducing the *IVA* principle to the epidemiological application domain does aim to compensate its weaknesses rather than replacing the existing workflow. In the current state the workflow treats the data like a black box. Statistical tests on features associated to a hypothesis yields a value deciding whether the data supports the hypothesis or not. Features not included in the analysis may potentially support the chosen hypothesis by discriminating the population in the expected way, but are not highlighted in any way. This becomes even more important when the workflow is adapted to the analysis of the medical image data, where domain experts annotate landmarks tediously to derive metrics such as diameters. This leaves out the majority of information in the 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.

IVA tries to illuminate the black box by making the domain experts part of an iterative feature list selection process, which is shown in Figure 1 (b) as part of the epidemiological workflow. 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 data is an imminent danger as described by Turkay and colleagues [39].

Domain and Range Variables. In the *IVA* context, data is divided into two major view types. The human body exposing shape information for the *physical view* [28]. This information space is usually displayed via volume rendering techniques [27]. These variables spanning the 3D-space are also referred to as *domain variables*. *Range variables* in the epidemiological context can be divided twofold:

- Variables derived from the image data. These measures quantify shape information to allow for comparison and can also be used to brush in the image space.

[2]Owned by Tableau Software; www.tableausoftware.com
[3]Owned by TIBCO; spotfire.tibco.com

- Epidemiological socio-demographic or medical attribute data. These values are associated with every subject represented in the image space. They do not describe shape information. This is the data epidemiologists usually want to correlate with image data.

4.1 IVA Patterns

IVA knows three different exploration patterns, handling associations of domain and range variables.

Local Investigation This pattern projects information from domain space to the range perspective. This step is more complicated in epidemiological context compared to other *IVA* application domains. Shape information can not be brushed using ROI-selections but rather has to employ techniques that specify local deformation changes [19] or subjects that belong to a shape class. *Feature selection* methods strongly depend on the type of segmentation used 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 [22]. These algorithms can be used to investigate associations between shape-groups and other non-image based variables. Analysis of outliers can indicate segmentation errors or an group of subjects sharing a pathology.

Feature Localization The vast majority of features 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 interested in the shape of subjects within a range of a set of variables that describe the current hypothesis. Epidemiologists may 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.

Multivariate Analysis Multivariate analysis incorporates brushing and linking of of views displaying non-image parameter. The type of the parameters which are compared determines the information visualization technique used. Statistic measures are needed which describe how variables correlate given the selected groups. These associations are also summarized using pivot tables which are popular in epidemiology.

4.2 Data Preprocessing

Transformation operations on the data to prepare it for a *IVA* system are denoted as data preprocessing.

Non-Image data. Multimodal features require different techniques. 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. Exporting the data in the respective tool to a CSV file and then convert it to file types that are easily manageable such as JSON or XML makes it readable for all modern programming languages. This can be achieved using data wrangling tools such as `OpenRefine`[4], which also validates the data (find missing data, clean up bad formatting, transform scales). Including the data dictionary, which stores information about each manifestation of a feature, helps to get a detailed description of data variables and the meaning and unit of measurement of their values. Missing data is denoted using Error codes indicating its cause ranging from ethical to medical and personal issues.

Image data. Processing the image data associated to each subject consists for the most part of information extraction about a anatomical structure. This is either done manually by experts setting landmarks (sometimes supported by algorithms connecting the land marks such as graph cuts [14]) or by a (semi)-automatic detection, registration and segmentation. These algorithms have to deal with a large inter-subject variability of the anatomical structure and need to produce reproducible results [30]. Model-based approaches have shown in principle to be effective for segmentation [12, 13] and detection [31]. 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 [10]. Intensity-based comparisons can be achieved using rigid image registration, but model based results however are preferable [23]. Grey value comparison is usually used to measure the quantity of fat, water, and–application specific–iron content (liver) or the distribution of grey and white brain tissue.

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

[4]Developed by Google, Open Source; `http://openrefine.org/`

4.3 Analysis workflow

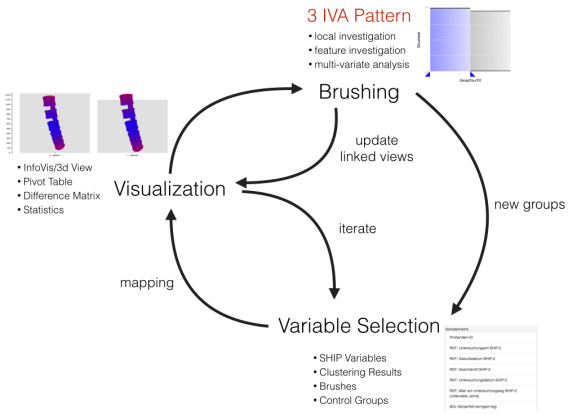

Fig. 2. Detailed Version of the *IVA Loop*. Usually starting with a selection of Feature of interest (user-driven or via data mining techniques), the data is mapped using a visualization technique appropriate for the selected data types. The data is visualized in the range and domain space, which can then be brushed, yielding new groups which can be investigated using further features. Note that adjacent steps are directly connected via feedback loops allowing for iterative refinement and give as much freedom to the user as possible.

Our proposed *IVA* workflow knows three major steps as illustrated in Figure 2: Feature selection, -visualization and -brushing. An hypothesis-driven analysis usually starts with the selection of features. A feature selection can also be derived from a shape-based clustering which creates shape groups. Hypothesis-generation with focus on image-data starts with a shape based clustering. The feature is mapped using a automatically chosen visualization appropriate for its data type. The visualization techniques have to combine both image- and non image data in order to set domain and range data in relation to each other. Subsequently the visualization can be either brushed or new variables can be added to the analysis. Brushes act as feature as they divide the subject space just like categorical variables and lead therefore to a new variable selection. Selecting features also triggers a contingency analysis highlighting variables interacting with selected features, which is described in the following section.

A sample workflow using interaction- and visualization techniques described in the following section can be seed in Figure 3.

4.4 Interaction- and Visualization Techniques

Suitability of an interaction- and visualization technique for epidemiological data depends on its ability to intuitively compare multiple data features at once while highlighting new interesting associations. The chances of a method to be adapted by epidemiologists increases highly if their domain language is used and the methods reflect routines they are used to. Visual evaluation of data is therefore as important as methods allowing for numerical data analysis. In the following sections we present the different parts of our proposed *IVA* system for image-based cohort study analysis.

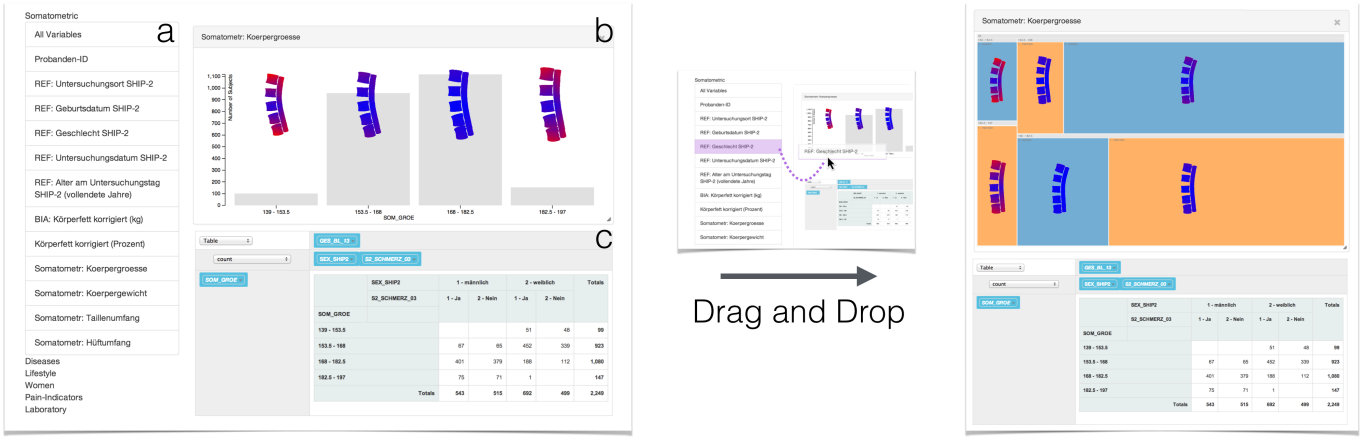

Fig. 3. (Left) Screenshot from the front-end which is divided as follows: (a) Sidebar which contains all features as well as the groups defined in the analysis process; (b) Canvas area where features can be added via drag and drop and the visualization chosen automatically according to the data type; (c) Interactive Pivot Table which shows exact numbers for each displayed variable combination. The data displayed is used to analyze the lumbar spine. Features can be added freely on the canvas via drag and drop. Dropping the *age* parameter on the already plotted *body size* container creates a mosaic plot combining both features (right).

4.4.1 System Structure

We divide the workspace into four major parts as seen in Figure 3.

- The sidebar, which contains all epidemiological features. Cluster results are treated like features and are part of the sidebar as well.
- The canvas which holds all visualizations. Elements can be added, arranged, resized and removed freely.
- The interactive Pivot Table gives detailed numerical information of the features in the canvas view. This view on the data come natural to epidemiologists.
- The contingency view depicting relations for features in the canvas.

Sidebar. An overview of all features 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 brushing or automated shape-clustering. Groups are treated like features since they work exactly the same way by dividing the subject space into labeled categories. Features can be dragged from the sidebar into the canvas area. This triggers an adaptive feature visualization suitable for the current data type.

Adaptive Feature Visualization. Inspired be previously the discussed `GPloms` [20] the visualization type is chosen dynamically based on the variable types and number of visualizations which need to be displayed. Following Tuftes concept of *small multiples* [38] the medical image data is directly included into the plot by including color coded mean shapes for each manifestation (Figure 3 (b)). The 3D-plots can be navigated using standard mouse inputs, the camera is synchronized between all views to enable direct comparison. Mapping the distance from a group mean shape to the global mean to color allows to assess local shape changes. 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. Brush selections are propagated to all visualizations allowing for fast feature querying.

Pivot Tables. Pivot tables are frequently used to present the data in epidemiological publications. Epidemiologists are used to process groups based on table representations so decided to introduce an interactive pivot table. As seen in Figure 3 (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. Another way to avoid clutter is the user-driven selection of displayed variables. To allow better comparison with respect to features the values of each cell can also be displayed as percentage of the feature represented of either the row or the column.

Automated Feature Suggestion using a Contingency Matrix. 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 [39]. Since the majority of our data are categorical features, we have to employ different solutions. The *Cramérs' V* contingency coefficient can be used to calculate coherences between categorical variables [8]. It is based on *Pearson's* X^2 distribution test [29], which uses contingency tables holding the counts of subjects for all possible manifestations of two variables. *Cramérs' V* is defined as:

$$V = \sqrt{\frac{X^2}{N(k-1)}}, \tag{1}$$

where X^2 equals *Pearson's chi squared*, N is the total of observations and k either the row or column count, depending on which one is lower. V assumes values between 0, meaning two variables are completely independent and 1 indicating they are the same. *Cramérs' V* is always positive and does therefore not allow statements about dependency direction.

It shares the same restrictions as *Pearson's* X^2 which include next to others expected count larger than 5 for 80% of the contingency table entry and no expected value can be smaller than one. Therefore are manifestations who are only exposed by small subject groups not assessable using this technique. This makes them not assessable for epidemiological analysis as well since statistical validation needs a minimum count to be seen as valid. The contingency matrix highlights contingencies of visualized features with features in the data set. This aims to highlight features which may interact with the focused hypothesis as well as trigger new hypotheses. Contingency is visualized using an interactive adjacency matrix with association power mapped to color hue. The distinction whether an association is a confounder or and effect depends on the context defined by the hypothesis and is a decision to be made by the domain expert.

In the following Section we discuss how we implemented the presented methods using free on open Web standards.

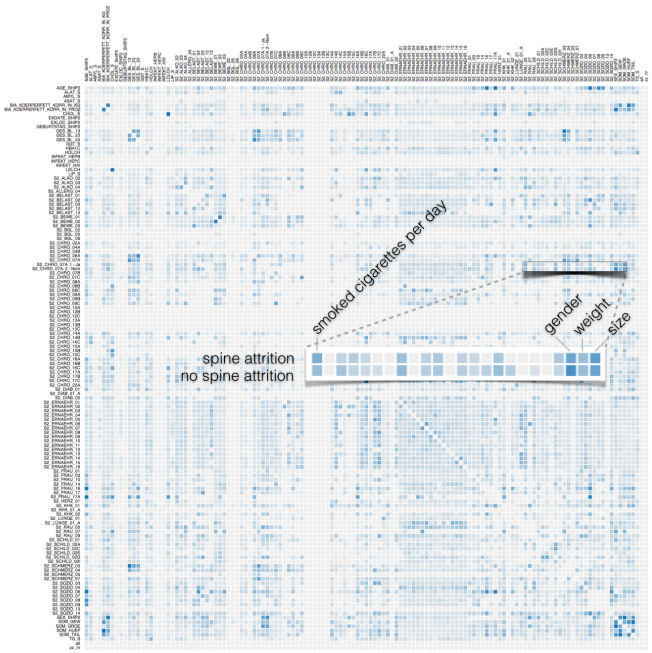

Fig. 4. Adjacency matrix of 131 features with a grand total of 17.161 combinations. Similarity is calculated using the *Cramérs' V* contingency value. Color brightness encodes association strength. The enlarged excerpt shows associations for shape clusters of subjects with and without diagnosed spine attrition, which show associations between gender, weight, body height and smoking behavior.

4.5 Implementation

To provide a fast communication loop between method development and expert input, we decided to rely 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 for employing 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
- 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.
- Web technologies are free.

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 `NodeJS`[5], 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.

[5]Developed by Joyent Inc, `nodejs.org`

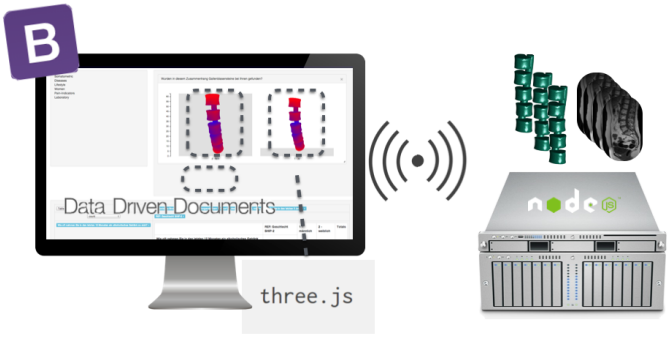

Fig. 5. *Image is not final.* The front-end solution (left) uses state of the art web technologies such as HTML5/CSS3, WebGL and SVG to display the data. The NodeJS based back-end (right) stores all image- and non-image data and transfers it to connected clients. All computation heavy operations like 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.

Non-image data for all subjects including the data dictionary is stored in a JSON file on the Server. 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 *Cramérs' V* values 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`[6] 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 (D3.js)` [3] library, which works well for attaching data to visible elements like vector graphics and provides powerful transformation and mapping tools. `WebGL` rendering is done using the `Three.js`[7] and allows GPU Accelerated data rendering. Communication between Client and Server runs through the WebSockets protocol. Since our clustering algorithms are written in `MatLab`[8] we hat to access them using the `NodeJS` Server. We accomplished this by converting it to a parameterized standalone console application, that is spawned by `NodeJS` 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.

The following section describes how our presented *IVA* workflow works for a epidemiological use case.

5 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 [40]. Epidemiological analysis of lumbar back pain such the work of Harreby and colleagues [16] 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 [25]. 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 [11] (evtl. weglassen, Dopplung bei Epidemiological Background?):

- a better understanding of effects of preventive measures such as occupational health and safety regulations

[6]Developed by Twitter, `getbootstrap.com`
[7]Originally developed by Ricardo Cabello, `threejs.org`
[8]Owned by The MathWorks, `www.mathworks.com`

- 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).

5.1 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) [18].

5.2 Data Preprocessing

Following methods described in Section 4.2, the data is preprocessed as follows for the presented prototype.

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.

Image-Data. The lumbar spine was detected in the image data using a hierarchical finite element method according to Rak and colleagues [31]. 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 [22]. Due to incorrect initialization, strongly deformed spines, contrast-differences and artifacts, the model was not able to detect lumbar spines for all subjects. We obtained and work with 983 models. 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 [22].

5.3 Shape Visualization and Clustering

The tetrahedron-based model detection model described in Section 5.2 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 4.4.

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 6).

Shape based clustering is carried out via Agglomerative Hierarchical clustering of the spine canal, centerlines which are described in Section 5.2 [22]. 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 [32]. For details, see [22]. The method has proven to produce comprehensible results on a preliminary data set and was able to reproduce textbook knowledge [22].

5.4 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.

5.4.1 Hypothesis free Analysis

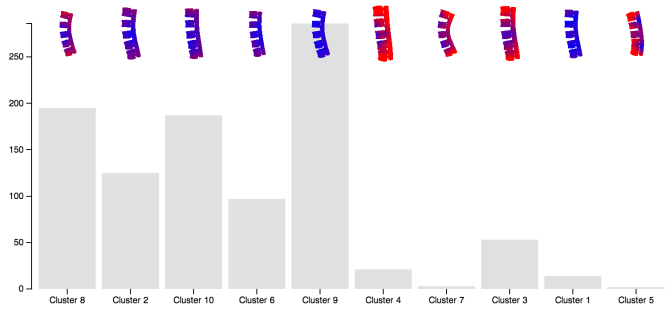

Fig. 6. *Image is not final.* (a) Clustering result of all subjects. The bar chart height indicates the number of subjects in the cluster. The difference to the mean-shape is color-coded where blue represents no difference and red a large difference.

Analyzing the data set without a prior hypothesis requires a starting point which gives a overview over the data first [33]. 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 6 (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 reveal associations of this group with employment status, body size, age, thyroid nodules and blood-fat value. TODO Feedback Greifswald, While all observed features seem as plausible associations related to back pain, the values indicate subjects with chronic back pain which radiates into the legs. Metabolic parameters such as blood-fat and blood-sugar are also possibly associated features. Employment status may be a confounder describing the lifestyle of a subject.

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 [42].

5.4.2 Hypothesis based Analysis

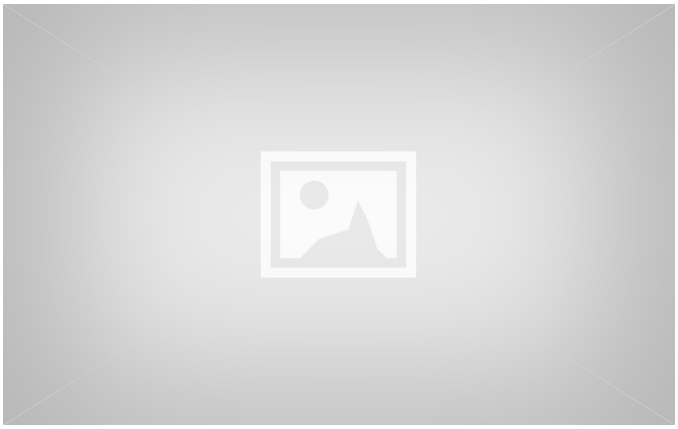

Fig. 7. *Image is not final.* (a) "Did you experience back pain in the past three month" Yes No; (b) Clustering of Yes

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 7 (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 7 (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.

5.4.3 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 [37]. 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, they are cross check for associations in SHIP-TREND as a second, independent population sample, which was examined with methods identical to SHIP.

The presented methods guide attention to features which where not in the focus of attention and expectance of clinical researchers. The explorative nature of the methods work well for gathering associations 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.

Displaying MRI scans for the subjects in the outlier cluster is promising because they are highly likely to exhibit pathologies.

6 Summary and Conclusion

The presented are classified as extension to the existing statistics driven epidemiological. Image-centric epidemiological analysis benefits from a IVA approach by illuminating the data-black box. Visualization of multivariate data using connected views and different views allows to get fast visual feedback about subject groups. Brushing and linking makes the data tangible and adaptable to formulated hypotheses. The use of pivot tables is familiar to epidemiologist while embracing the power of interactive adjustment of the features shown. Automatic suggestion of correlations using contingency methods like *Cramérs' V* trigger *hypothesis generation* by highlighting features potentially overlooked by the experts. Shape based clustering assesses variability of a anatomical structure in the context of non-image features such as disease indicators or lifestyle factors. The iterative nature of the *IVA* approach reflects versatility as both *hypothesis driven analysis* and *hypothesis generation* is allowed.

Our future work comprises a number of improvements:

- dynamic discretization of continuous variables to better fit the data distributions
- shape brushing methods to intuitively query subjects using image-data
- incorporate more statistical methods and views that are familiar to the epidemiologists (odds ratios, box plots)
- calculation of contingency differences for selected sub groups– highlight how feature correlation change for a given groups
- more visualization techniques which incorporate both image and non-image data
- support control-groups to check resilience of observations as suggested by Fletcher and Fletcher [11].

To reduce the number of false positive findings, the data space can also 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 the number of image-centric cohort studies, participating subjects, gathered features and imaging modalities rises and advances towards comparability between cohort studies are made, the gap between data complexity and assessability increases. Our work focuses on closing this gap, allowing the domain experts to dig deep into the data and potentially obtain unexpected findings. We believe web technologies pave the way to allow as many experts as possible to analyze this data and provide a fast exchange between users and developers employing many different devices. Visual analysis shows to be a promising way to clear the view on complex epidemiological data to uncover its secrets.

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] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[4] 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.

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

[6] 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.

[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] H. CRAMÉR et al. Mathematical methods of statistics. *Mathematical methods of statistics.*, 1946.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] K. Hegenscheid, J. Kuhn, H. Völzke, R. Biffar, N. Hosten, and R. Puls. Whole-Body Magnetic Resonance Imaging of Healthy Volunteers: Pilot Study Results from the Population-Based SHIP Study. *Proc. of RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 181(08):748–759, 2009.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] U. Niemann, H. Vlzke, J.-P. Khn, and M. Spiliopoulou. Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis. *Expert Systems with Applications*, 2014.

[27] 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.

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

[29] K. Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] S. Thew, A. Sutcliffe, R. Procter, O. de Bruijn, J. McNaught, C. C. Venters, and I. Buchan. Requirements Engineering for e-Science: Experiences in Epidemiology. *Software, IEEE*, 26(1):80–87, 2009.

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

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

[38] E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. CT: Graphics Press, 1983.

[39] 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.

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

[41] 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.

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

[43] 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.