# Interactive Visual Analysis of Image-Centric Cohort Study Data

Paul Klemm, Steffen Oeltze-Jafra, Kai Lawonn, Katrin Hegenscheid, Henry Völzke, Bernhard Preim

**Abstract**—Epidemiological population studies impose information about a set of subjects (a *cohort*) to characterize disease-specific risk factors. Cohort studies comprise heterogenous variables describing the medical condition as well as demographic and lifestyle factors and, more recently, medical image data. We propose an Interactive Visual Analysis (*IVA*) approach that enables epidemiologists to rapidly investigate the entire data pool for *hypothesis validation* and *generation*. We incorporate image data, which involves shape-based object detection and the derivation of attributes describing the object shape. The concurrent investigation of image-based and non-image data is realized in a web-based multiple coordinated view system, comprising standard views from information visualization and epidemiological data representations such as pivot tables. The views are equipped with brushing facilities and augmented by 3D shape renderings of the segmented objects, e.g., each bar in a histogram is overlaid with a mean shape of the associated subgroup of the cohort. We integrate an overview visualization, clustering of variables and object shape for data-driven subgroup definition and statistical key figures for measuring the association between variables. We demonstrate the *IVA* approach by validating and generating hypotheses related to lower back pain as part of a qualitative evaluation.

**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. The introduction of preventive measures in medicine and beyond is also based on epidemiological research, where, for example, subgroups with increased risk are identified [12]. Observations made by clinicians in the daily routine are translated into hypotheses for epidemiological research. These are used to determine environmental and lifestyle factors as well as medical examination results that may influence a disease. Potentially useful data variables 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*). The cohort consists of several thousands of people, randomly selected to avoid any bias. The subjects are selected without focus on a certain disease. A large cohort size is essential to investigate differences between healthy and diseased people. Cohort studies often include medical image data. The concurrent analysis of image data and non-spatial epidemiological factors requires techniques that reach beyond standard statistical methods. For instance, segmentation of the image data is required for an analysis of anatomical structure and of possible correlations between this structure and epidemiological factors. Semi-automatic segmentation techniques are promising but also challenging, since the employed modalities, such as magnetic resonance imaging (MRI) and ultrasound, are subject to inhomogeneity and noise.

Compiling a list of variables for tests of statistical resilience based on experience-driven hypotheses leaves out other variables in the data which potentially interact with a disease. This also applies to the chosen landmarks used to quantify medical image data information. The standard workflow lacks methods for automatically identifying correlations possibly buried deep in the data or overseen by the expert. Also, only a small subset of factors can be concurrently analyzed.

We propose an Interactive Visual Analysis (*IVA*) approach [37] for the combined analysis of image and non-image data. Visual queries and direct feedback of Visual Analytics systems allow for a fast exploration of the data space incorporating many different variables. Intended as an extension to the well-established epidemiological tools it provides a way to rapidly validate hypotheses and to trigger *hypothesis generation* using data mining methods, such as clustering. *Hypothesis generation* gains importance since the number of epidemiological variables increases and the focus shifts towards more complex relations involving more than two variables. Our contributions are:

- an IVA workflow for cohort study data to allow both, hypothesis-driven analysis and hypothesis generation,
- visualization techniques, which incorporate both information visualization and 3D rendering of organ shapes as well as combining them with epidemiological graphics and key figures,
- highlighting subject groups and variable associations using shape-based clustering and statistical contingency measures.

We applied our approach to a data set compiled to analyze lower back pain and aim to determine variables, which indicate pathological changes. This data set comprises 127 variables and 2 sequences of MRI data from 6,753 subjects. The method implementation is web based to allow a fast feedback loop with domain experts.

## 2 Epidemiological Background

This section covers the epidemiological workflow and 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 et al. [36], the workflow can be characterized as follows:

1. A hypothesis is derived from observations made by clinicians in their daily routine.
2. A set of variables depicting conditions affected by the hypothesis is compiled accordingly.
3. Confounding variables are identified and taken into account (for example using stratification).
4. Statistical methods, such as regression analysis, assess the association of selected variables with the investigated disease.

- *Paul Klemm, Steffen Oeltze-Jafra, Kai Lawonn, Bernhard Preim are with Otto-von-Guericke University Magdeburg, Germany. E-mail: {klemm,oeltze,lawonn,preim}@ovgu.de*
- *Katrin Hegenscheid, Henry Völzke are with Ernst-Moritz-Arndt University Greifswald, Germany. E-mail: {katrin.hegenscheid,voelzke}@uni-greifswald.de*

*Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.*
*For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.*

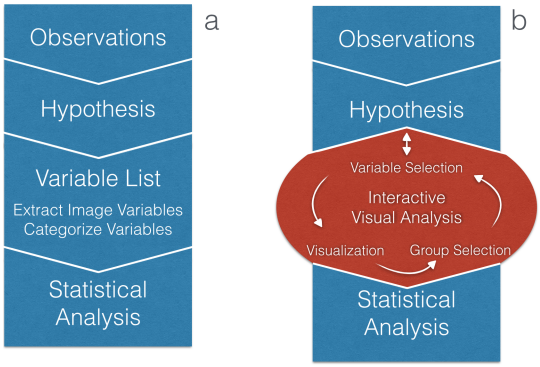


Fig. 1. (a) The standard epidemiology workflow consists of four steps. (b) *IVA* tools complement parts of this workflow instead of replacing them. The combination of statistical and interactive analysis shows promising potential to unveil information in the data. We call the iterative red highlighted part *IVA Loop*, described in detail in Figure 2.

The workflow is shown in Figure 1 (a) and serves as orientation for our approach. We focus on the potential of image data and attempt to support *hypothesis generation*.

*Reproducibility* of results is an epidemiological key requirement. It is difficult to achieve, since many physicians are involved when thousands of subjects are examined and interviewed. Thus, both intra- and inter-observer variability needs to be low for all aspects of a cohort study examination. Longitudinal studies require the acquired variables to be comparable for evaluation. Grouping subjects using epidemiological variables is essential in cohort studies 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. These groups strongly depend on the condition of interest and therefore there is no standard for their categorization.

*Relative risks* are determined to detect if a subject is prone to be affected by a certain disease. This includes confidence intervals indicating the certainty of that variable being a risk factor.

Statistical tools such as SPSS[1] play a major role for analyzing epidemiological data. Epidemiologists employ static graphical data representations primarily at the very end of an analysis session for presenting results or observing trends in the data.

### 2.2 Epidemiological Data

Epidemiological data are strongly 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 or medical reasons some variables cannot be gathered for each subject, e.g. women-specific questions about menstrual status or number of born children. Follow-up examinations or questions about conditions such as medications taken after a diagnosed disease also yield variables only available for a small amount of subjects.

Indicators for medical conditions as well as questions about a subject's lifestyle are often *dichotomous*–they have two manifestations (*Yes* or *No*). Dichotomous data can also be derived by aggregating variables to yield only two manifestations (e.g. subjects younger or older than 50 years). Medical examinations comprise categorical (e.g. levels of back pain) and continuous values (e.g. age or body size). Data analysis is usually carried out by calculating correlations, which is challenging due to the data type heterogeneity. Parameter correlation can also be associated with confounding, which cannot be automatically predicted. It has to be judged by a domain expert. Sparse populated variables are hard to assess statistically. Too few data samples may distort the real underlying distributions. Statistical correlations are prone to *confounding*, meaning that the association of two variables is influenced by a third variable, which needs to be isolated. A famous example is the association between shoe size and mortality, where it can be observed that people with larger shoe size have a smaller life expectation. The shoe size is actually associated with gender, where women have smaller feet and a longer life expectation.

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

**Image Acquisition.** Imaging techniques involving ionizing radiation for the subject are not suitable for ethical reasons. Therefore, MRI is the main method for collecting cohort study imaging data. The image quality is a tradeoff between accuracy and affordability [31]. This often yields image resolutions inferior to those of clinical practice.

**Image Analysis.** Decisions have to be made about *comparison* and *quantification* of image data. Segmentation masks representing the voxels of an anatomical structure would be ideal, since key figures, e.g., volume, largest diameter or aspect ratio, can be derived from them. Since reliable and efficient segmentation techniques are not available in general, epidemiologists are forced to measure the data by hand. Information derived by landmarks, such as top and bottom point of a vertebra, are by far not as expressive and versatile as segmentation masks describing its whole shape. They are also prone to a high inter-observer variability. Morphometric information from landmarks comprises thickness, diameter or length of a structure as well as grey value distribution in an area.

### 2.3 The Study of Health in Pomerania (SHIP)

After the pioneering Rotterdam study (started in 1990), several MR imaging study initiatives were initiated. They slightly differ in clinical focus, acquired data and epidemiological research questions. Starting in 1997 with a cohort of 4,308 subjects, the SHIP, located in Northern Germany, aims to characterize health and disease in the widest range possible [41]. Data are collected without focus on a group of diseases. This allows to query the data regarding many diseases and conditions. Subjects were examined in a 5-year time span, continuously adding new parameters including MRI scans in the last iteration [16].

## 3 Prior and Related Work

This section describes prior and related work and covers visual analysis methods incorporating both image and non-image data.

**Visual Analysis of Image and Non-Image Data.** Our work is closest to that of Steenwijk [35], Turkey [39], and Angelelli [1] and colleagues, who employ multiple coordinated view systems for the analysis of cohort study data.

Steenwijk et al. [35] propose a relational database to organize cohort study data for a visual analysis based on linked views such as parallel coordinates, scatter plots and time plots. Information about medical image data is incorporated via mappers, which extract comparable metrics about the data. Medical image data can be displayed individually for subjects, e.g., for analyzing outliers. While we use a similar approach when analyzing non-image data, our process also includes *overview visualizations* and statistical suggestions of potentially interesting variables.

Turkey et al. [39] present *hypothesis generation* based on descriptive statistics of the data dimensions. Key figures describing the distribution of data values, e.g., standard deviation and interquartile range, are computed per dimension and analyzed by pairs in a *deviation plot*. The *dual analysis* of data items and dimensions in multiple linked views led to several hypotheses in analysis sessions with domain experts. Hypotheses based on observations in the deviation plot may impose *overfitting* to the data because the measures highlight only parts of the statistical changes. Our approach uses information extracted from the segmented image data (such as 3D meshes) and variable associations with non-image epidemiological factors.

Angelelli et al. [1] focus on the data organization for an interactive visual analysis of heterogeneous cohort study data. The proposed data-cube model facilitates the seamless integration of image-based and non-image data. In a demonstration of the model, brain image data was integrated into the analysis by first segmenting brain regions and tracking neural pathways and then deriving attributes from both, e.g., volume and fractional anisotropy. A multiple coordinated view

framework then linked spatial and non-spatial data views. Our integration of image data into the analysis is similar to the work of Steenwijk [35] and Angelelli [1] and colleagues. While they offer a single spatial view for visualizing image-based information of one subgroup of the cohort, we provide multiple views showing the information of subgroups and their respective deviation from the entire cohort.

Gresh et al. [14] proposed WEAVE, one of the first systems which concurrently analyzed medical image and non-image data using linked views. Blaas et al. [2] presented a similar system, which also analyzed medical image data and variables derived from them using views from the variable and physical space. Both works are restricted to the analysis of one case at a time and to non-image data with a unique spatial reference, e.g., voltage simulated across the heart muscle. In epidemiology, multiple cases must be concurrently investigated and non-image data often lacks a spatial reference, such as gender or age.

Visual Analysis of Heterogenous Non-Image Data. Zhang et al. [44] 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. Their definition of a cohort differs from the understanding of the term in an epidemiological context by denoting every parameter-divided subject group as individual cohort. Due to the short paper length, detail is missing on the data types and their algorithms of identifying similar subjects or whether they employ statistical measures. We employ the idea of adding variables via drag and drop into a canvas area.

Generalized Pairs Plots (GPLOMS) are an information visualization technique comparing heterogenous variables pairwise using a plot-matrix grouped by type [10, 19]. They are useful to gain an overview over numerous variables and their distributions. Histograms, bar charts, scatter plots and heat maps are used to visualize variable combinations with regard to their type. The resulting matrix provides an *overview visualization*, but requires a lot of screen space for many variables (127 in our application scenario). We incorporate the idea of adaptive type-dependent visualizations. Dai et al. [9] explored risk factors by incorporating choropleth maps of epidemiological variables (e.g., mortality rates in a region) with parallel coordinates, bar charts and scatter plots with integrated regression lines. Their findings yielded a *Concept Map*, which linked cancer-related associations via graph edges. While their goal to identify possible risk factors using socio-economic and health data is similar to ours, they focus on iteratively refining defined hypotheses and on geographical data. We employ the use of small multiples for incorporating heterogenous data types for comparability. Chui et al. [6] visualized associations in time-dependent epidemiological data using time-series plots highlighting risk factor differences in age and gender. While the work shows how different visualization techniques provide insight into these data sets, it focuses on the time aspect, which is not present for our data.

Visualizing Shape Variance. Comparing tissue between many subjects requires shape variance visualizations. Caban et al. [5] investigated the suitability of variance visualizations of shape distribution models and concluded that users favor spherical glyph representations over deformation grids and likelihood volumes. The distribution of shapes in a space derived from a PCA is plotted by Busking et al. [4] in a 2D-projected plane of the space. Interpolated views can be created by the user in a separate view as well as comparisons in a contour view. Interpolation is carried out by mesh morphing. The distance to the mean shape is color-coded. We incorporate the idea of combining 3D shape rendering with information visualization techniques. Applying this technique to our data yielded a cluttered shape space due to the high subject count. The data needs to be abstracted to work in this context. Hermann et al. [18] identify local deformation changes by investigating shape-related differences on rodent mandibles. User-specified regions of interests are mapped to associated anatomic covariation using tensor visualization. This method enables rapid hypotheses validation and was able to reproduce textbook knowledge. It requires a spatial colocation of associated variables.

Prior Work. We visualized lumbar spine variabilities based on a semi-automatic shape detection algorithm of 490 participants of the SHIP-2 cohort [21]. Hierarchical agglomerative clustering divided the population into shape-related groups. As proof of concept, a relation between the size of the segmented shape and the measured size of the subjects was shown. This work focuses on incorporating these derived data as new variables, enabling to include it into the hypothesis validation and generation process. When applying clustering techniques to the non-image data it was found that k-Prototypes and DBSCAN are appropriate, but are strongly dependent on the chosen variables and distance measures [20]. Niemann et al. [26] presented an interactive data mining tool for the assessment of risk factors of hepatic steatosis, the fatty liver disease. Association rules created by data mining methods can be analyzed interactively with their tool and highlight potentially overlooked variables.

Interactive Visual Analysis. The strength of the *IVA* approach is its versatility with respect to the application field [22]. Oeltze et al. proposed a multiple coordinated view approach for the analysis of medical perfusion data [27] and biological multi-channel fluorescence microscopy data [28]. The approach is restricted to the investigation of a single subject at a time.

Lammarsch et al. [24] provided a workflow and terminology definition of Visual Analysis techniques. They define a model as a representation of system entities, phenomena, processes and hypotheses as models whose outcomes are not compared with real-world data (*validation*). The VA process is also reflected in our *IVA* loop.

Baldonado et al. [42] presented rules for designing multiple coordinated views. They point out the cost-benefit tradeoff introduced by the cognitive overhead by mentally connecting multiple views over more complex single views. Weaver et al. [43] extracted guidelines for cross-filtering multiple views by incorporating *views* mapping data to visual elements, *brushes* for selecting these elements and *switches* for linking brushing results between *views*. Our system follows the same rules for selecting subject groups, but our goal is to judge variable relations and potential outcomes.

The uniqueness of our workflow compared to the discussed work is threefold. (1) We incorporate 3D models abstracting shape information fused with non-image data visualizations, allowing to analyze local physiological changes related to non-image parameters. (2) We focus on hypothesis generation by discovering new relationships associated with shape information. (3) Overview visualizations using statistical abstractions aim to provide an unbiased variable relationship assessment.

## 4 Image-Centric Cohort Study Data in an Interactive Visual Analysis Context

We described the epidemiological workflow and emphasized the reproducibility and statistical integrity (recall Subsection 2.1). Introducing the *IVA* principle to the epidemiological domain aims to compensate the weaknesses of the existing workflow rather than replacing it (recall Fig. 1). In the current state, the workflow treats the data like a black box. Statistical tests on variables associated to a hypothesis yield a value for deciding whether the data supports the hypothesis. Variables not included in the analysis may potentially support the chosen hypothesis by discriminating the population in the expected way, but are not highlighted. 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 measures, such as diameters. This leaves out the majority of information in the image data by abstracting it to single values. Considering more complex parts of the data would make those results more trustworthy and could also 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 variable selection process (see Fig. 1 b). 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 [39].

Domain and Range Variables. In the *IVA* context, data are characterized by a combination of independent variables, such as space and/or time, and dependent variables, like temperature or pressure. Two kinds of views are employed to inspect the data:

- *physical views* [29], e.g. volume rendering, show information in the context of the spatio-temporal observation space [27], while
- *attribute views*, such as scatter plots and parallel coordinates, show relationships between multiple data attributes.

Transferred to epidemiological data, the residential area of cohort subjects could be interpreted as *space*, the different assessment cycles of a longitudinal study as *time*, and the image and non-image data as *dependent variables*. Our current work neglects geographical and temporal aspects. Instead we employ an abstract model and consider the subjects as living in a joint image space where each of them is represented by a segmented organ or structure. For instance, the lumbar spine is segmented over all subjects and all lumbar spines are co-registered spanning a joint space. Then, two types of dependent variables exist: the socio-demographic data and medical examination results, and variables derived from the segmented structures, e.g., spinal curvature or misalignment of the vertebrae. An alternative of the image space would be the shape space generated by extracting the major modes of variation from all segmentation results [4]. Based on our abstract model, the three analysis patterns of *IVA* can be employed.

Local Investigation refers to the inspection of dependent variables with respect to subsets of the image or shape space. For instance, the epidemiologist selects several lumbar spines with a common characteristic in the image or shape space and inspects the associated dependent variables in an attribute view [18]. The selection step requires dedicated interaction techniques for defining a subset. Alternatively, derived shape-related variables opposed in an attribute view or automatic techniques for shape clustering may be employed [21]. Clustering algorithms can be used to investigate associations between shape groups and other non-image variables. Analysis of outliers can indicate segmentation errors or a group of subjects sharing a pathology.

Feature Localization refers to the search for structures in the image or shape space with a defined characteristic. The epidemiologist may be interested in all female subjects with lower back pain and wishes to see the corresponding spines in a physical 3D view.

Multivariate Analysis refers to an investigation of multi-variate properties of the dependent data by specifying a variable in one attribute view while analyzing the value distribution with respect to other variables in other attribute views. Epidemiologists may define a variable in a scatter plot of the body mass index (BMI) and age to inspect the result in a histogram of body height. These associations may also be summarized using pivot tables, which are widely used in epidemiology.

### 4.1 Data Preprocessing

Non-Image Data. Data obtained using questionnaires or medical tests are often stored using statistical packages such as `SPSS`, which have a proprietary data format. Exporting the data in the respective tool to a `CSV` file and then converting it to file types that are easily manageable, such as `JSON`, makes it readable for modern programming languages. A data dictionary stores information about each manifestation of a variable. Detailed description of data variables, its meaning as well as unit of measurement are stored as a lookup table. Missing data are denoted using error codes indicating their cause ranging from ethical to medical and personal issues (recall Subsection 2.2).

Image Data. Information about anatomical structures, such as diameter or volumes, is extracted from the image data. This is either done manually by experts setting, landmarks or by a (semi-)automatic detection, registration and segmentation. These algorithms have to deal with a large inter-subject variability of the anatomical structure [31]. In principle, model-based approaches are effective for detection [32] and segmentation [13]. If a segmentation yields only binary masks, algorithms such as *Growing and Adaptive Shapes* can be applied for creating a surface grid where each point is comparable throughout the population [11]. Grey value comparison is used to measure the quantity of fat, water, and–application-specific–the 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 incorporates mostly positions, diameters, volumes, relative distances and alignment to other structures.

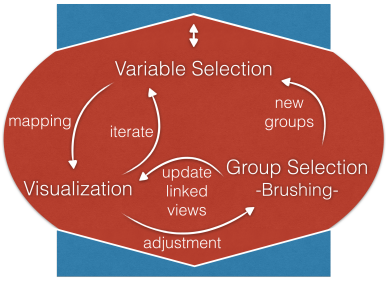


Fig. 2. Detailed *IVA Loop* as extension from Figure 1. Usually starting with a selection of a variable of interest (user-driven or via data mining techniques), the data are mapped using a visualization technique appropriate for the selected data types. The data are visualized in the range and domain space, which can be brushed, yielding new groups to be investigated using further variables. Note that adjacent steps are directly connected via feedback loops, allowing for an iterative refinement and giving as much freedom to the user as possible.

### 4.2 Analysis Workflow

Our proposed *IVA* workflow consists of three major steps, as illustrated in Figure 2: Variable selection, visualization and brushing. A hypothesis-driven analysis usually starts with the selection of variables or shape groups derived from a shape-based clustering. Hypothesis generation with focus on image data starts with a shape-based clustering or an *overview visualization* of all variables. The variable is mapped using an automatically chosen visualization appropriate for its data type (described in detail in the following section). The visualization techniques have to combine both image- and non-image data to set domain and range data in relation to each other. In our system, the visualization can either be brushed or new variables can be added to the analysis. Brushed regions are treated like categorical variables, as they divide the subject space in the same way. Selecting variables also triggers a *multivariate analysis* using contingency values (described in the following section) to highlight associated variables. A sample workflow using interaction and visualization techniques described in the next section can be seen in Figure 3.

## 5 System Design and Implementation

The suitability of visualization techniques for epidemiological data depends on their ability to compare multiple data variables while highlighting associations. Visual evaluations of data are therefore as important as methods allowing for numerical data analysis. In the following sections we present the different parts of our system.

### 5.1 Design and Visualization Techniques

Early it became clear that we have to rely a lot on online communication due to the large spatial distance towards each other. Hence, we built our system using web technologies. By running the prototypes on server machines, software exchange became as easy as sharing a weblink, giving us the opportunity to include the clinical experts in the development process with little effort. Incorporating the *IVA* workflow for image-centric cohort study data requires *overview visualizations* as well as *multivariate visualizations*, which bring image-derived information in context to non-image variables.

The focus on web technologies is not without tradeoffs. Classical UI elements, such as the menu bar or custom right-click menus, are technically possible, but not common in this domain. In favor of a clean layout, we designed the system without such components. Since

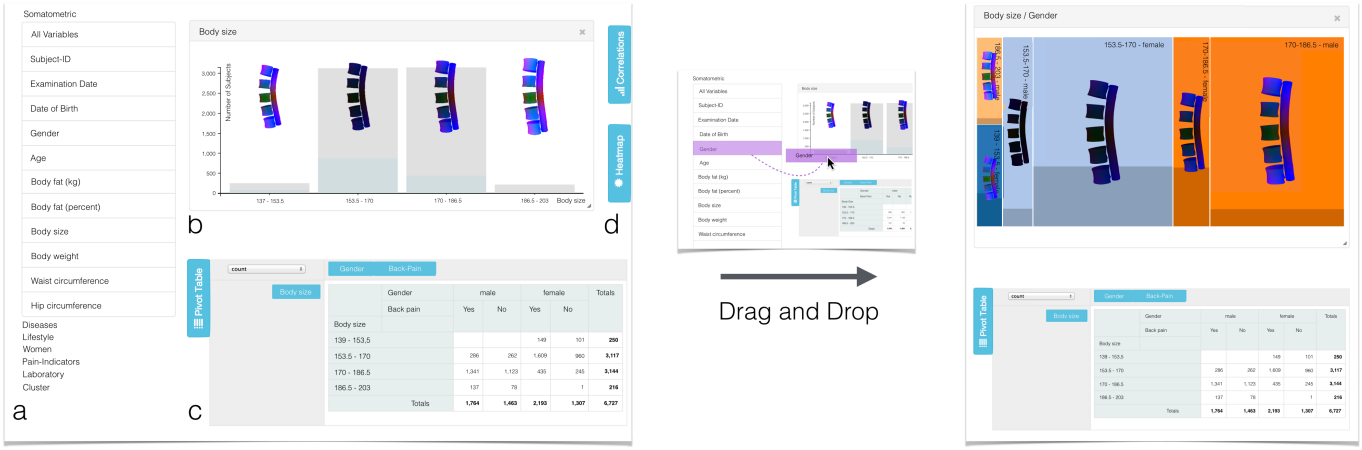


Fig. 3. (Left) Screenshot from the front-end, which is divided as follows: (a) The sidebar containing all variables as well as the groups defined in the analysis process; (b) the canvas area where variables can be added via drag and drop and the visualization is chosen automatically according to the data type; (c) the interactive pivot table showing the exact numbers for each displayed variable combination; (d) buttons to open panes containing the contingency matrix, contingency pane and pivot table. The data displayed is used to analyze the lumbar spine. Variables can be added freely on the canvas via drag and drop. Dropping the *gender* parameter on the already plotted *body size* container creates a mosaic plot combining both variables (right). In a prior step, the user selected all subjects with diagnosed thyroid disorder. These subjects are shown as shade in the visualizations, denoting their share. Subjects between 153.5-170 cm body size are more affected by thyroid disorder (box plot) and are mostly female (mosaic plot). Distance to the mean mesh of subjects with thyroid disorder is encoded as red for x axis, blue y axis and green z axis.

the previously described *IVA* workflow allows for many different ways to analyze the data, we designed the interface as minimalistic as possible, treating the resulting space as *canvas* for the data. We divided the workspace into four major parts, as illustrated in Figure 3 and 4.

- The *sidebar*, which contains all epidemiological variables. The cluster results group variables like categorical variables and are part of the sidebar as well (Fig. 3 a).
- The *canvas* holding all visualizations. Elements can be added, arranged, resized and removed freely (Fig. 3 b).
- The interactive *pivot table* gives detailed numerical information of the variables in the canvas view. This view on the data is familiar to epidemiologists (Fig. 3 c).
- The *contingency view* depicts relations for variables in the canvas in an contingency matrix (Fig. 4) and a *contingency list*.

System Layout. We experimented with several layouts. The initial idea was to make all components freely arrange- and resizable on a large *canvas* area. This idea was soon dropped, since domain experts reported a cluttered workspace, which required a lot of scrolling. The introduction of separate panes for the contingency matrix, pivot table and sidebar, displayed with a mouse click on the corresponding button and sliding on top of the *canvas* was considered more feasible (Figure 3 shows the system with reeled-out pivot table pane). All user-generated visualizations are part of the *canvas* and can be arranged freely.

Sidebar. Only the *sidebar* is visible at system start. It categorizes all variables into different types, such as somatometric (measurements of the human body dimensions), disease- or lifestyle-related, pain indicators and laboratory data (Fig. 3 a). It also contains subject groups defined by automated shape clustering. Groups are treated like dichotomous variables. Variables can be dragged from the sidebar into the canvas area for a *feature localization*, which works as follows.

Adaptive Variable Visualization. The visualization type, inspired by GPLOMS [10, 19], is dynamically chosen based on the variable types and number to allow for *multivariate analysis*. Categorical data are either mapped to bar charts (single variables) or mosaic plots (multiple variables). Figure 3 describes this dynamic adjustment. Continuous data can be visualized using scatter plots (two variables) or parallel coordinates (multiple variables), but in epidemiology, this data type is usually categorized into ordinal groups of *equal size*. Since the number of categories often depends on the hypothesis, the discretization steps can be adapted dynamically. Too many groups potentially generate sparse bins not suited for statistical evaluation. Not enough groups overgeneralize information. Adaptive discretization is an option, but imposes possible overfitting to the data. Conclusions based on statistical relationships derived from groups already biased by variable distribution are heavily influenced by the used discretization. Therefore, we follow the convention to use bins of equal size.

Following Tufte's concept of *small multiples* [38], information derived from the medical image data are incorporated into the plot by including color-coded mean shapes for each manifestation (Figure 3 b). The 3D plots can be navigated using standard mouse input, the camera is synchronized between all views to enable direct comparison. The distance from a group mean shape is mapped to the global mean using color. This allows to assess local shape changes (Fig. 3) and is an important information to the epidemiologist. Until now, epidemiologists were not able to inspect shape differences based on non-image variables. Dropping a variable on an existing plot adapts the visualization dynamically to allow for comparison (Fig. 3 right).

To support *feature localization*, subject groups can be brushed via a double-click on its representative in the visualizations. Holding down the shift key allows to select multiple manifestations. Brushed groups act as reference for the shape visualization, calculating distances based on the mean-shape of the brushed selection. The share of subjects of this subgroup is linked to all other views (Fig. 3 left). If the user selects all female subjects in a visualization of gender distribution, all other displayed meshes are color-coded with their distance to the female mean and the share of female subjects is highlighted in the information visualization.

Pivot Tables. Epidemiologists are used to perform *multivariate analysis* of groups based on table representations. Thus, we decided to introduce an interactive pivot table. These tables clearly convey the subject count in each group (see Figure 3 c). However, they quickly get confusing when they are divided into many subgroups. We tackled this problem by making the order and number of displayed variables adaptable. This also applies to the assignment of row or column variables. Another way to avoid clutter is the user-driven selection of displayed variables. To allow for better comparison with respect to variables, the values of each cell can also be displayed as percentage of the variable represented of either the row or column.

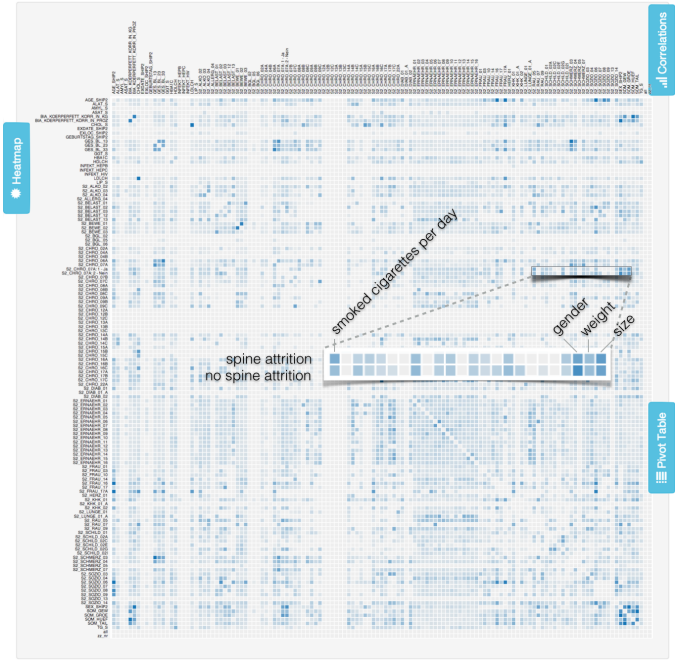


Fig. 4. Contingency matrix of 129 variables (127 data set variables, 2 cluster results) showing 16,641 combinations. Similarity is calculated using the *Cramér's V* contingency value. Color brightness encodes association strength. Moving the mouse over an entry enlarges the variable names for better readability. 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. The contingency pane is not shown here.

Automated Variable Suggestion using a Contingency Matrix. Highlighting potentially interesting associations in the data set is one major benefits of the *IVA*-powered approach and is part of the *multivariate analysis* pattern for analyzing variables outside the shape space. Turkay et al. [39] used the approach to calculate statistical key figures based on the distribution functions of each variable derived from the image data. Since the majority of our data are categorical variables, we have to employ different solutions. The *Cramér's V* contingency coefficient can be used to calculate coherences between categorical variables [8]. It is based on *Pearson's* $X^2$ distribution test [30], which uses contingency tables holding the counts of subjects for all possible manifestations of two variables. *Cramér's 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 number of observations and $k$ is either the row or column count, depending on which one is lower. $V$ yields values between 0, meaning that two variables are completely independent, and 1 indicates that they are the same. *Cramér's V* does not allow to infer the dependency direction.

It shares the same restrictions as *Pearson's* $X^2$. The expected counts in the contingency table have to be larger than five for 80% of the entries and no expected value must be smaller than one [7]. Some manifestations and variable combinations, which are only exposed by small subject groups, cannot be assessed with this technique. They cannot be included into the epidemiological analysis, since statistical validation needs a minimum count to be valid. The contingency matrix highlights correlations between all variables. This aims to highlight variables possibly associated with the focused hypothesis and to trigger new hypotheses. Contingency is visualized using an interactive contingency matrix with association power mapped to color brightness. The distinction whether an association is a confounder or an effect depends on the context defined by the hypothesis and is a decision to be made by the domain expert. The contingency matrix visualization is an *overview visualization*–something the epidemiological community lacks and is in great need of.

Contingency Pane. Dropping a variable into the canvas area adds an entry for each manifestation of it to the *contingency matrix*. Testing sessions revealed that it was tedious to open the matrix every time a new variable is added. As a consequence, the *contingency pane*, a table containing correlating variables for the last added visualizations in descending order of the *Cramér's V* value was added. *Contingency pane* entries can be dragged and dropped into the *canvas area* just like variables in the *sidebar*.

Initialization and Clustering. Using variable suggestion allows to initialize the system with a set of potentially interesting visualizations. After testing and domain expert feedback we dropped this idea. Reasons for this are twofold. Very often, high correlations are obvious, such as gender with menstrual status. Also, we observed that the variables of interest are dependent on the specialization of the domain expert (explained in detail in Section 6).

Subject clustering is triggered automatically as *local investigation* for a variable after it was added to the canvas by the user. A status indicator at the bottom of the screen keeps the user informed about the pending clustering result, since the process can take up to ten seconds. Clustering results are listed in their own category in the *sidbar*.

### 5.2 Implementation


Fig. 5. The front-end solution (left) uses *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, such as calculation of mean shapes or distances, are performed on the server side. Client-server communication is accomplished via the *Websocket* protocol.

In this section, we discuss how we implemented the presented methods using open web standards. To provide a fast communication loop between method development and expert input, we decided to rely on modern web technologies. In addition to the obvious advantages of web technologies, the following aspects are crucial for our work:

- The client-server structure allows for employing heavy computation on a server machine and transferring results to the client.
- Disk-space demanding image data remains on the server and elements can be transferred on demand. High confidentiality standards of the data are met by a password protecting the access.
- Recent developments in *WebGL* applications running in browsers with near-native performance results in many open source libraries, which are well documented and driven by active communities. We use *WebGL* for rendering shape information.

These advantages do not come without drawbacks. Sophisticated libraries/languages, such as the `Visualization Toolkit`[2] or `R`[3] for statistics, are either not available at all or only accessible through complex client-server systems. Therefore, many standard methods had to be written from scratch. The back-end is realized using `NodeJS`[4],

[2]Developed by Kitware Inc; `vtk.org`
[3]Open Source; `r-project.org`
[4]Developed by Joyent Inc, `nodejs.org`

which is based on the Google V8 Javascript runtime environment. Due to its event-driven non-blocking I/O model it is fast and responding even with heavy workload, such as mesh processing. Non-image data for all subjects including the data dictionary is stored in a `JSON` file on the server. Image data are available as raw `DICOM` files. Segmentation masks of anatomical structures are represented as meshes, suited for comparing subjects. The requested data are transmitted when a client connects. The server performs heavy statistical tasks, such as calculation of *Cramér's 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 `Bootstrap`[5] as foundation for the layout and basic UI elements using `HTML5`, `CSS3` and `Javascript`. Information visualizations such as scatter plots and bar charts are created using the popular `Data-Driven Documents (D3.js)` library [3], which works well for attaching data to visible elements like vector graphics and provides powerful transformation and mapping tools. The pivot table implementation uses `PivotTable.js`.[6] `Three.js`[7] allows GPU-accelerated data rendering using `WebGL`. The WebSockets protocol handles the client-server communication. Since our clustering algorithms are written in `MatLab`[8], we had to access them using the `NodeJS` server. We accomplished this by converting it to a parameterized standalone console application, spawned by `NodeJS` on client request. The result is read from the console output and is returned to the client. All parameter-steered console applications can be incorporated in this context.

## 6 Application

This section describes how the presented *IVA* workflow is used in the epidemiological application. We conducted a qualitative evaluation with two domain experts on a data set compiled to analyze lower back pain. This is one of the most common diseases in the Western civilization [40]. Epidemiological analysis of lumbar back pain, such as the work of Harreby et al. [15], is largely focused on non-image information. In comparable studies, only a few shape-related variables are included [25]. Determining risk factors in this area can lead to particularly affected risk groups, prognostic variables for diagnosis and treatment of lumbar back pain and a better understanding of effects of preventive measures, such as occupational health and safety regulations [12]. Characterizing the healthy aging process of the spine is a long-term goal for determining age-normalized probabilities for spine-related diseases by incorporating individual risk factors.

### 6.1 The Lumbar Spine Data Set

There are 127 variables describing diagnosed diseases, lifestyle factors, women-specific factors, pain indicators, laboratory values and somatometric variables for 6,753 subjects (4,420 from `SHIP-Trend-0` and 2,333 from `SHIP-2`). Since data acquisition protocols between these two cohorts are identical, the variables between the two cohorts are comparable. The data contains 30 metric, 7 nominal, 29 ordinal and 62 dichotomous variables. Somatometric variables include measures of the human body, such as body height, weight and body fat percentage as well as gender. These measures are reliable and complete. Other variables, such as pain indicators or lifestyle indicators (e.g. physical activity) are more subjective and less reliable. There are also variables missing for each subject, such as variables building upon each other (e.g. Do you have high blood pressure? Which medication is prescribed against it?). Therefore, some manifestations are sparsely populated, which makes statistical evaluation challenging.

The MRI data was acquired for each subject 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 ($1.1 \times 1.1 \times 4.0$ *mm* voxels) [17].

[5]Developed by Twitter, `getbootstrap.com`

[6]Developed by N. Kruchten, `nicolas.kruchten.com/pivottable`

[7]Originally developed by R. Cabello, `threejs.org`

[8]Owned by The MathWorks, `mathworks.com`

#### 6.1.1 Data Preprocessing

The data processing follows the description in Section 4.1.

**Non-Image Data.** To ensure a fast and easy data access outside of statistical processors like `SPSS`, the data was exported to the `JSON` file format. Since it lacks export methods for data dictionaries, we used `SPSS` to export our data to the `SAS v9+` format, which saves the data labels, and exported the data values as non-labeled `CSV`. A short script combined both data sources to a `JSON` file. The data types had to be transferred manually. Each variable is stored as an object containing the data as array, its data type, a detailed description and data dictionary translating value or error IDs to values. Continuous variables are discretized to allow for *Cramér's V* contingency coefficient assessment. Following epidemiological publications, we set the number of groups to five, to allow for contingency assessment.

**Image Data.** The lumbar spine was detected in the image data using a hierarchical finite element method by Rak et al. [32]. 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 describe an initial model height estimation. The model uses a weighted sum of T1- and T2-weighted MR images to detect the lumbar spine shape. Once registered, it captures information about the shape of the lumbar spine canal as well as the position of the L1-L5 vertebrae [21]. 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 worked with 2,540 tetrahedron models of the lumbar spine. For clustering, we extracted the centerline of the lumbar spine canal, which captures information about lordosis and scoliosis (the medical terms for spine curvature) [21].

#### 6.1.2 Shape Visualization and Clustering

The tetrahedron-based detection model consists of corresponding grid points for each structure instance. This allows to calculate shape distance and similarity. This information is used to calculate mean shapes as described in Section 5. The shape distance between meshes is mapped to color (recall Fig. 3).

Shape-based clustering is carried out via agglomerative hierarchical clustering of the spine canal centerlines (recall Section 6.1.1 and [21]). Since the "correct" number of clusters in a given group is unknown, an estimate is computed by means of the knee/elbow method [33]. The method has proven to produce sound results on a preliminary data set and was able to reproduce textbook knowledge [21].

### 6.2 Participants, Setup and Procedure

Inspired by Lam et al. [23], we conducted an investigation of *Visual Data Analysis and Reasoning* (*VDAR*). This approach aims to characterize the system's ability to explore data, discover knowledge, generate hypotheses and help formulating decisions. Since it is hard to quantify these outcomes, Lam et al. suggest case studies for the *VDAR* by applying the think-aloud protocol to understand the domain expert's observations, inferences and conclusions when using the system.

Our participants are two epidemiological domain experts who also co-authored this publication. HV and KH are physicians with focus on epidemiological research. HV is a specialist in internal medicine (23 years of experience) and head of the SHIP, KH a radiologist (9 years of experience) and responsible for the SHIP MRI data acquisition.

**Setup.** Due to the large geographical distance, the evaluation was done completely web-based. The experts accessed the prototype by entering the weblink into their browser. User input was observed using screen sharing. Communication was enabled via webcam-supported voice over ip. The total setup time including installing the screen sharing application was about five minutes. Video recordings of the sessions allowed a detailed evaluation afterwards.

**Procedure.** At first, we controlled mouse and keyboard of the participants' PC and demonstrated the basic functionalities of the system. As they understood the concepts, we handed over the mouse and

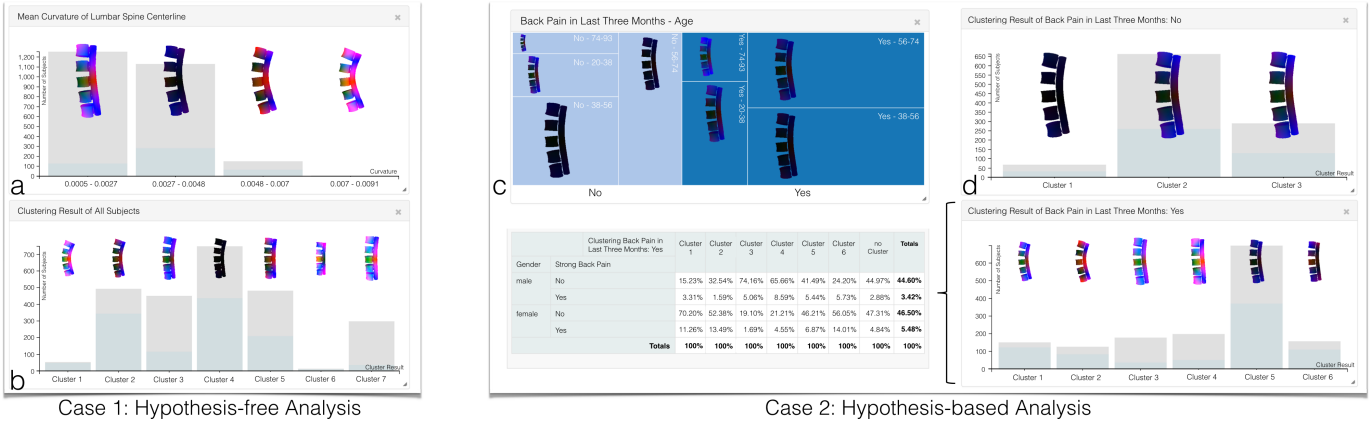


Fig. 6. Various case study results. (a) Mean curvature of lumbar spine canal plotted against the mean shape of 58-74 years old female subjects (light-blue bars). Note the high amount of this subject group relative to the total count in the third group. The last group contains four outliers. (b) Clustering of all subjects yields seven groups, whereas Cluster 4 assembles the mean. The light blue bars indicate the share of females in the group. (c) A mosaic plot mapping age against the dichotomous questionnaire answer to "*Did you experience back pain in the past three months?*". (d) Clustering result of "*Did you experience back pain in the past three months?*" Yes/no with female share in each group. Cluster 1 and 6 for answer "*Yes*" contain mostly women. The pivot table shows how many subjects with strong back pain are in each cluster for answer "*Yes*". Subjects in Cluster 1, 2 and 6 report strong back pain more often than subjects in other clusters.

keyboard control and only observed from this point on. The epidemiologists were given two tasks: one hypothesis-free analysis of the data and one starting with an assumption. For each case we conducted one analysis with each expert.

### 6.3 Case 1: Hypothesis-free Analysis

Analyzing the data set without prior hypothesis requires a starting point giving an overview over the data [34]. With our tool, there are two ways to achieve this. Performing a *multivariate analysis* by viewing the contingency matrix or a shape grouping step using shape-based clustering. The first was chosen by both experts. Before, they were not able to look at all variables in the context of each other. To cite one expert, the contingency matrix "illumunates the data black box", making it possible to look at the data unbiased of assumptions.

Analysis 1. The radiologist (KH) was looking for correlations with shape-related variables in the data, finding that *spine curvature* correlates with *leg pain*, *age*, *body height* and *hormone replacement therapy status*. Due to the dense mapping of information in the contingency matrix, KH suggested to make this visualization full screen.

After this initial overview, KH performed a *multivariate analysis* by introducing variables, such as *age*, *waist circumference*, *weight* or *lumbar spine canal curvature* as bar chart views into the canvas area and selected subgroups to see how they are distributed and if they could observe unusual behavior in the mean shapes. This pointed out problems with the used categorization method splitting numerical variables into equally-sized ordinal bins. If a variable contains outliers, such as *waist circumference* (e.g. by subjects with morbid obesity), this approach leads to sparse categories, making it hard to calculate associations. The proposed expert solution for this is categorization using quantiles/quartiles and is described in detail in Section 6.5.

A *multivariate analysis* using the *Cramér's V* contingency values for subjects with strong lumbar spine curvature showed, that these subjects are primarily females between 58-74 years who also report pain radiating from their back into other body regions Figure 6(a).

Analysis 2. HV also started with a *multivariate analysis* using the contingency matrix to analyze non-image variables, such as age-associated parameters like *income*, *blood fat values* or *number of born children*, but found no associations of interest. Therefore, he applied the *local investigation* pattern by a shape grouping step using shape-based clustering via dragging *All subjects* from the sidebar into the canvas area, triggering the shape clustering (Fig. 6 b).

Cluster 4 represents subjects with average shape. Other shapes differ with respect to size, such as cluster 2, 3, 7, whereas the last one and cluster 5 also represent a more straight spine, which is usual for subjects with larger body size. Cluster 1 and 6 contain outliers, characterized by their unusual shape and small number. To get an overview of the suggested variables, the user opened the contingency pane (not shown here) to perform a *multivariate analysis* by looking at *Cramér's V* contingency values of all clusters, revealing a strong correlation with *gender* and *body size*. Therefore, another *multivariate analysis* was carried out by dragging the *gender* variable to the canvas and selecting all female subjects (Fig. 6 b). Cluster 1 contained primarily female subjects. Contingency values for this cluster revealed correlations with *leg fatigue*, *physically heavy work*, *body weight*, *dyspnoea* and *headache intensity*. Since it is a pain indicator, headache was of special interest and was further investigated by incorporating a pivot table setting *headache intensity* in relation to cluster affiliation. It was found that cluster 1 subjects report heavy headaches more frequently than other subjects.

The experts emphasized the importance of methods providing an overview over the data for hypothesis generation. With the presented *IVA* approaches they were quickly able to confirm medical knowledge and to elaborate new hypotheses. We observed that the domain experts are more likely interested in variables they are familiar with and have personal clinical experience with.

### 6.4 Case 2: Hypothesis-driven Analysis

If the user proposes a hypothesis about a relation between a non-image variable and shape, the workflow slightly differs from the hypothesis-free analysis. The starting point follows the *feature localization* pattern, where a variable of interest is selected by dragging it into the canvas area and viewing the subjects' distribution as well as their shape differences.

Analysis 1. Hypothesis: "*Back pain is associated with age and lumbar spine shape*". To validate this hypothesis, a *feature localization* was performed by combining the dichotomous variable "*Did you experience back pain in the last three months?*" with age in a mosaic plot by dropping both variables on the canvas area (Fig. 6 c). HV was not able to observe the expected effect in the visualization. Reasons for this are twofold. Age influences the lumbar spine shape, while the differences between subjects with and without back pain are small. The major differences seen in the visualization are therefore related to the age variable, masking differences related to the back pain parameter. The second explanation is the commonality of back pain in our society. As seen in Figure 6 (c), subjects reporting back pain are the

majority, which makes it difficult to extract parameters that reliably describe back pain. A *multivariate analysis* using the contingency table showed a strong association between *back pain* with both, *gender* and *body height*. *Body height* was explained as a confounder for *gender*, since female subjects on average are smaller than male subjects. The analysis solely based on shape-accentuated body height differences in *gender*, which clouded the differences of *back pain*.

The epidemiologists pointed out that they would like to see a more intuitive and fast way to select subgroups based on different variables to make full use of the analysis capabilities, as discussed in Section 6.5.

Analysis 2. Hypothesis: "*Back pain is related to lumbar spine deformation*". The previously discussed analysis questions the suitability of the lumbar spine segmentation for analyzing back pain, leading to this analysis. Therefore, the dichotomous variable "*Did you experience back pain in the past three months?*" is dropped into the canvas area. Figure 6 (d) shows the results of the automatically triggered shape-based clustering for subjects with and without back pain. The clustering algorithm finds only three homogenous clusters close to the global mean shape for subjects reporting no back pain. The cluster analysis for back pain yields diverse clusters with pathological shape classes. Cluster 5 represents most of the subjects and is similar to the global mean shape. Cluster 1 and 2 present a *hyperlordosis*, a strong curvature of the lumbar spine, while Cluster 3 and 4 present a more straight shape. A *multivariate analysis* using the pivot tables put gender and strong back pain in context to cluster affiliation (Fig. 6 d). It shows that subjects in Cluster 1, 2 and 6 reported strong back pain, while at the same time they also have a considerably higher share of females. To check for unusual correlations, the expert used the *Cramér's V* contingency table. It depicted strong associations with *body fat*, *body weight* and *blood pressure* (Cluster 1) *alcohol consumption* and *attentiveness disorder* (Cluster 2), and *amount of sleep* (Cluster 6). For the experts, these observations are a starting point for a number of new hypotheses about possible relationships, for example the association between overweight and Cluster 1.

In summary, it can be stated that the hypothesis-driven analysis leads to hypothesis generation by design of the framework. It is not suited and intended to statistically validate hypotheses like statistical processors. It rather triggers the analysis of potentially associated variables with a pathology of interest.

### 6.5 Further Feedback and Lessons Learned

Both domain experts rated the *IVA* approach positively. KH emphasized the way the image data are included into information visualizations, which comes much more natural to her due to her background in radiology. Great potential is also seen in communicating insights efficiently using the presented visualizations.

**Multivariate analysis is most important for hypothesis generation.** Both experts emphasized the potential of the *multivariate analysis* capabilities of the contingency matrix for gaining insight into a large amount of variables simultaneously. It is also useful to verify established but still controversial risk factors, such as the metabolic syndrome for coronary heart disease and whether the data set provides more suitable risk factors. Creating contingency matrices for subgroups, such as different age bins can help to characterize the aging process by deriving age-specific risk factors. *Multivariate analysis* can be improved by more ways of brushing the data as well as creating subgroups for comparison as a result of the hypothesis-driven analysis case. Too small variable ranges yielding sparse groups could hinder the calculation of statistical resilient measures, since they require a minimum amount of subjects exhibiting the selected variable ranges.

**Segmentation quality is crucial.** KH pointed out the unusual strong similarity of the L3 vertebrae throughout the population. The medical explanation is that it represents an angular point of curvature of the lumbar spine. A second explanation is the use of the L3 vertebra as initialization point of the lumbar spine model. The experts also emphasized that associations related to shape strongly depend on the segmentation quality. The lumbar spine model used in this case study captures deformation of the spine canal well, but lacks precise definition in vertebrae height and shape. Since deformation of the spine canal is the last stage of pathological lumbar spine deformation and is preceded by vertebrae deformation, the system would strongly benefit from more precise segmentation results capturing these prior changes. For the visual comparison, KH proposed an abstraction of the representation into landmarks, such as centers of the vertebrae and cardinal points of the lumbar spine canal.

**Usage of different categorizations depending on expected outcome.** Categorizing numerical variables into equal groups possibly creates sparse categories due to outliers.These outliers are only of high interest for finding pathological subjects. The experts therefore suggested two modes of the tool. The outlier mode still creates categories of equal size, producing sparse categories for outliers. Balanced categories are created in the second mode, which uses quartiles or quintiles to set borders between categories.

**Web technologies are well suited for rapid feedback.** The web-based approach for both implementing the prototype and getting feedback via voice over ip conference calls worked very well. Since the software does not need to be compiled, small changes can even be made on the fly during a testing session. The large data base associated with image-based epidemiological data remains on the server machine and has not to be moved tediously using external hard disks. This approach is well suited for the *VDAR* approach to assess user thought processes using the think-aloud technique.

## 7 Summary and Conclusion

We presented an *IVA* framework for the analysis of image-centric epidemiological data. Hence, the framework allows both for hypothesis-driven analysis and hypothesis generation. The visualization of multivariate data using multiple connected 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 epidemiologists while embracing the power of interactive adjustment of the shown variables. The automatic suggestion of correlations using contingency methods, such as *Cramér's V* triggers *hypothesis generation* by highlighting variables potentially overlooked by the experts. Shape-based clustering assesses the variability of an anatomical structure in the context of non-image variables, such as disease indicators or lifestyle factors.

Epidemiologists are for the first time able to assess shape information of the lumbar spine and its influence to diseases. Findings from analyzing lumbar back pain using the *IVA* approach range from deriving shape-based groups of subjects to detailed descriptions of variables potentially associated with the disease, such as waist circumference, alcohol consumption and attentiveness disorder. A number of improvements is left open for future work, such as shape brushing methods to intuitively query subjects using image data or the inclusion of more statistical methods and views that are familiar to the epidemiologists (odds ratios, box plots).

As the number of image-centric cohort studies, participating subjects, gathered variables and imaging modalities rises, and advances towards comparability between cohort studies are made, the gap between data complexity and analyzability 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 that web technologies pave the way to analyze this data in a convenient way.


## Acknowledgments

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grant no. 03ZIK012), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania. Whole-body MR imaging was supported by a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg-Vorpommern. The University of Greifswald is a member of the Centre of Knowledge Interchange program of the Siemens AG. This work was supported by the DFG Priority Program 1335: Scalable Visual Analytics. We thank Marko Rak and Klaus Toennies for providing the image detection data.

## REFERENCES


[1] P. Angelelli, S. Oeltze, C. Turkay, J. Haasz, E. Hodneland, A. Lundervold, B. Preim, and H. Hauser. Interactive visual analysis of heterogeneous cohort study data. *IEEE Computer Graphics and Applications*, 2014. in print.

[2] J. Blaas, C. Botha, and F. Post. Interactive visualization of multi-field medical data using linked physical and feature-space views. *Proc. 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] 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] 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.

[7] W. G. Cochran. The $\chi 2$ test of goodness of fit. *The Annals of Mathematical Statistics*, pages 315–345, 1952.

[8] H. Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 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] J. W. Emerson, W. A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H. Wickham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91, 2013.

[11] 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.

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

[13] 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.

[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, 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.

[17] 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.

[18] M. Hermann, A. C. Schunke, T. Schultz, and R. Klein. A visual analytics approach to study anatomic covariation. In *IEEE PacificVis 2014*, pages 161–168, 2014.

[19] 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.

[20] 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.

[21] 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 *VMV 2013 - Vision, Modeling, Visualization*, pages 121–128, 2013.

[22] Z. Konyha, K. Matkovic, and H. Hauser. Interactive visual analysis in engineering: A survey. *Proc. of Spring Conference on Computer Graphics (SCCG 2009)*, pages 31–38, 2009.

[23] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[24] T. Lammarsch, W. Aigner, A. Bertone, S. Miksch, and A. Rind. Towards a concept how the structure of time can support the visual analytics process. In *Proc. of the Int. Workshop Visual Analytics*, pages 9–12, 2011.

[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. Völzke, J.-P. Kühn, 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, W. Freiler, R. Hillert, H. Doleisch, B. Preim, and W. Schubert. Interactive, graph-based visual analysis of high-dimensional, multi-parameter fluorescence microscopy data in toponomics. *IEEE Trans. on Visualization and Computer Graphics*, 17(12):1882–1891, 2011.

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

[30] 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.

[31] 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. in print.

[32] 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, 2013.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] E. Tufte. *The visual display of quantitative information*. 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] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proc. of the Working Conference on Advanced Visual Interfaces*, pages 110–119. ACM, 2000.

[43] C. Weaver. Cross-filtered views for multidimensional visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):192–204, 2010.

[44] 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.