# Interactive Visual Analysis of Image-Centric Cohort Study Data

TBA

**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. Using well established statistical methods the data is hypothesis driven analyzed to find statistically significant variable correlations (*interactions*). Modern cohort studies also incorporate medical image data. Analyzing these data requires image segmentation, extraction of key figures and shape based subject grouping.

We propose a Interactive Visual Analysis approach that enables epidemiologists to examine both image-based as well as sociodemographic and medical attribute data. It allows for both hypothesis validation approaches as well as hypothesis generation by incorporating data mining methods. Adaptive linked information visualization views and 3d-shape renderings are combined with epidemiological techniques. Similarity measures between data variables are used to compute interesting changes in variable interactions for the current variable selection. Shape based grouping of subjects is facilitated using clustering techniques which operate on surface meshes extracted from the image segmentation.

**Index Terms**—Interactive Visual Analytics, Epidemiology

## 1 Introduction

Epidemiology aims to characterize health and disease by determining risk factors. Clinical problems and questions answered using epidemiological methods comprise diagnosis accuracy, disease frequency, risk factors, disease prognosis, effectiveness of treatments or preventions and cause of diseases [9]. Observations made by clinicians in the daily routine are translated into hypothesis. These are used to determine environmental and lifestyle factors are as well as medical attributes which may influence a condition of interest. The data variables necessary are gathered using structured interviews and clinical examinations. Statistical methods like regression analysis aim aim to check the attribute list for plausibility.

Longitudinal population-based studies like the Study of Health in Pomerania [28] aim to gather as much information as possible about a defined sample of people (a *cohort*). The sample is drawn randomized to avoid selection bias which prohibits statements based on statistical correlations in the cohort from being inferred to the whole population (*external validity*) [9]. Also a information bias needs to be avoided by strictly standardizing the data acquisition. Statistical correlations are also prone to *confounding*, meaning that two factors are dependent and therefore should be normalized with respect to each other. When for example one investigates risk factors for prostate cancer in male subjects, the outcome is strongly dependent on the age. Therefore results need to be age adjusted to be comparable. Confounding variables, are often not obvious at all and characterizing them is already an epidemiological result.

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

We propose a Interactive Visual Analysis approach [26] to provide a way to analyze both image- and non-image data. Visual queries and direct feedback of Visual Analytics systems allow for a fast exploration of the data space. Intended as a extension to the well established epidemiological tools it provides a way to rapidly validate hypothesis as well as trigger hypothesis generation using Data Mining methods such as clustering. In order to characterize the healthy aging process we aim to determine changes for subjects which indicate unusual pathological changes.

Our contributions are:

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

## 2 Medical and Technical Background

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

### 2.1 Epidemiological Workflow

Since it is a diverse science, many different experts work at epidemiological studies, ranging from specialized doctors, medical computer scientists with focus on biometrics to statisticians. Epidemiologists follow a strict workflow mainly driven by statistic tools to validate hypothesis about disease specific risk factors. Following Thew and colleagues publication on this matter, the workflow can be characterized as follows.

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

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

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

The workflow is shown in Figure 4 (a).

Reproducibility of results is an key requirement. 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, disallowing inference between acquisition cycles. This underlines the high quality standards to methods processing the data, whether to extract additional parameters or gain insight. 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. Statistics tools such as SPSS and STATA play a major role for analyzing epidemiological data. Graphic data representation is largely used to present results rather than gaining insight.

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

### 2.2 Epidemiological Data

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

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

Image acquisition. Imaging techniques emitting an hazardous amount of radiation for the subject are not suited for ethical reasons. MRI data is more expensive to obtain than CT data but does not affect the subjects health and is therefore the main method for collecting cohort study imaging data. The image quality is a tradeoff between accuracy and affordability [22]. This often yields image resolutions inferior to those of clinical day-to-day practice, which makes their analysis more challenging.

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

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

Starting 1997 with a cohort consisting of 4.308 subjects this cohort study located in northern Germany aims to characterize health and disease in the widest range possible [28]. Data is collected diseases-independent. This allows the data set to be queried regarding many different diseases and conditions. Subjects were examined in a 5-year time span, continuously adding new parameters including MRI scans in the last iteration of 2012. The MRI protocol features a rich number of different sequences. Also for women, breast MRI scans were acquired. A second cohort `SHIP-Trend` was established in 2008 to acquire data about a younger population. The protocols for analyzing the subjects between `SHIP` and `SHIP-Trend` remained the same, making them comparable. The overall examination time for each person attending the study is two days.

## 3 Prior and Related Work

Einfuehren von Helwigs Terminologie?

Designing a visualization which communicates all aspects of the data equally is challenging. Given the number of features of epidemiological data sets and their different manifestations, it is often a good solution to combine the strength of different visualization techniques in a unified system [2, 19]. Data mining tools like the Principal Component Analysis are able to reduce the dimension by extracting most expressive components, but make the influence of each variable hard to determine.

The work of Turkay and colleagues is closest to ours albeit our focus on processing medical image data and variables with categorical manifestations [27]. Investigating Data on an norwegian aging study their methods aim to amplify a hypothesis generation process. Statistical measures of metric variables such as mean, standard deviation, skewness, or inter-quartile range are used to create *dimension plots*. These transform dimensions into data points and make them comparable with respect to the derived descriptive measures. This not only allows for comparing all continuous variables in a single plot but make their distribution change comprehensible. This requires a good descriptive measure which captures the kind of change the user is interested in or which reflects unexpected data behavior. The technique was applied to variables generated by segmenting the brain into 45 parts and measure the voxel number, volume and properties of the intensity values. The method is strongly dependent on the descriptive measures of the epidemiological factors. Hypothesis based on observations of changes in these plots may impose over-fitting to the data because the measure highlights only subsets of statistical changes. Our approach sticks more to the information extracted from the segmented image data and derive variable interaction with non-image epidemiological factors.

Gresh and colleagues proposed `WEAVE`, one of the first systems which analyzed medical image and non-image data using linked views [12]. 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 [1]. This approach already incorporated Data Mining approaches like dividing the data space using a k-nearest-neighbor technique and Principal Component Analysis. Steenwijk and colleagues employ a relational database to organize the data to visualize subject data using linked views like parallel coordinates, scatterplots and time plots [25]. 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 [29]. Their understanding of a cohort differs from the understanding of the term in an epidemiolgical context.

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

Klemm and colleagues used lumbar spine variabilities based on an semi-automatic shape-detection algorithm of 490 participants of the SHIP-2 [17]. Hierarchical agglomerative clustering divided the population into shape-related groups. As proof of concept a relation between size of the segmented shape and measured size of the subjects was measured and behaved as expected. This work focuses on incorporating these 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 is was found that k-Prototypes and DBSCAN is appropriate in the epidemiological context but is strongly dependent on the chosen variables and distance measure [16].

Generalized Pairs Plots (GPLOMS) are an information visualization technique that allows for heterogenous variables to be pairwise compared using appropriate plots in a plot matrix grouped by type [15]. This technique is also useful to gain an overview over numerous variables and their distributions. It uses histograms, bar charts, scatterplots and heat maps to visualize the different variable combinations with regard to their type. Brushing capabilities allow for brushing and linking as well as filtering, but has limitations like making only one category brushable at a time. We applied this technique to our data and it shows promising potential for simultaneously visualizing many different variable but does not fit in the scope of this paper. The inspiration on the chosen visualization techniques stems from this publication. A similar approach was taken by Dai and colleagues for risk factor exploration as they also incorporate choropleth maps of epidemiological factors (e.g. mortality rates in a region) with parallel coordinates, bar charts and scatterplots with integrated regression lines [7]. From their findings regarding the interaction of cancer-related socio-demographic factors are drawn in a *Concept Map* where related factors are connected via graph-edges.

Chui and colleagues visualized interactions in time-dependant epidemiological data using time-series plots highlighting risk factors differences in age and gender [6].

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 user favor spherical glyph representations over deformation grids and likelihood volumes [4]. The distribution of shapes in a space derived from a principal component analysis is plotted by Busking and Colleagues in a 2D-projected plane of the space [3]. We incorporate the idea of combining 3D-Shape rendering with information visualization techniques. Differences between structures are highlighted using color mapping of the difference to the mean shape, but is rather hard to recognize due to small renderings of each subject in the shape-space. Via mesh morphing interpolated views can be created by the user in a separate view as well as comparisons in contour-view. Applying our data sets to this technique yielded a cluttered shape space due to the many subjects. The data needs to abstracted or summarized in order to work in this context. In order to detect local deformation changes, Hermann and colleagues investigating shape related difference by letting the user specify a deformation of interest and showing corresponding changes in the shape using covariance tensors [14]. This method allowed for rapid hypothesis validation and was able to reproduce textbook knowledge. By plotting p-values in ventricle surfaces, Chou and colleagues were able to map disease-associated values directly on a 3D tissue representations [5]. This requires a geographic colocation of associated features.

The strength of the Interactive Visual Analysis approach described in the next section is it's versatility with respect to the application field [19]. Oeltze and colleagues combined a linked view representation of results from a statistical analysis with feature localizations of the human blood flow with the goal of its evaluation [20]. While we take similar steps when analyzing the data like employing statistical tests, our data is mostly independent from the medical image data and is not describing it–except the variables derived the segmentation model itself.

## 4 Interactive Visual Analytics in Cohort Study Data

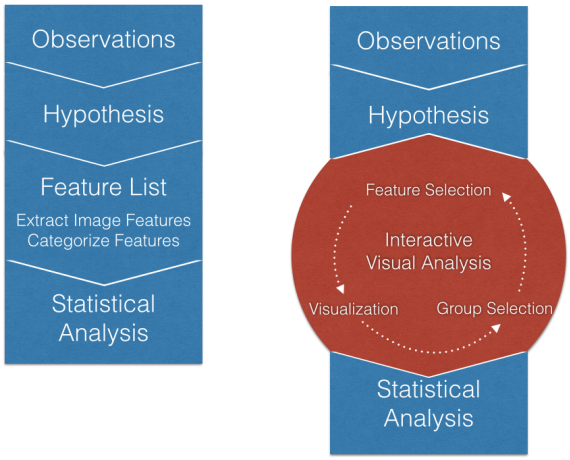


Fig. 1. ToDo Abbildung noch nicht final. Visual Analytics systems are able to complement parts of the epidemiological workflow, not replace it. The appropriate combination of statistical- and interactive driven analysis shows promising potential to unveil the information in the data. (a) shows the standard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

As described in subsection 2.1, the epidemiolgical workflow is a strict sequence of steps taken by domain experts and needs to be reproducible and comprise statistical integrity. Figure 4 (a) describes this workflow as consecutive series of steps. The workflow we propose by introducing the *IVA* principle into the epidemiological application domain does not aim to replace the existing workflow but to complement its weaknesses. In the current state the workflow treats the data like a black box. A list of features describing the hypothesis is compiled and analyzed using statistical tests. The resulting value decides whether the data supports the hypothesis or not. It would be possible that there actually are features of the data set which support the hypothesis by discriminating the population in the expected way, but with this approach they are not highlighted in any way. This becomes even more imminent when the workflow is adapted to the analysis of the medical image data. Domain experts annotate tediously landmarks which allow to derive metrics like distances which are then handled like other features and analyzed using the same set of statistical tools. Not only does this leave out the majority of the information in the medical image data by abstracting it to single values, it is easily possible that information left out would heavily influence the result. Considering more complex parts of the data would make those results more trustworthy and also could identify possible anatomical confounders–an epidemiolgical research result in itself. Statistical tests check for validity of the number but not for their completeness or plausibility!

*IVA* tries to illuminate the black box by making the domain experts part of the feature list selection process. Figure 4 (b) highlights the iterative process as part of the epidemiolgical workflow. Note that it also aims to project back into the hypothesis formulation step to amplify hypothesis generation This has to be handled with care since overfitting of expectations to the available data is an imminent danger as described by Turkay and colleagues [27].

### 4.1 Image Centric Cohort Study Data in Interactive Visual Analytics Context

In the *IVA* context, data is divided into two major view types. The human body exposing shape information for the *physical view*, also referred as the *independent variable* [21]. This information space is usually displayed via volume rendering techniques [20]. *Dependent variables* in the epidemiological context can be divided twofold:

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

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

### 4.2 Data Preprocessing

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

Processing the image data associated to each subject consists for the most part of information extraction about a structure of interest. This is either done manually by experts setting land marks (sometimes supported by algorithms connecting the land marks like Graph-Cuts) or by a (semi)-automatic detection, registration and segmentation. Algorithms applied to the data do not only have to deal with a large inter-subject variability of the structure of interest but also needs to be reproducible [22]. Model based approaches have shown to be effective for this task [10, 11, 23]. If a segmentation yields only binary masks separating the structures, algorithms like Growing and Adaptive Shapes can be applied creating a surface grid where each point is comparable throughout the population [8]. Intensity based comparison can be achieved using rigid image registration, but model based results however are preferable [18]. Comparison based on grey values is usually carried out to to measuring the quantity of fat, water, and–application specific–iron content (liver) or distribution of grey and white brain tissue.

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

### 4.3 IVA Patterns

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

#### 4.3.1 Local Investigation

This pattern projects information from image space to the range perspective. As opposing to other *IVA* application domains this step is more complicated in the epidemiological context. Shape information can not be brushed by incorporating ROI-selections but rather has to employ techniques that specify local deformation changes [14] or subjects that belong to a shape class. Methods available for *feature selection* strongly depend on the type of registration that was applied to extract the tissue of interest. Model based segmentations or masks yield data structures capable of calculating mean shapes and distances between individuals or subject groups. Feature selection is also possible by applying clustering algorithms in order to get shape-groups [17]. These algorithms can be used to investigate interactions between shape-groups and other non-image based variables. Another application is the outlier analysis. Outliers can indicate segmentation errors or an outstanding group of individuals who may share a pathology.

#### 4.3.2 Feature Localization

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

#### 4.3.3 Multivariate Analysis

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

## 5 Interaction- and Visualization Techniques

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

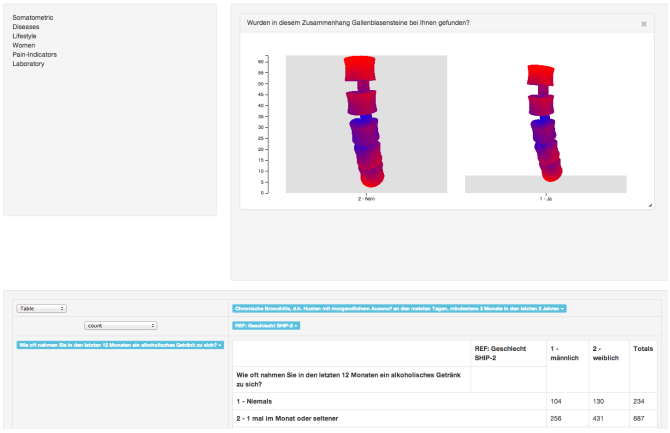


Fig. 2. *Image is not final* Screenshot from the front-end which is divided as follows: (a) Sidebar which contains all features as well as the groups defined in the analysis process; (b) Canvas area where features can be added via drag and drop and the visualization chosen automatically according to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

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

### 5.1 Structure and Workflow

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

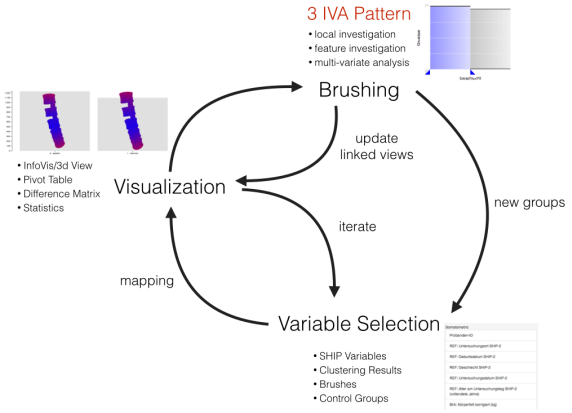


Fig. 3. Detailed Version of the *IVA Loop*. Usually starting with a selection of Variable of interest (user-driven or via data mining techniques), the data is mapped using a visualization technique appropriate for the selected data types. The data is visualized in the independent and dependent domain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

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

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

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

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

### 5.2 Sidebar

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

### 5.3 Adaptive Feature Visualization

Inspired be the previously discussed *GPloms* [15] the visualization type is chosen dynamically based on the variable types and number of visualizations which need to be displayed. If possible the medical image data is directly included into the plot as well by including mean shapes for each manifestation (Figure 2 (b)). The 3D-plots can be navigated using standard mouse inputs and the camera is synced so that a direct comparison is given. If a feature is dropped on an existing plot, the visualization changes dynamically to properly make them comparable. Each plot can be brushed using widgets. It is able to duplicate brushes in order to create new groups which are evenly spread out. A use case for this is when a continuous feature has to be divided into even groups.

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

### 5.4 Pivot Tables

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

### 5.5 Automated Feature Suggestion

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 [27]. Since the majority of our data are categorical features, we have to employ different solutions. A solution for this problem are odds ratios, which are a standard statistical tool for stating relations between features. Odds ratios can only be calculated for $2 \times 2$ contingency tables which usually represent the presence of a condition in a population divided by a characteristic (e.g. male/female subjects with presence or absence of back pain). To calculate odds ratios for variables with more than two manifestations, we calculate local odds ratios for each possible combination, yielding a matrix for each feature combination [24]. When the subjects are divided into groups and the calculation is carried out again for all feature combination, the difference in sum of the odds weighted with the number of manifestations per feature can be used to indicate if the feature combination yields a difference. These difference are then highlighted in a separate tab of the side bar "Interactions". Interesting interactions then can be assessed creating a linked view using the standard drag and drop workflow.

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

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

### 5.6 Implementation

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

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

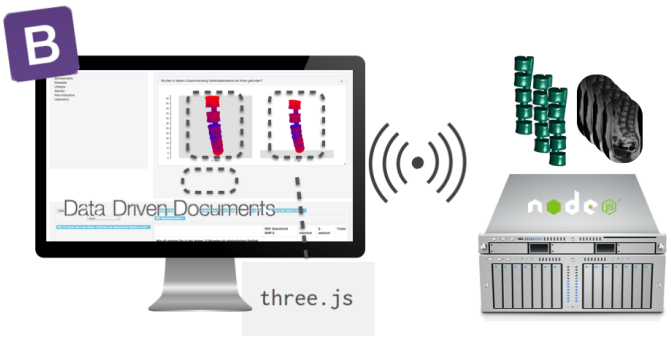


Fig. 4. *Image is not final.* The front-end solution (left) uses state of the art web technologies like HTML5/CSS3, WebGL and SVG to display the data. The NodeJS based back-end (right) stores all image- and non-image data and transfers it to connected clients. All computation heavy operations like calculation 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.

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

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

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

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

## 6 Application

### 6.1 The Spine Dataset

- Describe steps from gathering Information from the raw image files (segmentation, abstraction, visualization)
- Problem of sparse differences - Visualization has to be more abstract to emphasize the differences.
- Input of Epidemiologists goes here!

**From VMV'13 Paper**

All whole-body MRI scans were acquired on a 1.5 Tesla scanner (Magnetom Avanto; Siemens Medical Solutions, Erlangen, Germany) by four trained technicians in a standardized way. Subjects were placed in the supine position. Five phased-array surface coils were placed to the head, neck, abdomen, pelvis, and lower extremities for whole-body imaging. The spine coil is embedded in the patient table. 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). First, both sequences were placed over the cervical and upper thoracic spine. Then, they were placed over the lower thoracic and lumbar spine. The MRI software automatically composed a whole spine sequence from the two T1-weighted and T2-weighted sequences [13]. We were provided with 490 data sets.

The model is placed in the scene using an empirically chosen initialization point. The force acting on the model stems from aggregation of loads, which are derived from a potential field resulting from a weighted sum of the T1- and T2-weighted MRI images, see [23]. After detecting all spines, we register the models because in a later clustering step we only want to capture the local deformation of the lumbar spine, not different locations in world space. The models are registered using the Kabsch Algorithm, which is designed to minimize the root mean squared deviation between paired sets of points. The model-based detection captures information about the spine canal curvature as well as the alignment of the vertebrae. It is not meant to capture information about vertebrae deformation and differences in spine canal extent.

## 7 Summary and Conclusion

## References

[1] 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.

[2] A. Buja, J. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In *Visualization, 1991. Visualization '91, Proceedings., IEEE Conference on*, pages 156–163, 419, Oct 1991.

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

[4] 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.

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

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

[8] 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.

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

[10] 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.

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

[12] 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 *Visualization 2000. Proceedings*, pages 489–492. IEEE Computer Society Press, 2000.

[13] 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.

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

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

[16] 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 fr die Medizin (BVM)*, page in print, 2014.

[17] 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.

[18] 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.

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

[20] 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, 28. October - 1. November 2007.

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

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.

[28] 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.

[29] Z. Zhang, D. Gotz, and A. Perer. Interactive visual patient cohort analysis. In *IEEE VAHC Workshop*, 2012.