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 data variables describing the medical condition as well as demographic and lifestyle factors of a subject. 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 Analytics approach that enables epidemiologists to examine both image-based as well as sociodemographic and medical attribute data. It allows for both classical 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 hierarchical agglomerative clustering. (Remove Clustering reference?)

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 [**?**]. 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 are believed to influence a condition of interest. The data variables necessary are gathered using 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 [**?**] 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 prohibit statements based on statistical correlations in the cohort to be inferred to the whole population. 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 influence each other 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, however, 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 segment it. Manual segmentation via 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 other epidemiological factors require techniques which reach beyond standard statistical methods.
We propose a Interactive Visual Analysis approach 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. **Characterization of the healthy aging process! Which change indicates a unusual pathological change?**

Our contributions are:

- Applying the Interactive Visual Analysis approach to the epidemiological problem domain by characterizing special affordances of this context.
- Provide an overview over the workflow for analyzing cohort study data to gain insigth into the large 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 in a Web Framework based on WebGL, D3js and Nodejs as backend.

2 Medical and Technical Background

Wer ist an epidemiologischen Studien beteiligt?
Arzte (Facharzt fur offentliches Gesundheitswesen, Gene@ker)
Medizinische Informa@ker mit Fokus auf Biometrie und Sta@s@k
Bei klinischen Studien: Arzte des entsprechenden Fachs

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. ToDo Define Epidemiological Outcome

2.1 Epidemiological Workflow

Epidemiologists follow a strict workflow mainly driven by statistic tools to validate hypothesis. 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. 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.

Reproducibility of results is an epidemiological key requirement. Longitudinal epidemiological studies require the acquired attributes

• *Otto-v.-Guericke-University Magdeburg*

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.

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 a expressed through the evaluation of p values which indicate statistical significance. Statistics tools such as SPSS and STATA play a major role for analyzing epidemiolgical data. Graphic data representation is largely used to show results rather than gaining insight.
Group subjects using epidemiological factors is essential in order to make statements about their statitical power. Grouping is carried out hypothesis driven. Age for example is also divided into groups (*binned*) 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, questionaire answers, and image data yield complex data spaces 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 as CT data but does not affect the subjects health and is therefore the main method for collecting cohort study imaging data. The quality of medical image data acquired for cohort studies is a tradeoff between accuracy and affordability [**?**]. This often yields image resolutions inferior to those of clinical day-to-day practice, which makes their analysis more challenging.

Image analysis. When analyzing medical image data there have decisions to be made on how they are *compared* and *quantified*. Segmentations masks describing all parts of the 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 [**?**]. Data is collected independent of diseases in mind. 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

- VMV Paper
- Analysis of data in the Rotterdam study
- VA in Epidemiology Part of Bernhards Paper
- Commercial visual Analytics Systems
- GPloms

Unterlagen Wissenschaftliche Projekt

4 Interactive Visual Analytics in Cohort Study Data

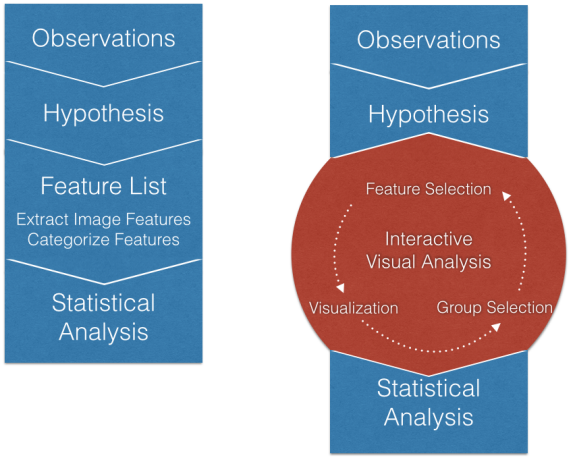

Fig. 1. Workflow Comparison

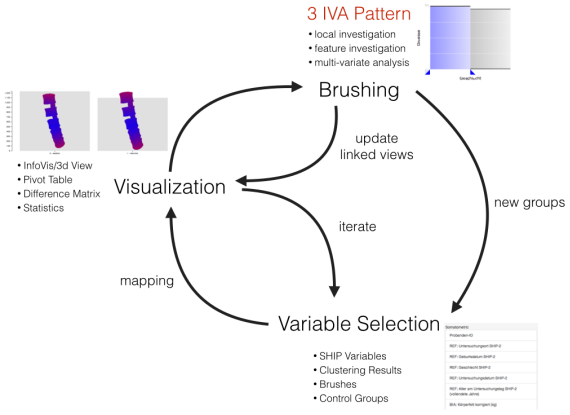

Fig. 2. Interaction Overview

According to Steffens Terminology in Interactive Visual Analysis of Perfusion Data

Object Space
- Medical Image Data / Spine Segmentations

Attribute Space
- SHIP-Variables
- Derived visualizations

? Where to incorporate Levels of IVA

4.1 Feature Selection

- Automatic using Clustering
- Automatic using Statistical Recommendations
- Via User Input

4.2 Brushing the information space

- derive new groups which serve as input for the feature selection

4.2.1 Feature Localization

- Projection from Image Space to Attribute data. This is more difficult in our application, because we currently only brush on derived features.
- Clustering in Image Space yields Groups which can be analyzed using the attribute space
- Create a Pipeline Overview over different Levels of different IVA Patterns and Stages

4.2.2 Local Investigation

- Selection of Information using Bar Charts, Scatterplots or Parallel Coordinates which projects the selection into the Object space
- this selection is for categorical data already given implicitly bei projecting the 3D-View onto the Bar Charts/Mosaik Plots!
- this aims to locale features of the data!
- Gain Information about SHIP-Variables by putting them into the context of each other. The Pivot Table allows for direct numerical analysis, while the information visualizations allow for better insight of the combination

4.2.3 Multivariate Analysis

- Selection of Elements in Scatterplots, Barcharts or Parallel Coordinates Views - observe how selection changes another view - this allows for multivariate analysis
- Becker, R.A., Cleveland, W.S.: Brushing scatterplots. Technometrics 29(2) (1987)
- Wang Baldonado, M.Q., Woodruff, A., Kuchinsky, A.: Guidelines for using multiple views in information visualization. In: AVI 00: Proceedings of the working conference on Ad- vanced visual interfaces, pp. 110119. ACM Press, New York, NY, USA (2000). DOI http://doi.acm.org/10.1145/345513.345271 - By brushing individual parameters or create new binnings of parameter it is possible to see how they change in coordinated views. This is already implemented in the Cargo framework
- by creating comparative 3d Visualizations it is possible to assess the influence of non-image parameters to the visual space.

4.3 Implementation

Diagram of used Technologies

5 Application

5.1 The Spine Dataset

- Describe steps from gathering Information from the raw image files (segmentation, abstraction, visualization)
- 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 [**?**]. 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 [**?**]. 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.

6 Summary and Conclusion

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.