Regression Cube Analysis of Cohort Study Data

Paul Klemm, Kai Lawonn, Sylvia Glaßer, Uli Niemann, Katrin Hegenscheid, Henry Völzke, Bernard Preim

Abstract—Epidemiological studies comprise heterogeneous data about a subject group (a *cohort*) to define disease-specific risk factors. These data contain information (*features*) about a subjects' lifestyle, medical conditions and also medical image data. These features are analyzed using statistical regression analysis to identify feature combinations indicating a disease (the *target feature*). Although there is a strong demand for overview visualizations of a whole data set towards a target feature, no suitable tool is available for epidemiological researchers.

We propose an analysis approach of epidemiological data sets by incorporating all features in an exhaustive regression-based analysis. This approach combines all *independent features* with respect to a *target feature* and provides a visualization, which reveals insight into the data by highlighting relationships. A 3D-visualization of all combinations of two to three independent features towards a target acts as an overview of the whole data set, the *Regression Cube*. Slicing through the *Regression Cube* allows for the detailed analysis of features towards the target disease. Expert knowledge about disease-specific hypotheses can be included into the analysis by adjusting the regression model formulas. Furthermore, the influences of features can be assessed using a difference view comparing different calculation results. We applied our *Regression Cube* method to a hepatic steatosis data set to reproduce results from a data-mining driven analysis. A qualitative analysis with three domain experts was conducted on a breast fat data set. We were able to derive new hypotheses about relations between breast fat density and breast lesions towards breast cancer. With our work, we present a visual overview of epidemiological data with the *Regression Cube*, which allows for the first time a interactive regression-based analysis of large feature sets with respect to a disease.

Index Terms—Interactive Visual Analysis, Epidemiology, Breast Cancer, Hepatic Steatosis

1 Introduction

Epidemiology aims to characterize health and disease conditions in defined populations (*Cohorts*). Insights about risk factors allow to characterize disease-specific high-risk groups and act as important diagnostic key figures [4]. They can also be used to give recommendations regarding a healthy lifestyle and provide information about wide spread diseases. In the standard workflow, physicians translate observations into hypotheses, which are depicted using epidemiological features and then assessed using regression analyses.

An important epidemiological tool for deriving such features are *Cohort studies*, such as the Study of Health in Pomerania (SHIP) [17]. To reduce any selection bias, subjects are invited at random and without a focus on a specific disease. The acquired features range from social and lifestyle factors to prior or current diseases and medications as well as medical parameters, such as blood pressure and also comprises of non-radiating medical image data. e.g. magnetic resonance imaging (MRI). Medical image data quantified using user-defined landmarks, describing for example shape, volume or diameter of a structure.

To assess the statistical resilience of a hypotheses using regression analyses rarely involves more than three features due to the required subject count. Due to missing overview techniques, possibly interesting correlations lie within the data, but are not made apparent. Explorative analyses and first overview visualizations of the data set as presented by Klemm et al. [12] are not custom tailored to a specific target variable and mostly highlights correlations between variables, which are known to the domain expert (e.g. correlation between body size and spine shape). We incorporate the regression analysis, which is familiar to the domain experts into a overview visualizations, which can either be used for an hypothesis-free analysis or a analysis towards a specific disease or hypotheses. This is achieved by providing template regression formulas, which are applied to all potential variable combinations. Since the notation is familiar to epidemiologists, they can rapidly include their domain knowledge into the analysis process. Difference views between regression formulas allow to assess the influences of individual variables in the process.

Our contributions are:

- A overview visualization technique describing feature interactions using target features.
- Visualization techniques, which incorporate overview visualizations of all regression analyses at once as well as details on demand techniques for detailed investigations of feature relationships.
- freely adjustable regression formulas provide a simple, yet powerful way to adjust the regression analysis to specific hypotheses about the data.
- analysis of confounding variables by providing comparison views between different formula results
- The open and web based approach of the system allows for analysis of any data using the presented method.

2 Epidemiological Background

This section covers the epidemiological workflow and requirements.

2.1 Epidemiological Workflow

Epidemiology unites experts from different academic disciplines, such as physicians, statisticians and medical computer scientists focusing on biometrics and image segmentation. Their goal is derive disease-specific risk factors by assessing epidemiological features using statistical methods. As described by [15] is divided into three different steps:

- Clinicians make observations in the daily practice which are translated into hypotheses.
- Epidemiologists compile a list of variables depicting the hypothesis and include confounding variables.

- *Paul Klemm, Kai Lawonn, Sylvia Glaßer, Uli Niemann, Bernhard Preim are with Otto-von-Guericke University Magdeburg, Germany. E-mail: {klemm,lawonn,niemann,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.

- Statisticians assess the association of the derived features towards the investigated disease.

Relative risks can be determined if a statistical resilient association of features towards a condition are extracted. They indicate the per-subject chance of developing the disease. Reproducibility of results is a epidemiological key requirement and guides all analyses steps. Statistical programs, such as SPSS are used to analyzed the data using the classical sequential epidemiological workflow, which uses images largely only to communicate results, rather than providing insight. A alternative data driven approach is described as follows.

2.1.1 Data Driven Hypothesis-Generation-based Analysis

Klemm et al. describe an approach for a Interactive Visual Analysis approach for image-centric cohort study data, which connects to the variable listing step [12]. Their methods aim to derive hypotheses through analysis of the data and observations about previously unknown feature correlations. Employing *hypotheses generation* requires overview visualizations of feature correlations, which are not supported by standard statistical processors.

2.2 Epidemiological Data

Cohort studies are a major epidemiological tool for gathering epidemiological features. They yield a highly heterogenous and incomplete information space. Subject data are collected with the widest range possible, allowing the data set to be assessed towards different diseases. The feature space comprises information about lifestyle, somatometric variables, medical parameters, genetic data as well as medical image data derived through different modalities, such as questionnaires or medical examinations or laboratory analyses. Many features are sparse, such as follow up questions about a medication or treatment of a certain disease. Other features are exclusive, such as women-specific questions, e.g. number of born children or period status.

Restriction of amount of data because it must not be triangulated and because the ethics committees are very restrictive

Data Types. Medical status variables or lifestyle factors are often of *dichotomous* (binary) type. The data space comprises continuous variables (somatometric variables, such as body weight or BMI or laboratory values) as well as categorical variables (e.g. graduation type). The data heterogeneity has to be taken into account when the analysis method is chosen. Continuous data are often discretized (e.g. 10 year steps for age) to equalize the variable types and to simplify the method selection. This is however avoided if possible, since it reduces the information space and also introduces a new information bias, as assumptions are modeled through the discretization.

Image Data. Since the Rotterdam study, many modern cohort studies include medical image data. For ethical reasons, the imaging modalities must not include ionizing radiation. The image quality is often inferior to clinical standard, which is a tradeoff between time and cost [14]. These data are hard to analyze as they require segmentation highlighting the structures of interest. This process is prone to inter and intra-observer variability when carried out manually. Automatic or semi-automatic solutions bypass this problem, but are costly and need to be custom tailored to the structure of interest [16]. Image-derived variables are either dichotomous variables indicating a medical finding by a radiologist or continuous, describing the shape of segmented structures (e.g. volume or diameter).

Confounding Features Features influencing the exposure as well as the outcome of a analysis are called confounders. The analysis model has to be adjusted by normalizing all included features towards the confounding feature. *Age* is included as confounder in almost any epidemiological analysis, since most diseases (such as different cancer types) are more likely to happen with increasing age. It also influences the general body condition and therefore almost all features acquired through cohort studies. Another important confounder is *Gender*. Other confounder have to be selected by epidemiologists specific to the investigated condition.

2.3 Regression Analysis

Regression analyses are the most important statistical tool when analyzing epidemiological data. They are also the foundation of this work. A regression analysis assesses the influence of one or more (*independent*) features to one target (*dependent*) feature. The regression model yields a function describing the target feature by weighting the independent features. Metrics, such as the weightings itself and associated p values and describe the resulting function (the *model*). R^2 values describe the goodness of fit; in other words how well the dependent features describe the target feature. The value is in the range [0,1], while 1 encodes a perfect fit.

Regression Analysis Notation. Regression formulas are usually denoted as follows:

$$Dependent \sim Independent_1 + Independent_2 + ... + Independent_n \quad (1)$$

The most used regression operators comprise:

- $+ / -$ inclusion/exclusion of the variable,
- : inclusion of interactions between the variables (e.g. $x : y$),
- $*$ inclusion the variables as well as their interactions (e.g. $x * y$)
- $|$ (conditioning) inclusion of variable x, given y (e.g. $x|y$)

The type of the target feature restricts the regression type.

Linear Regression for Continuous Target. The basic type is the linear regression, creating linear weightings for the *independent* features. The dependent variable has to be of a continuous type.

Logistic Regression for Dichotomous Target. Logistic regression implies a dichotomous target variable. The target is described by fitting a logistic function. Logistic models do as opposed to linear models does not allow for extracting a R^2 goodness of fit value. Therefore, pseudo-R^2 values are extracted, such as the *Nagelkerke* R^2, which mimics the behavior of the R^2. *Nagelkerke* R^2 cannot be compared to R^2 values extracted from linear regression model.

Move this to Regression Cube part?

2.4 The Study of Health in Pomerania (SHIP)

The SHIP, located in Northern Germany aims to aims to characterize health and disease in the widest range possible [17]. It does not focus on a specific disease, making the data set open for many diseases. Unique for the SHIP is the acquisition of medical image data per subject. A second cohort, SHIP-TREND was introduced in 2012. Data for both cohorts are examined in a in a 5-year time span. New parameters are added in each iteration, extending the range of investigated diseases. For the last acquisition (SHIP-2 and SHIP-TREND-0), MRI scans are included into the cohort [6, 8].

3 Prior and Related Work

From VAST'14 Paper:

Visual Analysis of Heterogenous Data. Zhang et al. [18] 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 [3, 7]. 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. [2] 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. [1] 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.

Prior Work. We visualized lumbar spine variabilities based on a semi-automatic shape detection algorithm of 490 participants of the `SHIP-2` cohort [11]. 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 [10]. Niemann et al. [13] 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.

4 Regression Cube Analysis of Cohort Study Data

The basic idea of our *Regression Cube* is to provide a overview visualization of large cohort study data sets towards target variables. Overview visualizations of feature relationships as presented by Klemm et al. [12] are often focused on relationships between the visualized features. Correlation metrics, such as the *Pearson product-moment correlation coefficient* or *Cramér's V* contingency values are incorporated to achieve this goal. In epidemiology, these relationships are also of interest, but rather towards their explanatory power towards the target feature. These target features often indicates the presence of the investigated disease. As described in Section 2.3, regression analyses are the statistical tool of choice for analyzing these relationships. A regression model is based on expert knowledge, there is no unified rule on how to apply them to a given set of features, so they have to be applied with care.

4.1 Cube Description using Regression Formula Notation

Expert knowledge is introduced into regression analyses using the regression formulas. As described in Section 2.3, the formula input influences the type of the chosen regression method as well as the *independent* features describing the target.

Since we want to use the regression analyses with a overview visualization, we are are interested in all possible combinations of (two or more) independent features describing a target. We achieve this by introducing dynamic variables X, Y and Z into the regression notation. The method then replaces the dynamic variables with all features in the data set. In a data set with 100 features, the regression formula

$$Cancer \sim X + Y \tag{2}$$

would yield 10.000 regression models, describing all possible combinations of two features in the data describing the target *Cancer*. The major advantage of this notation is that it comes natural to anyone familiar with regression analysis, because it uses the same notation, all operators can be used as before. This allows for a fast adaptation and in the epidemiological application domain. With simple adjustments to the formula, different results can be achieved:

- $Z \sim X + Y$ calculates all combinations of two features towards all possible target features.
- $Cancer \sim X + Y + BodyWeight$ includes the *BodyWeight* feature into all regression models as feature.
- $Cancer \sim X + Y + Z$ calculates all combinations of three features towards the target.

The problem with this brute-force approach lies in its complexity. The number calculated regression models increases exponentially for each dynamic variable added. If we assume a data set with 100 features with the formula $Z \sim X + Y$ we calculate 1,000,000 regression models. When each regression takes about 50 ms calculation, we have to wait roughly 14 hours for the calculation to complete. Therefore, the computational complexity needs to be reduced.

4.2 Target-Variable-dependent Dimension Reduction

Uli, Can you please proofread this? The vast majority of features in an epidemiological data have no or a very low relation towards the target feature. Identifying these features und excluding them from the calculation can reduce the number of dimension significantly. The *Correlation based Feature Selection* (CFS) algorithm is very popular in data mining for achieving this task [5]. It uses information entropy to select the features which have the most explanatory power towards the target feature. At the same time it tries to reduce the correlation between the independent features. When for example the *body weight* has a strong explanatory power towards the target it is likely that *BMI* or *waist circumference* behave similar towards it. They, however, correlate strongly with each other. The CFS algorithm would then select the feature which has the largest explanatory power and discards the other dimensions.

We apply the CFS algorithm for each target feature in a regression formula with dynamic variables. The formula $Cancer \sim X + Y$ would yield one initial CFS information space reduction. For $Z \sim X + Y$ the CFS algorithm is applied to the data every time Z is replaced with another feature.

The number of features calculated by the CFS algorithm is dependent on the information entropy in the data. In our epidemiological data we observed a usual number of 10 to 20 features. For features, such as age or gender, which affect most other features (Recall "Confounder" in Section 2.2) the number is significantly larger (about half the features in the whole data set).

Now we have a method to derive the interesting regression models in a reasonable time span. The next section shows ways of abstracting the results to make them visually feasible.

4.3 Abstracting Regression Results using R^2

The goal of a overview visualization is to provide a comprehensive view on the data, which is easy to understand. As described in [12], correlation values scaled between 0 (no correlation) and 1 (perfect correlation) can be encoded using color on a mosaic plot. Regression models are more complex, having many associated describing metrics. For the *Regression Cube* analysis we are interested in the goodness of fit of the resulting model, which allows to infer about the predictive quality of the independent features included in the model. As described in Section 2.3, the R^2 value is the metric allowing for this kind of assessment.

2D (Slice) View. Since R^2 is scaled between [0,1] it also allows for comparison between regression models. We can apply the same mosaic plot mapping by translating the R^2 values to color saturation (Fig. 3 a). This describes a 2D regression square for dynamic variables X and Y (e.g. $Age \sim X + Y$).

3D (Cube) View. Introducing Z creates a 3D *Regression Cube* (Fig. 3 b). R^2 values of each cube entry (*voxel*) is mapped to opacity to reduce the overlap. The visualization of R^2 values derived from different regression cubes (e.g. $Z \sim X + Y$) is misleading, as they can be compared relatively, but not in precise numbers. Therefore, the R^2 results of different regression methods are encoded using

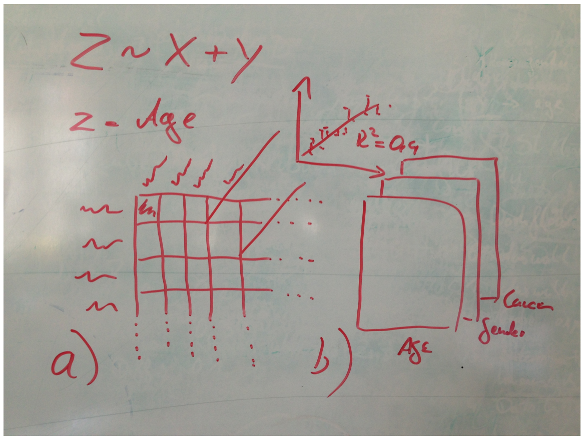

Fig. 1. (a) Overview visualization using a heatmap for the formula $Z \sim X+Y$, where Z assumes feature Age. The R^2 values extracted from the regression formulas depict the goodness of fit and are mapped to color saturation. A saturated color shows a strong correlation. (b) Since Z assumes all features n as given by the formula, it yields n heatmap visualizations. These represent the slices in our cube visualization. The R^2 values of each slice voxel is mapped on opacity in the 3D-view, reducing the occlusion of other values.

different colors (e.g. blue for linear regression and red for logistic regression). This way, the cube can easily be extended using other regression types. For cubes having one fixed target feature, such as $Cancer \sim X+Y+Z$ no such encodings is required and the z dimension can be compared directly.

The goal is to create a overview visualization for a data set, but on the other hand we also want to incorporate expert knowledge into the visualization by adapting the underlying formulas. These two approaches do not exclude each other, they rather underline the difference in purpose of the chosen formula. Different analysis approaches require different starting points using the *Regression Cube*.

4.4 Analysis Workflow

I'm not satisfied with the structure in this subsection, please proofread it with focus in the train of thought! The analysis workflow follows the Visual Analytics Mantra of Keim et al [9]:

Analyze First. Choosing an initial regression formula triggers the *Regression Cube* calculation on the given data set, filtering the dimensions of the dependent feature through the CFS algorithm.

Show the Important. The 3D-cube visualization acts as overview over the whole data set. Regression models with large R^2 values can be spotted fast here, steering the attention to the respective slice.

Zoom, Filter and Analyze Further. The slices of interest can then be analyzed using the 2D mosaic plot of the slice.

Details on Demand. Precise information about the individual regression models (coefficients, associated confidence intervals and p-values) can be retrieved based on the data point representatives (e.g. in a hover-modal on a currently selected data point).

As seen in Fig 2, the workflow is highly iterative. Observations in the 2D-mosaic plot or simply the CFS-based features can trigger new analyses by adjusting the underlying regression formulas. This can be carried out either to refine the current formula based on observations, or creating a new regression cube for a difference view.

Formula Types. Input formulas reflect *hypotheses* about the data. Using the operators, dynamic variables and dataset features, many different assumptions can be formulated. Here we show just a few examples on how formulas can serve different purposes.

$Z \sim X+Y$ is the default if the user did not specify a formula. This cube represents all possible combinations of two independent features towards all features in the data set, since we do not know which feature(s) are of interest. Each slice represents a different target feature. It is therefore suitable for a *hypothesis-free* explorative analysis to give a general impression about relationships in the data set.

$Cancer \sim X+Y+Z+feature_1 : feature_2$ is an example for the formulation of a hypothesis, where a specific feature is analyzed. The interaction between $feature_1$ and $feature_2$ is also included in all regression formulas.

$Cancer \sim X+Y+Z$ subtracted with results from $Cancer \sim Age$ excludes the confounding effect age has towards the target $Cancer$ feature. TODO: divide using analysis approaches, e.g. Hypothesis-based & Hypothesis-free with follow up on the formulas used?

Cube Comparison.

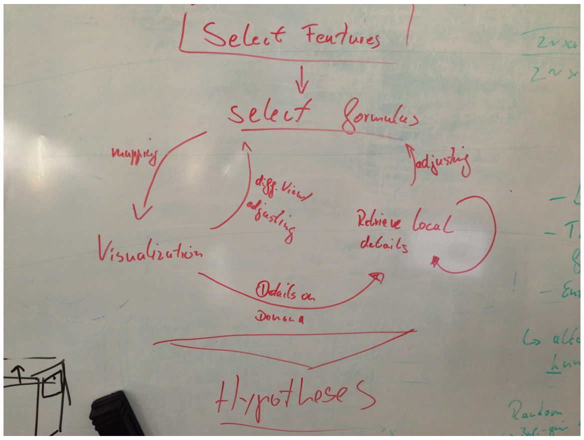

Fig. 2. Workflow of the analysis using *Regression Cubes*. The analysis starts with a set of features given by the data set. Then, the user may specify a formula regarding to specify a hypothesis, or use a predefined formula to start a explorative analysis of the data set. The Regression Cube is then visualized, where the user has either the option to adjust the formula or to derive details on demand on a specific regression. Insights into the data yield either an adjustment of the current formula. By applying a difference view, regression cubes can be compared to each other. The analysis yields insights and hypothesis about feature relations.

5 System Design and Implementation

5.1 Design and Visualization Techniques

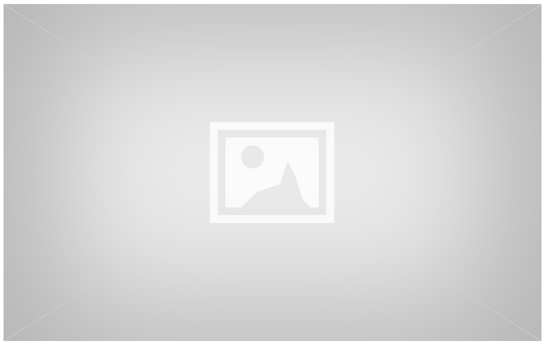

Fig. 3. Screenshot of the cube visualization embedded in the framework Also include active tooltip

Simplicity of interface to allow for steep learning curve. Visualization as skewed cube to reduce visual clutter and remove double entries. Cube acts as visual mini map to give a global impression over the data. current pane shown using a heat map visualization. details on demand for selected regression formula. comparison view using reference cubes.

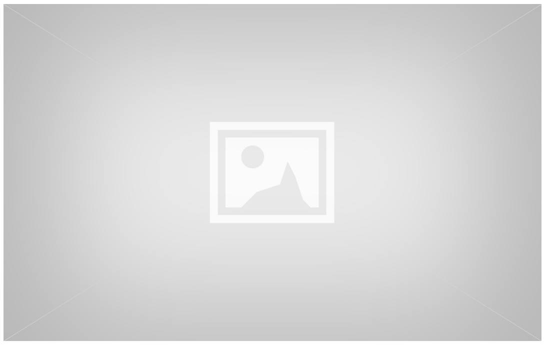

Fig. 4. Screenshots of Cubes from the Evaluation

System Layout.

5.2 Implementation

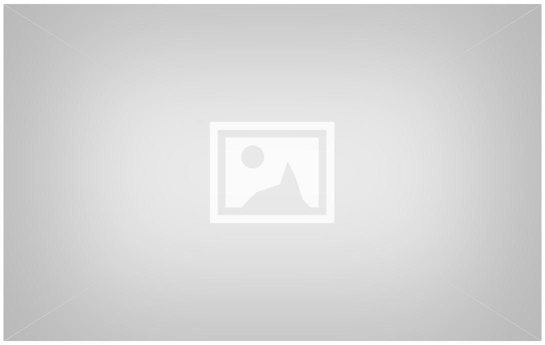

Fig. 5. Describe the Bootstrap/Angular/D3 Frontend, Node backend, R/OCPU Backend.

Use VAST'14 as Guideline Speed is merely a matter of available server machines due to parallelization process. Security enabled by hashing files. Bootstrap/Angular/D3 Frontend, Node backend, R/OCPU Backend (Fig. 5)

6 Application

6.1 The Breast Fat Data Set

Describe how the data was acquired and preprocessed

6.1.1 Data Preprocessing

The data processing follows the description in Section **??**.

6.2 The Hepatic Steatosis Data Set

Uli: Describe the data set and also prior work on it.

6.2.1 Data Preprocessing

The data processing follows the description in Section **??**.

6.3 Participants, Setup and Procedure

Use of VDAR Technique

Setup. Due to the large geographical distance, the evaluation was done completely web-based.

Procedure.

6.4 Case 1: Hypothesis-free Analysis of the Breast Cancer Data Set

6.5 Case 2: Hypothesis-driven Analysis of the Hepatic Steatosis Data Set

6.6 Further Feedback and Lessons Learned

Feedback from the evaluation goes here. Time-aspect critical, interactive analysis requires for fast response. The method needs to be speeded up.

7 Summary and Conclusion

Future Work. Implementing of regression analysis on the graphics card.

Acknowledgments

Omitted due to blind review

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[6] 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.

[7] 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.

[8] T. Ivanovska, R. Laqua, L. Wang, V. Liebscher, H. Völzke, and K. Hegenscheid. A level set based framework for quantitative evaluation of breast tissue density from mri data. *PloS one*, 9(11):e112709, 2014.

[9] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. *Visual analytics: Scope and challenges*. Springer, 2008.

[10] 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.

[11] 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.

[12] P. Klemm, S. Oeltze, K. Lawonn, K. Hegenscheid, H. Völzke, and B. Preim. Interactive visual analysis of image-centric cohort study data. *IEEE Trans. on Visualization and Computer Graphics*, 20(12):1673–1682, Dec 2014.

[13] 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.

[14] 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.

[15] 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.

[16] K. D. Toennies, O. Gloger, M. Rak, C. Winkler, P. Klemm, B. Preim, and H. Völzke. Image analysis in epidemiological applications. *it-Information Technology*, 57(1):22–29, 2015.

[17] 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.

[18] 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.