Regression Cube Analysis of Cohort Study Data

Paul Klemm, Kai Lawonn, Sylvia Glaßer, Uli Niemann, Katrin Hegenscheid, Henry Völzke, Bernhard 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 subject's lifestyle, medical condition as well as medical image data. Statistical regression analysis is used to evaluate these features and 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 that reveals insights 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 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 the *Regression Cube*, we present a visual overview of epidemiological data that allows for the first time an 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 [10]. Furthermore, the insights can be used to derive recommendations regarding a healthy lifestyle or to provide information about widespread diseases. During the standard workflow, physicians transform observations into hypotheses, which are depicted using epidemiological features and then are statistically analyzed.

An important epidemiological tool for deriving such features is the evaluation of a *cohort study*, such as the Study of Health in Pomerania (SHIP) [40]. To reduce any selection bias, subjects are randomly invited without a focus on a specific disease. Hence, a wide range of features is acquired. Social and lifestyle factors, prior or current diseases and medications as well as medical parameters, such as blood pressure, are gathered. Medical image data from non-radiating modalities, e.g., magnetic resonance imaging (MRI), is also acquired in modern studies. These data may be quantified based on user-defined landmarks, which describe attributes like the shape, volume or diameter of a structure.

Testing features for association with diseases using regression analysis is one of the most important epidemiological tools. Using it to assess the statistical resilience of a hypothesis rarely involves *more than three features* due to the higher dimensional problem and the required subject count. Due to the amount of data and only limited overview techniques, possible correlations may be missed. Explorative analyses and overview visualizations of the data set as presented in prior work [22] are not tailored to a specific target feature and mostly highlight correlations between features, 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 overview visualizations to support a hypothesis-free analysis or an analysis towards a specific disease or hypothesis. This is achieved by providing template regression formulas, which are applied to all potential feature 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 features on the process.

Our contributions are:

- An overview visualization technique representing feature interactions using [illegible]t features.
- Visualization techniques, which incorporate overview visualizations of all regression analyses at once as well as details on demand techniques for in-depth investigations of feature relationships.
- Freely adjustable regression formulas to provide a simple, yet powerful way to adjust the regression analysis to specific hypotheses about the data.
- The analysis of confounding features by providing comparison views between different formula results.
- The expansion and adaption of our method to any data with our open web-based system.

The term 'Regression Cube' has been used for the detailed analysis of one or a small number of regression models [1, 6]. They extend 2D scatter plots to display detailed information about a regression model. In behalf of that, we declare our visualization technique for comparing and analyzing a large number of different regression models with different input features as *Regression Cube*.

2 Epidemiological Background

This section covers the epidemiological workflow and requirements.

2.1 Epidemiological Workflow

Epidemiological research is performed by experts from different academic disciplines, such as physicians, statisticians and medical computer scientists focusing on [illegible]etrics and image segmentation. Their

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

goal is to derive disease-specific risk factors by assessing epidemiological features with statistical methods. As described by Thew et al. [36] the epidemiological workflow is divided into these different steps:

- Clinicians make observations in the daily practice, which are translated into hypotheses.
- Epidemiologists compile a list of features depicting the hypothesis and include confounding features.
- 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 is extracted. They indicate the per-subject chance of developing the disease. Reproducibility of results is an epidemiological key requirement and guides all analyses steps. Statistical programs, such as ~~SPSS,~~ are used to analyze the data regarding the classical sequential epidemiological workflow. Hence, images are mostly used for the communication of results, rather than for providing insight into the data.

In [22] we described an Interactive Visual Analysis approach for image-centric cohort study data, which connects to the feature listing step. The methods aim to derive hypotheses through analysis of the data and observations about previously unknown feature correlations. Employing *hypothesis generation* requires overview visualizations of feature correlations, which are not supported by standard statistical processors. In this work, we focus on a similar approach to derive insight ~~and~~ even new hypotheses through the data rather than only using it for a confirmatory analysis.

2.2 Epidemiological Data

There are many data acquisition modalities available to epidemiologists, which are suitable for different analysis types. We focus on cohort studies, which impose the largest data sets and yield a highly heterogenous and incomplete information space. Subject data are collected either towards a specific disease or with the widest range possible. The latter allows the data set to be assessed towards different diseases. The feature space comprises information about lifestyle, somatometric features, medical parameters, genetic data as well as medical images. These features are derived through different modalities, such as questionnaires, 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 for a sub-group, such as women-specific questions, e.g., number of born children or period status. Medical status features or lifestyle factors are primarily of *dichotomous* (binary) type. Continuous data are often discretized (e.g. 10 year steps for age) to equalize the feature types and to simplify the method selection. However, this should be avoided, since it reduces the information space and also introduces a new information bias, as assumptions are modeled through the discretization.

Medical image data is also acquired and analyzed in modern cohort studies. We incorporate image-derived features, but do not focus on analyzing image data. More discussions of cohort study data types and characteristics can be found in [29, 37].

Features influencing the exposure as well as the outcome of an analysis are called *confounders* and have to be specially treated. The analysis model has to be adjusted by normalizing all included features towards the confounder. *Age* is included as confounder in almost any epidemiological analysis, since most diseases (such as different cancer types) are more likely for older subjects. It also influences the general body condition and thereby almost all features acquired through cohort studies. Another important confounder is *gender*. Other confounders have to be selected by epidemiologists specific to the investigated condition.

2.3 Regression Analysis

Regression analysis is the most important statistical tool when analyzing epidemiological data and impose the basis 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. Different metrics, such as the weightings itself and associated p values, describe the resulting function (the *model*). R^2 values describe the quality of the fit; in other words how well the dependent features describe the target feature. The value ranges between $[0,1]$, where 1 encodes a perfect fit.

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

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

An example of a regression formula would be $KidneyDisorder \sim Smoking + Obesity$. The most commonly used regression operators comprise:

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

The class of the target feature restricts the regression type. Different regression types are available, we focus on the following:

Linear Regression for Continuous Target. The basic type is the linear regression, creating a linear map from the space comprising the *independent* features to the *dependent* 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, as opposed to linear models, do not allow for extracting an R^2 quality 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 behaves different than R^2 values extracted from the linear regression model. Comparisons have to be handled with care.

2.4 The Study of Health in Pomerania (SHIP)

The SHIP, located in Northern Germany, aims to characterize health and disease in the widest range possible [40]. 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 cohort subject. A second cohort, SHIP-TREND was started in 2012. Data for both cohorts are examined in a 5-year time span. New parameters are added in each iteration, extending the range of investigated diseases. For the latest two acquisitions (SHIP-2 and SHIP-TREND-0), MRI scans are included into the cohort [15, 18]. However, most examinations occur in all stages and are performed according to the same instructions to enable comparisons over the different stages of the study.

3 Prior and Related Work

In 1977, Tukey already stated that data is too often analyzed solely using confirmatory data analysis [38]. He emphasized the need to use data to derive hypotheses instead, which can then be tested again. In this section we present prior and related work trying to achieve this goal.

The work of Piringer et al. [28], Chan et al. [6], Angelelli et al. [3] and Zhang et al. [41] are closest related to ours regarding different aspects. Akin to our approach, Piringer et al. [28] propose methods for visualizing regression analysis results and properties for developing car engines. Their main goal is to assess the pairwise influence of independent features towards the target feature using a plot matrix displaying models as contours. They also incorporate 3D visualizations for each pairwise combination, but mainly because of its popularity with the target domain engineers. Linked views of model deviations allow to select outliers. This limits the method to

comparing a few models at once as the plot matrix gets complex with increasing feature number. The main difference is their focus on analyzing one complex model in detail, yielding extensive plots, while we analyze a large amount of simpler models in terms of *different* features. They also focus on metric features, while we process categorical data as well.

We use the term 'Regression Cube' for a comparative 3D representation of multiple regression models. Chan et al. [6] use the same term to describe an extension of the 2D scatter plot representation of a linear regression model (incorporating solely metric features) to a 3D Cube. They group subjects using a set of interaction techniques as well as clustering algorithms to calculate sub-groups, which can then be compared using their cube representation. Similar to [28], they focus on highlighting details of the included models rather than comparing models consisting of different features. Insight is derived by subject grouping, which spawns new cube correlations and therefore allows drilling down to the data. In contrast to this approach, we focus on comparing models using quality-of-fit measures and do not focus on subdividing the data sets to gain insight. We model expert knowledge through the definition of regression formulas, which is not the focus of Chan et al.

Zhang et al. [41, 11] is closest to our work regarding the application of visual analysis on cohort study data. They present *Cohort Analysis via Visual Analytics* (*CAVA*), a framework that distinguishes three major elements of cohort study data analysis: *cohort data* (and its manipulation using operations), *views* and *analytics*. They use the system to find longitudinal pathways for diseases on the basis of health records. Sub-groups are automatically created by incorporating *analytics* and through selection in expert-guided *views*. We use their requirements as guidelines, which incorporate *flexible* and *iterative analysis*.

Angelelli et al. [3] visualize image-derived an non-image data using cube data structures with focus on comparison and knowledge extraction. Their proposed engine represents multiple heterogenous cohort study data sets as normalized data cubes. Thus, data redundancy is avoided and runtime measure aggregation can be carried out when measures of the different cubes have to be combined. They use *Pearsson's r* to depict relationships towards target features and employ list views and scatter plots to visualize and rank them. In contrast, we aim for a fast large scale correlation analysis incorporating many features to derive insight into the data.

Visual Analysis in Public Health. Shneiderman et al. [33] highlight the necessity of interactive visualizations in healthcare data. The presented challenges in this application require the systems to be sufficiently comprehensible and intuitive to fit into the short time span that clinicians have for analyzing data. Furthermore, the systems should provide comparative relationship visualization, characterization of similarity and presenting risks.

In their survey on interactive information visualizations for health record data, Rind et al. [31] identified future work addressing the usability of such frameworks. We follow their recommendation to design an open system available to everyone and applicable to various data sets. Schreck and Keim [32] present a method for analyzing social media data with focus on characterization and cause of an epidemic. They employ data aggregation techniques and use geographical data to map social media events, focusing on word clouds to derive insight into possible causes. They derive hypotheses using their high variation average message density algorithms, steering the user's focus to particular events.

Steenwijk et al. [34] focus on *hypothesis-free* exploration of cohort data. They employ a framework consisting of feature extraction and visualization to derive dependencies between image- and non-image features. They incorporate linked views using scatter plots, bar charts, parallel coordinates as well as time plots to display and brush the data. Generalized Pairs plots (GPLOM's) extend the idea of scatter plot matrices by the pairwise depiction of heterogeneous data using type-combination-dependent visualizations [9, 17]. The techniques are useful for selected features or a data set with a small number of features due to their increase in complexity with every additional parameter. They are therefore suitable for *small* epidemiological data sets. Dai et al. [8] incorporate a GPLOM-like visualization using choropleth maps mapping spatial data, such as mortality rates together with scatter plots augmented with *Pearsson's r* values. A *concept map* summarizes features related to a specified disease. Time-dependent epidemiological data is visualized by Chui et al. [7] using multi-panel graphs highlighting risk factor differences with age and gender with regard to influenza- and salmonellosis-associated hospitalizations.

Statistical Analysis. Bertini et al. [4] present an overview of quality metrics describing high dimensional-data. Their research agenda acts as guideline for our design, comprising perceptual tuning towards human pattern recognition of important aspects, scalability between different data set sizes and application testing with domain experts. Ahmadi et al. [1] define the *Sparse Regression Cube* that partitions sparse high dimensional data into subspaces, which are then described by their most reliable linear regression model. They focus on an algebraic representation for efficient regression-model calculation to find the best fit for a subspace.

Albuquerque et al. [2] present an interactive exploration framework displaying quality metrics of high dimensional data sets with brushing facilities to create subsets. They analyze subsets using a drilling-down approach by incorporating scatter plot matrices (SPLOM's) of quality metrics. Turkay et al. [39] follow a similar approach by using both descriptive metrics for features as well as the features themselves and incorporate them into linked plots. The approach was applied to a cognitive aging study, where new hypotheses could be derived.

Niemann et al. [26] investigate risk factors of hepatic steatosis using decision trees with interactive data mining tools. They extract classification rules that serve as basis for our proof-of-concept tests. In [25], Niemann et al. improve the classification performance by generating features (called *evolution features*) that describe latent temporal information across the study waves. We try to reproduce their results and investigate findings presented in [26] further.

Prior Work. In our prior work, we analyzed the healthy aging process of the lumbar spine. We utilized semi-automatic algorithms do detect the lumbar spine shape [21, 30].

We defined an Interactive Visual Analysis workflow for image-centric cohort study data, by extending the feature selection step (recall Sec. 2) with an iterative analysis loop incorporating group selection and visualization using expert input as well as clustering methods [22]. Information visualizations of cohort study features were augmented with extracted image data, yielding linked views with both image and non-image information. The mosaic plot presented in [22] highlights the pairwise correlation between features without a target disease by depicting the *Cramér's V* contingency values. Its great popularity among our domain experts was the inspiration for this work.

While the presented techniques were well received by the epidemiological experts, the explanatory power towards back pain was limited, yielding an analysis based on image-derived features, such as extracting and analyzing curvature, torsion and angle of the lumbar spine [20]. We concluded that the model quality is insufficient to characterize back pain.

The difference of the presented approach towards our previous and related work is twofold:

1. We focus on the large-scale analysis of a vast number of linear and logistic regression-models by assessing their quality-of-fit using descriptive metrics.

2. The analysis is conducted w.r.t. a target feature and incorporates expert knowledge via the regression model definition rather than subdividing the underlying data.

4 Regression Cube Analysis of Cohort Study Data

The basic idea of our *Regression Cube* is to provide an overview visualization of large cohort study data sets towards target features. Overview visualizations of feature relationships as presented by Angelelli et al. [3] 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 w.r.t. their explanatory power towards the target feature. These target features often indicate the presence of the investigated disease. As described in Section 2.3, regression analysis is the statistical tool of choice for analyzing these relationships. A regression model is based on expert knowledge; there is no singular rule how to apply models to a given set of features. Thus, they have to be applied with care.

4.1 Cube Description Using Regression Formula Notation

Expert knowledge is integrated into the analysis using 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 associate the regression analyses with an overview visualization, we 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. Our method replaces the dynamic variables with all features in the data set. In a data set with n (e.g., 100) features, the regression formula

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

would yield n^2 (10,000) regression models, describing all possible combinations of two features in the data describing the target feature *Cancer*. The major advantage of this notation is that it comes natural to anyone familiar with regression analysis, since it is the standard way of expression. With simple adjustments to the formula, different results can be achieved:

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

The problem with this approach lies in its complexity. The number of calculated regression models exponentially increases for each dynamic variable added. If we assume a data set with 100 features and the formula $Z \sim X + Y$, we obtain 1,000,000 regression models. When each regression takes about 50 ms of calculation time, we roughly have to wait 14 h for the calculation to complete. Therefore, the computational complexity needs to be reduced. An approach for this is presented in the following section.

4.2 Target-Variable-Dependent Dimension Reduction

In epidemiological studies, manifold recordings lead to an abundance of features and thus a high-dimensional feature space. In general, many of them exhibit a low or no correlation at all in view of the target feature. Identifying irrelevant features and excluding them from the feature space considerably reduces computational costs and yields a comprehensible *Regression Cube* representation. The correlation-based feature selection (CFS) [13] aims to find a feature subset that maximizes the *merit value* M_F, which is the ratio between the average feature-class and feature-feature dependencies in the feature set F. The dependency of a set of features utilizes the entropy-based information gain to measure the explanatory power towards the target feature. Starting with an empty set of features F, the CFS algorithm iteratively adds the feature f to F that leads to the highest new merit value $M_{F \cup f}$ and halts when no feature is left that would increase the merit. For example, if the *body weight* has a strong explanatory power towards the target, it is likely that *BMI* or *waist circumference* exhibit similar correlations to the target. However, they strongly correlate with each other. The CFS algorithm will 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 usually observed a number of 10 to 30 features. The number of selected features using the CFS algorithm reflects their information entropy on the target. A large list of features is an expression of low correlation to the target feature.

With this method, we are able to derive the interesting regression models in a reasonable time span (seconds to minutes instead of hours). The next section shows ways of abstracting the results to make them visually feasible.

4.3 Abstracting Regression Results Using R^2

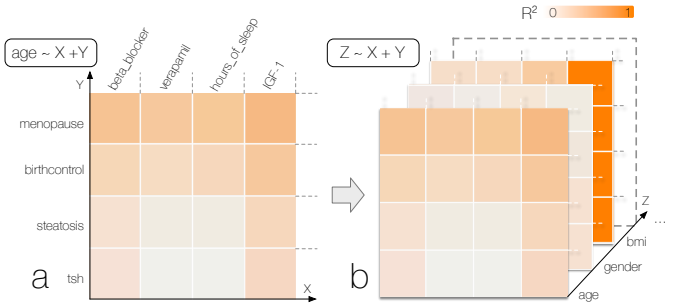

Fig. 1. (a) Overview visualization using a mosaic plot for the formula $Z \sim X + Y$, where Z assumes the feature *age*. The R^2 values extracted from the regression formulas depict the quality-of-fit and are mapped to color saturation (a saturated color shows a strong correlation). (b) Now, Z is set to all features n and yields n mosaic plot visualizations. These represent the slices in our cube. The R^2 value of each slice voxel is mapped on opacity in the 3D view later on, reducing the occlusion of other values.

The goal of an overview visualization is to provide a comprehensive view on the data (raw or using descriptive metrics [4]), which is easy to understand. As described in our previous work [22], correlation values scaled between 0 (no correlation) and 1 (perfect correlation) can be encoded with color in a mosaic plot. Regression models are more complex, having many associated describing metrics. For the *Regression Cube* analysis we are interested in the quality-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 allows for comparison *between* regression models. We can apply the same mosaic plot mapping by translating the R^2 values to color saturation (Fig. 1a). 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. 1b). R^2 values of each cube entry (*voxel*) are 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 different colors (i.e., orange for linear regression and blue for logistic regression). Thus, the cube can be easily extended using other regression types. For cubes with a fixed target feature, e.g., $Cancer \sim X + Y + Z$, no such encodings are required and the z dimension can be compared directly.

Our goal is to create an overview visualization for a data set, but 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. The different analysis approaches require different starting points using the *Regression Cube*.

4.4 Analysis Workflow

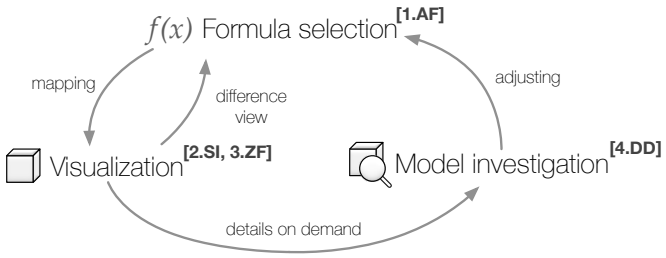

Fig. 2. Different workflow types of the analysis using *Regression Cubes* (based on [19]). [**1.AF**] The workflow starts by declaring a formula to specify a hypothesis, or to use a predefined formula for a hypothesis-free analysis. [**2.SI**] The *Regression Cube* is then visualized. The user has the option to either adjust the formula or to derive details on demand on a specific regression. [**3.ZF**] Insights into the data yield either an adjustment of the current formula or a selection of a difference view. The latter is used to compare regression cubes. [**4.DD**] Details about features using the 2D mosaic plot representation yield insights and hypotheses about feature relations.

Our *Regression Cube* model is well suited for different workflow analysis techniques, based on the Visual Analytics (VA) Mantra of Keim et al. [19]:

Analyze First [**1.AF**]. 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 [**2.SI**]. The 3D cube visualization acts as an overview over the whole data set. Here, regression models with large R^2 values can be spotted fast, steering the user's attention to the respective slice.

Zoom, Filter and Analyze Further [**3.ZF**]. The slices of interest can then be analyzed using the 2D mosaic plot of the slice.

Details on Demand [**4.DD**]. 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).

We use the squared bracket abbreviation for each step to denote the affiliation to the system design section later on. As shown 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 to create a new *Regression Cube* for a difference view.

Hypothesis-Free and Hypothesis-Based Analysis. Input formulas reflect *hypotheses* about the data. Using the operators, dynamic variables and dataset features, many different assumptions can be expressed. To support the *hypothesis-free* analysis, we provide a default formula:

$Z \sim X + Y$. This cube represents all possible combinations of two independent features towards all features in the data set, since we do not know which features are of interest. Each slice represents a different target feature. It is therefore suitable for an explorative analysis to give a general impression about relationships in the data set.

Hypotheses about the data are easily built up by relating dynamic variables with the regression operators. Furthermore, static features can be added for each regression formula. Here are a few examples:

$Cancer \sim X + Y + Z$ is the formulation of a hypothesis, where the specific feature *Cancer* is analyzed. All combinations of three independent features towards the target are analyzed through the cube.

$Cancer \sim X + Y + Z + feature_1 : feature_2$ encodes more assumptions. This formula models the hypothesis of an interaction between $feature_1$ and $feature_2$ (denoted with ':') being relevant for the target feature, but it is not clear how other feature combinations influence the result. Therefore, the cube incorporates this interaction for all X, Y and Z values as independent features.

$Cancer \sim X + Y + Z$ subtracted with the R^2 from $Cancer \sim Age$ excludes the confounding effect that age has in view of the target *Cancer* feature. This is achieved through cube comparison.

Cube Comparison. *Regression Cubes* can be compared by creating difference views. One cube (extracted with one formula) acts as reference. The absolute difference in R^2 values towards the second cube is calculated, yielding a difference cube showing only the differences between the two formulas. For example, it can be utilized for comparing the influence of a single feature towards the complete result (e.g., $Z \sim X + Y$ and $Z \sim X + Y + Income$)

5 System Design

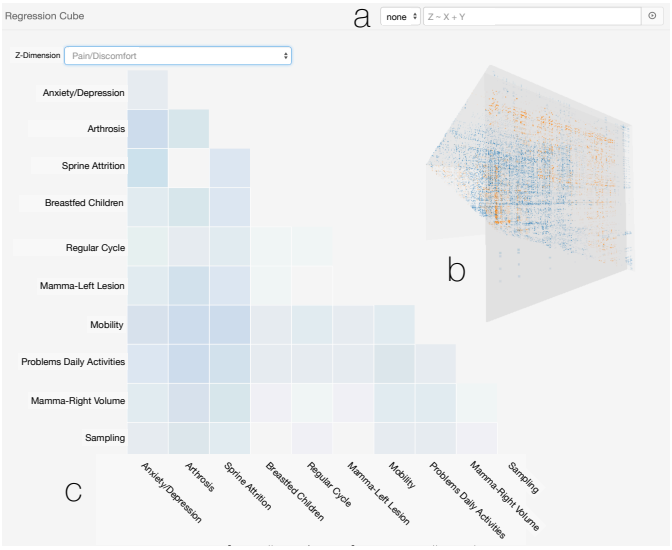

Fig. 3. Breast fat data set loaded into the *Regression Cube* prototype. (a) Using the formula input, the user specifies the dependent feature as well as its calculation rules. (b) 3D cube visualization, showing values above the cube matrix diagonal as overview. The values of the currently selected slice are mirrored and represented as orange data points on the slicing plane. (c) 2D mosaic plot visualization of the selected slice for feature *Pain/Discomfort*.

We designed our system to be freely accessible and easy to use. With open formats as input interfaces, the application can be extended to non-epidemiological data sets. The focus lies on creating an overview visualization and gaining insight into relationships of the data, which triggers further analyses, maybe with other (statistical) tools. Therefore, the system has to be intuitive and comprehensive in order to be adapted by domain experts.

Design choices and spaces are restricted by the underlying technologies and devices. The exchange of data and methods using web-based technologies offers many advantages. There is no set-up time involved and domain experts can use the methods from any computer connected with the web. By design, web technology is based on a client-server architecture, making it easy to outsource computationally heavy tasks on server clusters and transferring results to the client device. The design space spanned by web technologies is different to standard WIMP applications: right-click menus, modal windows and menu bars are no established user interface components in this context. Therefore, we have to adapt our design respecting the conventions of web pages.

5.1 System Paradigm and Components

The *Regression Cube* design focuses on a clean interface, reducing the amount of user-interface elements as much as possible. This allows for a fast learning of the individual system parts. Our prototype consists of three components:

- The *file upload* section starting the analysis with providing a comma-separated value (CSV) file [1.AF].
- The *cube visualization* consisting of the 2D mosaic view as well as a 3D representation of the whole cube [2.SI].
- The *formula editor* allows formula input w.r.t. a hypothesis or to conduct a *hypothesis-free* analysis. It also allows to select a reference formula for creating a difference cube [1.AF, 3.ZF].

File Upload and Classification [1.AF]. Popular analytics tools, such as WEKA [12], owe their popularity to their support of open file types. To allow other users even outside the epidemiological application domain to access to our tool, we use standard ASCII-based CSV files. The first line in a CSV file represents all features (columns) of the data set. Each line after that represents one subject (row) and its feature manifestations.

Encoding via CSV Files. Encoding variable types in CSV files is not standardized. However, we need to ensure the correct variable type classification and have to enforce some basic standards. All categorical values have to be enclosed by quotation marks. Continuous variables are denoted as digits without enclosing quotation marks. Although this seems obvious, many cohort study data sets encode categorical features using ID values that are denoted in a data dictionary. Variables with only two manifestations are classified as dichotomous, leading to three possible data types: numerical, categorical and categorical/dichotomous. Missing values are denoted by using no character at all, a whitespace, or an empty quotation mark encapsulated string.

Data Security. Security issues are raised by uploading data into an online service such as our prototype. The use of epidemiological data is preceded by a detailed description of the analysis purpose and has to be approved by ethics committees. Preventive steps have to be taken to restrict access to unauthorized subjects. We calculate a SHA-256 hash to derive the data set name using the data contents and disable directory listings on the web server to avoid data set downloads. Data sets are deleted from the server after closing a session.

Formula Editor [1.AF, 3.ZF]. After uploading the data, the user can specify a formula or use the default ($Z \sim X + Y$). Entering a formula is facilitated via text input. On formula input, a context panel displays all data set features as well as the available operators and their function. This allows to comprehend the function of the underlying formula for users without statistical background about regression analysis and its notation. Auto-completing input features also simplifies the approach and works as spell check of feature names.

Formula Validation and Calculation. The formula is checked for validity directly on input. The text input containing the formula is marked using a red halo to indicate invalid input, which turns green for valid formulas. This prevents processing errors on the statistical processor back-end. Confirming a formula triggers the cube calculation, which is preceded by determining all required formulas. These are then divided by the number of available statistical back-end processors, driving a *cloud-computing*-based approach. In theory, the calculation duration is reduced by a factor of 0.5 by every statistical processor. In practice, data transmission and differences in machine specifications always influence the speed.

Difference Cube. Adding a formula also adds it to the reference selection for a difference cube. Since all cells in the cube are represented using R^2 values, a difference cube is calculated by denoting the absolute difference of R^2 for each cell.

5.2 Regression Cube Visualization [2.SI].

The visualization and interaction with the *Regression Cube* is the prototype core. Results from the statistical processors are uploaded into the visualization slice by slice, allowing the assessment of the data as soon as parts of the calculations are finished while the rest is still in progress.

Usage of a Regression Prism for Information Reduction. Figure 1 shows that all values are mirrored along the diagonal of the mosaic plot matrix. This is due to the symmetry of basic regression operators. $Z \sim X + Y$ produces the same result as $Z \sim Y + X$. Therefore, we can discard half of the results to reduce visual clutter and repetition, yielding a *Regression Prism*. This opens up space for displaying additional information. Along the diagonal, X and Y assume the same feature, $Z \sim X + Y$ turns into $Z \sim X$ because the regression automatically ignores doublings. The diagonal therefore acts as reference on how strong the correlation for the given row (or column) feature is.

3D Prism as Data Mini-Map. The 3D cube representation acts as an overview over the whole data set, but its purpose is not to derive detailed information about data points. It serves as a function similar to a mini-map, guiding the attention towards points of interest in the data, as well as giving context information about adjacent data values when using the 2D mosaic plot. The displayed prism shows values above the matrix diagonal. For formulas with a dynamic target feature (e.g. exploratory analysis using $Z \sim X + Y$), the color encodes absolute R^2 values (Fig. 3b). Applying this strategy to a formula containing a static target (e.g. *Cancer* $\sim X + Y + Z$) yields many occlusions, since the CFS algorithm creates the same feature space for every slice. For such formulas, the 3D view encodes every data element as absolute difference between its R^2 features and the global mean along the z-axis. This highlights slices with unusually low or high results (Fig. 5).

Tackling the disadvantages of 3D information visualization. 3D information visualizations are often criticized for introducing occlusions and interaction problems, which often do not balance out the advantages of using the third dimension for visual mapping. We aim to minimize these problems. Since the R^2 values are mapped on data point opacity, large values are highlighted in the prism, guiding the focus to the respective slices. It also creates a sparse representation, since the majority of regression models yield (depending on the data set and the chosen formula) low R^2 values. Also, the preceding correlation-based feature selection reduces the information space significantly, leading to sparse cubes. Overlapping is still an issue, but this way greatly reduced in its effect to the readability of the visualization.

Rotating with the cube is restricted to the y-axis, preserving the mental map to position individual features. The cube is always oriented according to the 2D representation, allowing for an easy mental combination of the two representations. Allowing more degrees of freedom was confusing to our users and also did not add value to the visualization. Also we provide a zoom functionality using the mouse wheel input.

Cube Slice Selection [3.ZF]. In order to *Zoom, Filter and Analyze Further*, the user has to navigate towards different slices of interest. We propose two ways to achieve this.

- *Applying the slicing metaphor from 3D volume data.* In medical volume data renderings, slicing views are very common to view details on a selected plane in the scene. We employ this technique for selecting cube slices (e.g., by moving a plane via vertical mouse input while pressing the right mouse button). We, however, still display the whole 3D object instead of cutting away information towards the slice position.
- *Selecting the slice using a dropdown menu* provides fast access to plane selections when the user already knows the slices of interest.

The currently selected slice is denoted with a semi-transparent gray plane. The space available from visualizing only the prism generated from the upper half of the cube diagonal is used to display information about the currently selected plane. The values are projected on this plane to give an overlapping-free view on the data points, which makes it easier to identify the current slice.

2D Mosaic Plot Slice Visualization [4.DD]. The 2D mosaic visualization (Fig. 3c) shows all values below the matrix diagonal of the current slice, creating optical equivalence towards the 3D cube. To reduce visual clutter, the 2D view only shows dimensions, which are retrieved through the correlation-based feature selection. The free space above the matrix diagonal is used to display the 3D cube.

The purpose of this view is the detailed assessment of the underlying regression models. By hovering over a data entry in the plot, a tooltip displays detailed information about a model's coefficients, associated p values and confidence intervals.

6 Implementation

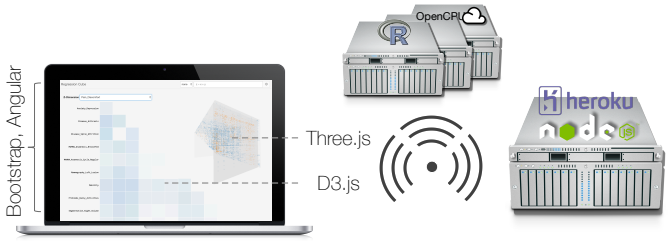

Fig. 4. The front-end (left) is realized with *HTML5*/*CSS3*/*Javascript* and different Javascript libraries, such as *Angular.js*, *Three.js* and D3.js. The web-server back-end (right) is written using *Node.js*, hosting is carried out using *Heroku*. *R* and *OpenCPU* constitute the statistical back-end (top) to compute the *Regression Cubes*. Additional statistical back-ends can be attached to the system to decrease the computation time.

We rely on web-based technologies for our prototype. The ongoing transition of open-science software into the web spawned numerous projects, making state-of-the-art algorithms available in this domain.

Front-End. The front-end is created using `HTML5`, `CSS3` and `Javascript`. `Angular.js`[1] abstracts web application into models and views, allowing for a responsive way to combine `HTML` and `Javascript`. It forces developers to write modularized code, which makes the components easier expandable while keeping the code maintainable by including unit tests. The page layout is handled using `Twitter Bootstrap`[2], which also provides a rich set of user interface elements with proper stylings. The 2D mosaic plot is implemented using `Data driven Documents` (`D3.js`), which is popular in information visualization using vector graphics [5]. It provides fast and easy methods for binding data to graphical elements. The 3D plot is created using the WebGL-based `Threejs`[3] library. We experimented with different ways for achieving the cube representation, including volume rendering, cube primitives for each data point and shader-based solutions. Open source volume rendering methods are available but do not satisfy our requirements. Creating a cube primitive for each data point resulted in non-interactive frame-rates for data sets larger than 30 features (creating 30^3 cube primitives). Therefore, we decided to use a shader-based solution by rendering the cube as a sprite-based particle system, allowing to customize color and opacity of every data point. It also is the fastest solution that we tested.

Back-End. Two server structures serve as back-end. The first one is the web-server, which is written in `Javascript` using `Node.js`[4], running on Googles V8 Javascript runtime environment. It is hosted on `Heroku`[5], a cloud application platform.

The statistical processors yield the second structure. They rely on the statistical programming language `R`.[6] It is widely adopted in the statistical analysis community, yielding a rich support of fast state-of-the-art statistics algorithms as well newly published methods. `OpenCPU` is an `R` package and provides an API for accessing it via HTTP calls [27]. This way, any computer, which runs `R` can be turned into a statistical processor for our project. The back-end functions necessary for all cube calculations are provided via an `R` package. It uses multi-core optimization to use all machine CPUs to speed up the calculation process. The server workload balances are managed by the front-end code.

Access and Source. A running instance of the *Regression Cube* prototype can be found under `regressioncube.herokuapp.com`. The source for the prototype is freely available at `Github`.[7,8] Instructions and code on how to create a setup running the *Regression Cube* statistical back-end through a `Ubuntu` server using `OpenCPU` are referenced in the repository. The front-end can be deployed using `www.heroku.com` by cloning the repository into a Heroku app.

7 Application

In this section, we describe how we applied the *Regression Cube* to two epidemiological data sets. The hepatic steatosis data set was analyzed using data mining algorithms, yielding risk groups, which we now analyze further. Also, we try to reproduce the prior results from this analysis as proof-of-concept of our method. The breast fat data set is the basis for an explorative analysis towards the influencing parameters of the parenchyma tissue of the female breast.

Both data sets are unusual for epidemiological analysis regarding their feature extent. Usually, only a few features depicting a hypothesis are compiled into a data set to assess them using statistical tools. The herein used data sets comprise several hundred features. Our method focuses on data exploration and knowledge extraction and requires a wide scope of sociodemographic, medical and lifestyle features.

7.1 Participants, Setup and Procedure

The ability of a system to discover knowledge is difficult to measure. Lam et al. [23] propose the *Visual Data Analysis and Reasoning* (*VDAR*) technique, which is focused on the characterization of a systems' ability to generate hypotheses and explore the data in order to extract information. *VDAR* can be carried out using case studies using thinking-aloud techniques to comprehend the reasoning and thought process of the user. We employ *VDAR* for analyzing our system.

Participants, Setup and Procedure. We conducted a web-based analysis by using an online-meeting software, which features voice chat as well as screen sharing. Starting an analysis using these techniques takes about 5-10 minutes of setup time.

The sessions started with an initial overview of the system, showcasing its features and functionality. Afterwards, the experts use the system on their own computers. The screen-sharing function was still used to observe the actions of the expert. All sessions were video recorded to be processed later on.

We conducted the analysis with three participants. *KH*, a clinician (10 years of experience) with focus on epidemiological research, is the domain expert for the breast fat data set analysis. She is a radiologist, responsible for the SHIP-MRI acquisition and also responsible of the mammography analysis. The hepatic steatosis data set is analyzed by *UN*, a data scientists responsible for prior analysis of the data. The third participant is *TI*, a statistician with focus on epidemiology (8 years of experience) assesses the statistical reliability of the tool and the underlying methods without a focus on a specific data set.

7.2 The Hepatic Steatosis Data Set

We employ the data set used in [26] to identify predictive features w.r.t. the reversible hepatic steatosis disorder. The dichotomous target feature is derived from the liver fat concentration measured using MRI

[1] Open Source; Maintained by Google, `angularjs.org`
[2] Open Source; Maintained by Twitter, `getbootstrap.com`
[3] Open Source; Originally developed by R. Cabello, `threejs.org`
[4] Open Source; Maintained by Joyent Inc, `nodejs.org`
[5] Owned by Salesforce.com, `heroku.com`
[6] Open Source; `r-project.org`

[7] R-based back-end: `github.com/paulklemm/regression-cube-r-package`
[8] Front-End and Node.js Webserver: `github.com/paulklemm/regression-cube-prototype`

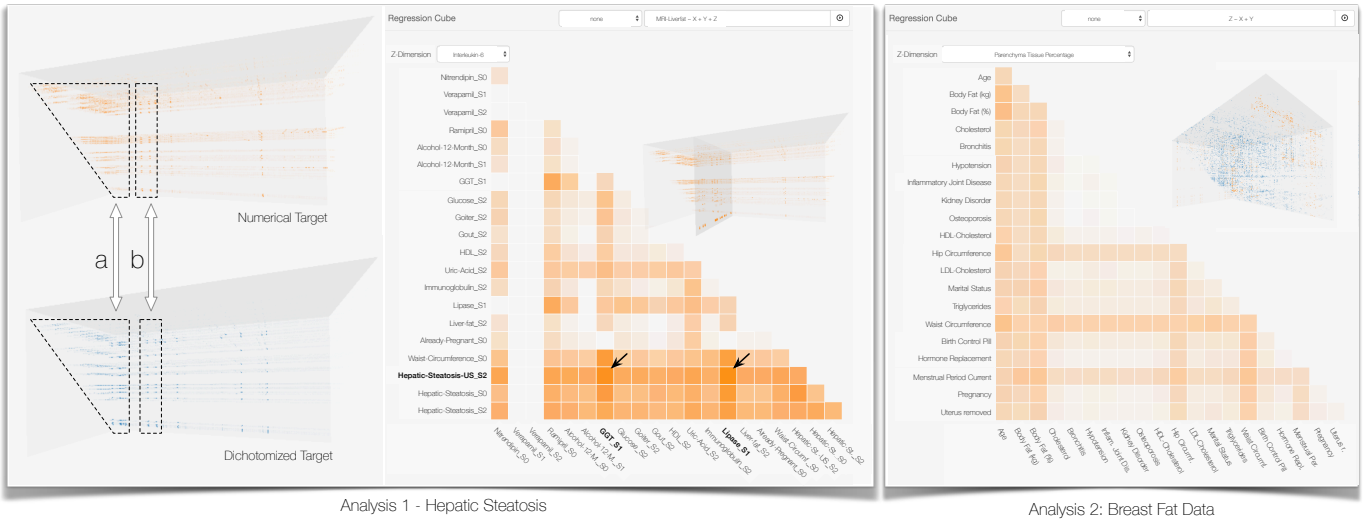

Fig. 5. The analysis of the numerical and dichotomized target feature depicting liver fat values yields similar results (left). In (a), hotspots for somatometrical features with high correlations were found. High correlations were also found for features depicting *hepatic steatosis* (b). A high correlation between *Interleukin-6*, *hepatic steatosis*, *GGT* and *Lipase* (highlighted using arrows) was revealed during the analysis using the 2D mosaic plot. The hypothesis-free analysis of the breast fat data set (right) w.r.t. the *parenchyma tissue percentage* of the breast displays correlations between *age*, *body fat*, *hip circumference* as well as *menstrual period*.

scans. Liver fat concentrations of no more than 10% are mapped to the 'negative' class; values greater than 10% are mapped to the 'positive' class to indicate absence or presence of the disease. The data set contains labels for 578 participants for which the MRI data was available to that point. The MRI scans for each subject have been introduced in SHIP-2, so the target feature is only present the last study wave.

Apart from the target feature, the data set contains 199 features, comprising sociodemographic features (gender, age), consumption behavior (e.g., alcohol or tobacco), laboratory data (e.g., sera concentrations), and two features depicting the liver ultrasound. The acquisition wave is denoted using the appendix; 85 features with appendix *s0* denote their affiliation to SHIP-0 (first study moment), 50 features for *s1* and 55 for *s2*, alongside with 10 time-independent SNPs (DNA base pairs).

In [26], the authors show different class distributions of liver fat concentrations of women and men. For women, an association between age and liver fat was identified. An appropriate cut-off value of 52 years was set, yielding the most homogeneous class distribution within the resulting subsets. Based on these observations, we perform our analysis on three populations: *males*, *females (all ages)* and *females older than 52 years*.

7.3 The Breast Fat Data Set

The breast fat data set was compiled to find associations between the parenchyma tissue proportion in the female breast compared to other features in the data. The ratio between parenchyma and cellular connective tissue (breast density) has been shown to be associated with breast cancer. Studies describe a four to five times increased risk of getting breast cancer for participants with a breast density above 50% [24].

The data comprises 1,186 female subjects (368 from `SHIP-2`, 818 from `SHIP-TREND-0` cohort). It contains 231 dimensions, holding information about somatometric features (e.g., body size or weight) consumption behavior (e.g., alcohol or tobacco), personal and medical history (e.g., occupation or prior diseases), women-specific features (e.g., number of born children or contraception type) as well as mammography features (e.g., fat content or parenchyma tissue proportion to volume). The latter were derived from MRI data for each subject, which was manually segmented by radiologists.

The data of each cohort was presented as individual `SPSS` files. All features related to the mammography attributes were stored in an additional file. We converted the `SPSS` data sets to CSV and used `R` to merge the data sets together using their ID. All features were renamed to be self-explaining, e.g., *chro_09a* is now denoted as *Disease_Osteoporosis*. This avoids the need of defining a separate data dictionary file for translating the feature names.

7.4 Case 1: Hypothesis-Driven Analysis of the Hepatic Steatosis Data Set

We refer to each analysis step with regard to its belonging in the VA-Mantra (recall Sec. 4.4). The analysis goal was reproducing results with our visual analysis framework that are in accordance to the data mining-based results presented in [26]. Therefore, *UN* started the [**1.AF**] step using the dichotomized MRI fat liver concentration and the formula $mrt\_liver\,fat\_s2 \sim X + Y + Z$ for *male* subjects. The [**2.SI**] step using the 3D cube representations locates hotspots at the end of the cube (Fig. 5a). The Zoom, Filter and Analyze Further Step [**3.ZF**] was realized by slicing through the cube using the mouse input to inspect the hotspots. Analyzing the mosaic plot [**4.DD**] revealed high correlations for somatometric features, hepatic steatosis indicator features as well as laboratory values, such as *creatinine* (used as renal retention parameter) and *uric acid* (among others used as gout and diabetes risk factors) magnitudes. Similar results were present for analyzing the *female* groups. *UN* could reproduce most results, some features exhibit lower correlations though, e.g., *creatinine* magnitudes. A slight influence of *age* on the target feature could be observed for women (R^2 of 0.09 for females compared to 0.02 for males). Relationships not described in [26] were found, such as enzymes indicating liver dysfunctions, e.g., *aspartate aminotransferase*. Due to the difference between our regression model approach and the decision tree approach presented in [26], a complete matching set of correlating features is not expected.

Analysis of Non-discretized Target Feature. Since our method can assess numerical target features, the analysis was conducted again for the non-dichotomized target using the same formula. The 3D cube representation already showed lower R^2 values in general. However, the analysis is now based on linear regression and the R^2 values cannot be directly compared. The correlation hotspots matched with the ones from the dichotomous target, but were generally lower (R^2 of 0.37 for somatometric features as opposed to 0.58). We assume that the bias introduced by dichotomizing the fat liver content enforces the findings of liver diseases, while using the numerical features is less expressive.

Interleukin-6 Correlation With Liver Fat. During the analysis, one hotspot was always observable in the [**2.SI**] and [**3.ZF**] steps, in-

corporating a high *Interleukin-6* (*IL-6*) correlation with liver fat values (R^2 of 0.8, see Fig. 5b). The correlation was high for both the dichotomized and continuous target feature. The literature described relations between *IL-6* and liver cancer [14] as well as chronic liver diseases [35]. For mice, strong effects of *IL-6* with hepatic steatosis were described [16]. The finding is subject of further analysis.

7.5 Case 2: Hypothesis-free Analysis of the Breast Cancer Data Set

The analysis aims to find relationships on the breast fat data using mammography analysis features. Relationships between the share parenchyma tissue on the overall breast volume are of high interest [24]. The [**1.AF**] was started by *KH* using the default formula for hypothesis-free analysis ($Z \sim X + Y$). At first, she was interested in correlations with the *parenchyma tissue* percentage, which was selected through the drop-down for the z-axis [**2.SI**]. She observed strong correlations with *age*, *body fat percentage*, *hip* and *waist circumference* as well as *menstrual period* or *pregnancy status* as expected (Fig. 5 right). Women with higher *body fat* also have a larger *breast fat percentage*, which also correlates with other somatometric features. *Age* is a strong influencing factor, as breast tissue and subsequently the parenchyma tissue degrades over time. *KH* proceeded using [**3.ZF**] and [**4.DD**] to check for relationships for different target features, such as current *hormone replacement therapy*, *BI-RADS* (classification of the mammography findings) as well as different diseases, such as *diabetes* or *gout*. She observed relationships matching her expectations and expert knowledge. One unexpected relationship was observed between *breast lesions* and *menstruation cycle* towards *spiral contraception* (R^2 of 0.77). *KH* proceeded with a detailed analysis of the parenchyma tissue.

Detailed Breast Parenchyma Analysis. The analysis was conducted by calculating the cube $Parenchyma\_Percentage \sim X + Y + Z$ [**1.AF**]. Using the 3D cube, *KH* observed several hotspots [**2.SI**]. Navigating to them using the slicing facility of the 3D visualization [**3.ZF**] highlighted features of high influence, such as image-derived features, as *glandular tissue density* and *parenchyma segmentation* metrics. Also, strong correlations were observed in the *diabetes* slice, confirming expectations of *KH* w.r.t. its strong influence to the parenchyma tissue. A surprising finding was the strong correlations with *kidney disorder* (R^2 values around 0.9). The [**4.DD**] analysis, however, showed only 8 subjects with this disease. Too few subjects impose the risk of a biased finding. The correlation was noted and will be further investigated in the near future using an extensive data set. Lastly, *KH* assessed the influence of contraception-related features, such as use of *birth control pills* or the *spiral*, but found no significant correlations with the parenchyma tissue. Other consumption behavior features, such as *alcohol intake* also yield no elevated R^2 values. *KH* remarked that these features are suspected to have an impact on the parenchyma tissue, but they are less reliable, since they are self-reported.

7.6 Further Feedback and Lessons Learned

The presented method was well received among the domain experts. For the first time, they were able to derive an overview visualization custom-tailored to underlying assumptions. *KH* noted the ease of use, which "*converts data sets into a feasible form*". She highlighted the efficiency of combining fast target feature selection with visually highlighting interesting results, enabling rapid analysis cycles. To get nearly similar results, she had to spend hours using `SPSS` and potentially missed interesting hotspots during this process. *TI* highlighted the ability to simultaneously analyze thousands of regression models while maintaining little time expenses for rating them. **Extracted Hypotheses Have to be Investigated Further.** We map results of complex statistical results into comprehensive visualizations. Agreeing with *TI*'s feedback, each finding and hypothesis has to be confirmed using a dedicated statistical analysis. An accompanying search for correlations potentially highlighting confounders can be carried out using our method. Statistical validation of an epidemiological end result still has to be carried out by statisticians using their respective tools. *TI* commented on the possibility of adding more regression types to model different correlation types.

Overview Visualizations are Preferred over Black-box Methods. Explorative analysis based on the data gains importance in epidemiology with increasing data set complexity. Results from automatic 'black-box' methods, such as data mining algorithms are more often obscure to the expert. Findings and hypotheses derived through overview visualizations, however, are met with more confidence, because the users actually observed the behavior themselves. The participation and steering of the analysis using human pattern detection and expert knowledge is preferred. Observing *expected* correlations matching the expert knowledge strengthens the confidence in the method and, subsequently, in the hypotheses generated from unanticipated relationships.

Using Non-discretized Features Reduces Information Bias. Discretization reduces the information space and introduces bias into the data and is therefore avoided in epidemiological research whenever possible. In contrast to many data mining algorithms, our method allows to use the concurrent analysis of heterogeneous data types. Investigations of the hepatic steatosis data set with both numerical and dichotomized liver fat values showed comparable results but slightly differed when it comes to details. The overall explanatory power towards the numerical feature was lower, supporting the hypothesis that the dichotomized target feature already models knowledge to bias the results towards the expected result.

Attention Steering is Crucial. Important events have to be highlighted in overview visualizations to direct the user's attention to interesting parts of the data. Poor guidance potentially leads to overlooked relationships. We found supporting mini-map visualizations, such as the 3D cube, most useful for this purpose, e.g., for highlighting differences rather than displaying absolute values (Fig. 5).

8 Summary and Outlook

We presented a technique for knowledge discovery in cohort study data sets with user-defined target features. Dimension reduction using the target restricts the analysis to the most important features. *Hypothesis-free* analysis employs default regression models. Modeling expert knowledge using regression formulas allows for a *hypothesis-based* investigation. A 3D *Regression Cube* allows to assess hotspots in the analysis by abstracting regression models using a quality-of-fit measure. These can then be analyzed further using the 2D mosaic plot for each cube slice. Details on demand for each model allow for a detailed assessment of regression models. We successfully applied the approach to find correlations in a hepatic steatosis as well as a breast cancer data set. The method was well received by our clinical partners, triggering detailed investigations of the findings.

As next step, we ~~want~~ to introduce more regression types, which model different kinds of correlations. We also ~~want to~~ extend the cube to time-dependent data by expanding the difference cube approach.

We published all associated code and provide a freely accessible analysis platform open to heterogenous data types. We want to support opening up knowledge discovery to allow a diverse group of domain experts to derive insight into their data and support the notion of open science.

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. This work was supported by the federal state of Saxony-Anhalt under grant number 'I 60' within the Forschungscampus STIMULATE.

REFERENCES

[1] H. Ahmadi, T. Abdelzaher, J. Han, N. Pham, and R. K. Ganti. The Sparse Regression Cube: a Reliable Modeling Technique for Open Cyber-physical Systems. In *Proc. of IEEE/ACM Second International Conference on Cyber-Physical Systems*, pages 87–96, 2011.

[2] G. Albuquerque, M. Eisemann, T. Löwe, and M. A. Magnor. Hierarchical Brushing of High-Dimensional Data Sets Using Quality Metrics. In *Proc. of VMV - Vision, Modeling & Visualization*, pages 119–126, 2014.

[3] P. Angelelli, S. Oeltze, C. Turkay, J. Haasz, E. Hodneland, A. Lundervold, B. Preim, and H. Hauser. Interactive Visual Analysis of Heterogeneous Cohort Study Data. *IEEE Computer Graphics and Applications*, 2014. in print.

[4] E. Bertini, A. Tatu, and D. Keim. Quality Metrics in High-dimensional Data Visualization: an Overview and Systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

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

[6] Y.-H. Chan, C. D. Correa, and K.-L. Ma. Regression Cube: A Technique for Multidimensional Visual Exploration and Interactive Pattern Finding. *ACM Transactions on Interactive Intelligent Systems*, 4(1):7:1–7:32, 2014.

[7] K. K. Chui, J. B. Wenger, S. A. Cohen, and E. N. Naumova. Visual Analytics for Epidemiologists: Understanding the Interactions Between age, Time, and Disease With Multi-panel Graphs. *PloS one*, 6(2), 2011.

[8] 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.

[9] 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.

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

[11] D. Gotz, A. Perer, and Z. Zhang. Iterative Refinement of Cohorts Using Visual Exploration and Data Analytics, Apr. 17 2014. US Patent App. 13/672,000.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

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

[14] G. He, D. Dhar, H. Nakagawa, J. Font-Burgada, H. Ogata, Y. Jiang, S. Shalapour, E. Seki, S. E. Yost, K. Jepsen, et al. Identification of Liver Cancer Progenitors Whose Malignant Progression Depends on Autocrine IL-6 Signaling. *Cell*, 155(2):384–396, 2013.

[15] 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. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 181(08):748–759, 2009.

[16] F. Hong, S. Radaeva, H.-n. Pan, Z. Tian, R. Veech, and B. Gao. Interleukin 6 Alleviates Hepatic Steatosis and Ischemia/Reperfusion Injury in Mice With Fatty Liver Disease. *Hepatology*, 40(4):933–941, 2004.

[17] 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.

[18] 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.

[19] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. *Visual Analytics: Scope and Challenges*. Springer, 2008.

[20] P. Klemm, S. Glaßer, K. Lawonn, M. Rak, H. Völzke, K. Hegenscheid, and B. Preim. Interactive Visual Analysis of Lumbar Back Pain-What the Lumbar Spine Tells About Your Life. In *Proc. of Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 85–92, 2015.

[21] P. Klemm, K. Lawonn, M. Rak, B. Preim, K. Tönnies, K. Hegenscheid, H. Völzke, and S. Oeltze. Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data. In *Proc. of VMV - Vision, Modeling & Visualization*, pages 121–128, 2013.

[22] P. Klemm, S. Oeltze-Jafra, 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, 2014.

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

[24] V. A. McCormack and I. dos Santos Silva. Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: a Meta-analysis. *Cancer Epidemiology Biomarkers & Prevention*, 15(6):1159–1169, 2006.

[25] U. Niemann, T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kühn. Can we Classify the Participants of a Longitudinal Epidemiological Study from Their Previous Evolution? In *Proc. of the 28th IEEE Int. Symposium on Computer-Based Medical Systems (CBMS15)*, June 2015. in print.

[26] U. Niemann, H. Völzke, J. Kühn, and M. Spiliopoulou. Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis. *Expert Systems with Applications*, 41(11):5405–5415, 2014.

[27] J. Ooms. The OpenCPU System: Towards a Universal Interface for Scientific Computing Through Separation of Concerns. *Computing Research Repository - arXiv*, abs/1406.4806, 2014.

[28] H. Piringer, W. Berger, and J. Krasser. HyperMoVal: Interactive Visual Validation of Regression Models for Real-Time Simulation. *Computer Graphics Forum*, 29(3):983–992, 2010.

[29] 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, 2015. in print.

[30] M. Rak, K. Engel, and K. Toennies. Closed-Form Hierarchical Finite Element Models for Part-Based Object Detection. In *Proc. of VMV - Vision, Modeling & Visualization*, pages 137–144, 2013.

[31] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Interactive Information Visualization to Explore and Query Electronic Health Records. *Foundations and Trends in Human-Computer Interaction*, 5(3):207–298, 2013.

[32] T. Schreck and D. Keim. Visual Analysis of Social Media Data. *Computer*, 46(5):68–75, 2013.

[33] B. Shneiderman, C. Plaisant, and B. W. Hesse. Improving Healthcare With Interactive Visualization. *Computer*, 46(5):58–66, 2013.

[34] M. Steenwijk, J. Milles, M. van Buchem, J. H. C. Reiber, and C. Botha. Integrated Visual Analysis for Heterogeneous Datasets in Cohort Studies. *Proc. of IEEE VisWeek Workshop on Visual Analytics in Health Care*, 2010.

[35] K. L. Streetz, F. Tacke, L. Leifeld, T. Wüstefeld, A. Graw, C. Klein, K. Kamino, U. Spengler, H. Kreipe, S. Kubicka, et al. Interleukin 6/Gp130-dependent Pathways are Protective During Chronic Liver Diseases. *Hepatology*, 38(1):218–229, 2003.

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

[37] 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.

[38] J. W. Tukey. Exploratory Data Analysis. *Reading, Ma*, 231:32, 1977.

[39] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser. Hypothesis Generation by Interactive Visual Exploration of Heterogeneous Medical Data. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 1–12. Springer, 2013.

[40] H. Völzke, D. Alte, C. Schmidt, et al. Cohort Profile: The Study of Health in Pomerania. *International Journal of Epidemiology*, 40(2):294–307, 2011.

[41] Z. Zhang, D. Gotz, and A. Perer. Iterative Cohort Analysis and Exploration. *Information Visualization*, 2014. Published Online First, doi:10.1177/1473871614526077.