

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 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, which 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 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.

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 [11]. Furthermore, they can be used to derive recommendations regarding a healthy lifestyle or to provide information about wide-spread diseases. During the standard workflow, physicians transform observations into hypotheses, which are depicted using epidemiological features and then assessed using regression analyses.

An important epidemiological tool for deriving such features is the evaluation of a *cohort study*, such as the Study of Health in Pomerania (SHIP) [38]. 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 using user-defined landmarks, which describe attributes like the shape, the volume or the 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 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 by Klemm et al. [21] are not tailored to a specific target variable and mostly highlight 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 overview visualizations, ~~which can~~

~~be used for~~ 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 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 on the process.

Our contributions are:

- An overview visualization technique representing 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 in-depth 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.
- The analysis of confounding variables by providing comparison views between different formula results.
- The expansion and adaption of our method to any data with our open and web-based system.

Our Regression Cube has been used for the detailed analysis of one or a small number of regression models [7, 2]. 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. Bernhard, does this suffice as reasoning?

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 biometrics and image segmentation. Their goal is to derive disease-specific risk factors by assessing epidemiological features using statistical methods. As described by Thew et al. [34] the epidemiological workflow is divided into these different steps:

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

- 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.
- 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 using the classical sequential epidemiological workflow. Hence, images are mostly used for the communication of results, rather than for providing insight into the data. An alternative data-driven approach is described as follows.

Data Driven Hypothesis-Generation-based Analysis. We already described an Interactive Visual Analysis approach for image-centric cohort study data, which connects to the variable listing step [21]. 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 heterogeneous 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 variables, medical parameters, genetic data as well as medical image data. 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 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. 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 put focus on analyzing image data. More discussion of data types and characteristics can be found in [21, 35, 28].

Confounding Features. Features influencing the exposure as well as the outcome of an 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 with increasing age. 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. The presented work is based on regression analysis as well. 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 is in the range [0,1], where 1 encodes a perfect fit.

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

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

The most used regression operators comprise:

- $\boxed{+, -}$ 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 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 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.

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 [38]. 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 ~~introduced~~ 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 [14, 17]. 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 ("*method loyalty*").

3 PRIOR AND RELATED WORK

In 1977, Tukey already stated that data is too often only analyzed using confirmatory data analysis [36]. He emphasized the need to use data to derive hypotheses, which can then be tested again. In this section we present prior and related work trying to achieve this goal. We put each paper in relation to our work, derive ideas and point out differences, starting with the works closest to ours:

The work of Piringer et al. [27], Chan et al. [7], Angelelli et al. [4] and Zhang et al. [39] is closest to ours regarding different aspects. Akin to our approach, Piringer et al. [27] propose methods for visualizing regression analysis results and properties for developing car engines. Their main goal is to ~~pairwise~~ assess the influence of independent features towards the target. This is achieved by using a plot matrix displaying models as ~~columns~~ towards the target feature. They also incorporate 3D visualizations for each pairwise

combination, but mainly because it is popular with the engineers in their target domains. Linked views of model deviations allow to select outlier residuals. Comparison of models is also included by mapping model values on line-style (e.g. solid for $model_a$ and dashed for $model_b$). This limits the method to comparing a few models at once. The plot matrix representation however gets complex and harder to understand with increasing number of independent features. The main difference is their focus on analyzing one complex model in detail, yielding in extensive plots, while we analyze a large amount of simpler models towards *different* features. They also focus on metric features, while we process also categorical data. We share the matrix-arrangement of results depicting relationships between independent features in a model.

We use the term Regression Cube to refer to a comparative 3D representation of multiple regression models. Chan et al. [7] use the same term to describe an extension of the 2D-Scatterplot representation of a linear regression models (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 [27], 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. Does this make sense as reason or not, we use the same term "Regression Cube" for different things?

Zhang et al. [39] is closest to ours regarding the application of visual analysis on cohort study data. ~~Their work 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 using subject health records. Sub-groups are created automatically by applying analytics and expert-guided using views. Unfortunately, they do not go into the detail into the analytics used, they seem to not incorporate confounder analysis (other than dividing groups by gender) and the user study lacks concrete goals. Their data differs strongly from ours, having multiple moments for each subject, yielding a approach focused on finding pathways of interest. We use their requirements as guidelines, which incorporate flexible and iterative analysis, meaning that the system has to adapt to different requirements in different analysis steps. We implement this by supporting both hypothesis-based and hypothesis-free analysis.~~

Angelelli et al. [4] ~~is similar to ours as they visualize both image-derived and 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 multiple normalized data cubes to eliminate data redundancy between them and perform runtime measure aggregation when measures of the different cubes have to be combined. This allows for saving memory space. Their presented aggregation functions allow for a fast data calculation to map it using different linked visualizations, such as scatter plots, bar charts or histograms. They use Pearson's r to depict relationships towards target features and employ list views and scatter plots to visualize and rank them. We use regression models and associated R^2 values to depict relationships between a target feature and multiple independent features instead of one and employ the information into one global visualization. We employ the idea of fast conducting large scale correlation analysis between many features to derive insight into the data.~~

Visual Analysis in Public Health. Shneiderman et al. [31] highlight the necessity of interactive visualizations in healthcare data, e.g. using *Health 2.0*, which is an umbrella for web-based health-informatics frameworks. They divide Health 2.0 in three domains, *Personal health information*, *clinical health information*, and *public health information*. We focus on the latter. The presented challenges

in this application comprise custom tailoring the systems to be comprehensible and intuitive to fit into the short time span clinicians have when analyzing data. Furthermore, the systems should provide comparative relationship visualization, characterization of similarity and presenting risks. We use the **challenges as design guidelines** for our web-based prototype, focusing on **functional clean designs** to allow fast insight into relationships towards diseases.

In their survey on interactive information visualizations for health record data, Rind et al. [29] identified future work addressing the usability of such frameworks. They recommend in opening up analysis systems to a diverse target group, e.g. including patients, who also can observe information about their own lifestyle and associated risks. We follow their recommendation in designing an open system available to everyone, applicable to a various data sets.

Schreck and Keim [30] presented 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 users focus on particular events. ~~The focus remains on geolocated social media data rather than subjects with multiple dimensions. These and similar systems follow the same analysis work as we do to derive insight into data.~~

Perer et al. divide cohort analysis in three major parts, present CAVA, a system **which allows**

Steenwijk et al. [32] 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 and time plots to display and select data. ~~Medical image data can be incorporated to investigate potentially interesting individuals. We distinguish to this and other linked-view analysis based approaches by showing abstracted information derived from a preceding analyze-first step to highlight relationships rather than displaying the actual data using various information visualization techniques.~~

Generalized Pairs plots (GPLOMS) extend the idea of scatter plot matrices by pairwise visualizing heterogeneous data using type-combination-dependent visualizations [10, 16]. 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. Dai et al. [9] incorporate small multiples using a GPLOM-like visualization using choropleth maps mapping spatial data, such as mortality rates together with scatter plots augmented with Pearson's r values. A *concept map* summarized features related to a specified disease. Time-dependent epidemiological data is visualized by Chui et al. [8] using multi-panel graphs highlighting risk factor differences with age and gender towards influenza- and salmonellosis-associated hospitalizations. We also include confounder features (age, gender) into the analysis while focusing on non-time-dependent cohort study data sets.

Statistical based Analysis. Statistical analysis ~~is an~~ inherent part of ~~Analysis~~. Here we present work showing a strong focus on statistical analysis in this context. Bertini et al. [5] present an overview over quality metrics describing high dimensional data. Their research agenda acts as guideline for our design, comprising application of domain-experts, perceptual tuning towards human pattern recognition of important aspects and scalability between different data set sizes. Ahmadi et al. [2] define the *Sparse Regression Cube*, which partitions sparse high-dimensional into subspaces, which are then described using by their most reliable linear regression model. ~~In their application, they try to find most fuel-efficient roads connecting two user-defined landmarks. They focus on presenting an algebraic for efficient regression-model calculation to find the best fit for a subspace. The dimensions for the calculations are fixed, while we assess different features for each model calculation and can therefore not benefit from a hierarchical calculation approach.~~

Albuquerque et al. [3] present **a** interactive exploration framework displaying quality metrics of high dimensional data sets with brushing facilities to create subsets. These subsets are then analyzed in detail, following a drilling-down approach. They incorporate scatter plot ma-

trices (SPLOMS) of quality metrics describing the features. We incorporate the idea of abstracting features for comparison. Turkay et al. [37] 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 studies, where new hypotheses could be derived.

Niemann et al. [25] investigate risk factors of the hepatic steatosis using decision trees with a interactive data mining tools. They extracted association rules, which serve as basis for this works proof-of-concept evaluation. In [24], Niemann et al. improve on 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 their findings further.

Prior Work. In recent publications, we focused on analysis on characterization of the healthy aging process of the lumbar spine. Therefore, we utilized semiautomatic detection algorithm to detect the lumbar spine shape [20]. The reliable part of the presented models are represented by the shape of the lumbar spine canal, which was extracted and described using line representation.

We defined a Interactive Visual Analysis workflow for image-centric cohort study data, by extending the feature selection step (see 2) with a interactive analysis loop incorporating group selection and visualization using expert-input as well as clustering methods [21]. Information visualizations of cohort study features were augmented with extracted image data, yielding linked views with both image-and non-image information. The popularity of the mosaic plot depicting *Cramér's V* contingency values with our domain experts was the inspiration for this work. The representation presented in [21] only highlights pairwise correlation between features without a target disease.

While the presented techniques were well received with epidemiological experts, the explanatory power towards back pain was limited, yielding a analysis based on image-derived features, such as extracting and analyzing curvature, torsion and angle of the lumbar spine [19]. These measures were analyzed using a novel decision tree quality plot towards non-image data. We concluded that the model quality is not enough to characterize back pain.

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

1. We focus on 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 towards 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 variables. Overview visualizations of feature relationships as presented by Angelelli et al. [4] 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 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 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

$$Cancer \sim X + Y \quad (2)$$

would yield n^2 (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 of n . 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 approach lies in its complexity. The number of 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

In epidemiological studies, manifold recordings lead to an abundance of features. Generally many of them exhibit low or ~~none~~ correlation towards the target feature. Identifying irrelevant features and excluding them from the feature space considerably reduces computational costs and contributes to **presentiveness** of the Regression Cube. 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 of features in F . The dependency of a set of features utilizes the entropy-based information gain to measure the explanatory power towards the target feature. Beginning 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 adding any feature does not increase the merit. 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). The number of features extracted is not directly associated with the overall number, but rather on their information entropy towards each other. **Uli, is this correct this way?**

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.

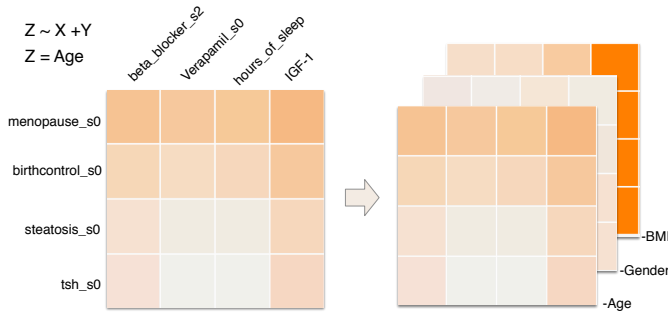


Fig. 1. (a) Overview visualization using a heatmap for the formula $Z \sim X + Y$, where Z assumes feature Age. The R^2 values extracted from the regression formulas depict the goodness of fit and are mapped to color saturation. A saturated color shows a strong correlation. (b) Since Z assumes all features 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.

4.3 Abstracting Regression Results using R^2

The goal of an overview visualization is to provide a comprehensive view on the data (raw or using descriptive metrics [5]), which is easy to understand. As described in our previous work [21], 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 allows for comparison between regression models. We can apply the same mosaic plot mapping by translating the R^2 values to color saturation (Fig. 1 a). This describes a 2D regression square for dynamic variables X and Y (e.g. $\text{Age} \sim X + Y$).

3D (Cube) View. Introducing Z creates a 3D *Regression Cube* (Fig. 1 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 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 $\text{Cancer} \sim X + Y + Z$ no such encodings are required and the z dimension can be compared directly.

The goal is to create an 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

The analysis workflow follows the Visual Analytics (VA) Mantra of Keim et al. [18]:

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. Regression models with large R^2

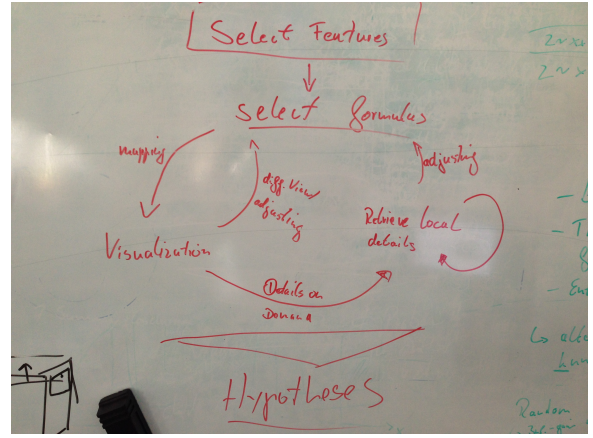


Fig. 2. Workflow of the analysis using *Regression Cubes*. [1.AF] The analysis starts with a set of features given by the data set. Then, the user may specify a formula regarding to specify a hypothesis, or use a predefined formula to start a explorative analysis of the data set. [2.SI] The *Regression 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. [3.ZF] 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. [4.DD] Details about features using the 2D-heatmap representation yields insights and hypothesis about feature relations.

values can be spotted here fast, steering the users 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 the system design section later on. 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.

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 formulated. To support *hypothesis-free* analysis by providing 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 feature(s) 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 reflected in either how dynamic variables are related using the regression operators. Also, static features can be added for each regression formula. Here are a few examples:

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

$\text{Cancer} \sim X + Y + Z + \text{feature}_1 : \text{feature}_2$ 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 combination 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 age has towards the target *Cancer* feature. This is achieved through cube comparison.

Cube Comparison. Regression Cubes can be compared by creating difference views. One cube (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. It can for example be utilized for comparing the influence of one 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 **current** selected slice are mirrored and represented as orange data points on the slicing plane. (c) 2D mosaic plot visualization of the selected slice. **TODO: Adapt formula to final image.** Context information is shown for the regression model of $Parenchym \sim Bronchitis + AlcoholLast30Days$, depicting details about the confidence intervals, coefficients and associated p-values.

We designed our system with openness and ease of use in mind. Using open formats as input interfaces does not restrict the application **towards** non-epidemiological data sets. The focus lies on creating an overview visualization and gaining insight into relationships into 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. Resting collaboration and exchange about data and methods on web-based technologies shows promising results, as there is no set-up time involved and domain experts can use the methods from any computer connected with the web. Web technology is by design 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 not established UI components in this context. Therefore, we have to adapt our design respecting the affordances 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. More importantly, it allows the user to focus all mental resources on the analysis tasks, **rather than wrangling with configuring the system to serve the current task or hypothesis.** Therefore, 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 to adjust the formula **towards** a current hypothesis or conducting a *hypothesis-free* analysis. It also allows to select a reference formula for creating a difference-cube [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 their feature manifestations.

Classification. Encoding variable types in CSV files is not standardized. We, however, 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. This seems obvious, but in fact many cohort study data sets encode categorical features using ID-values, which 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 apply a SHA-256 hashing on 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 provided via text input. On entering the text **are**, a context modal displays all data set variables, 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 variables also simplifies the approach and also works as spell-check of variable 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 available, driving a *cloud-computing*-based approach. In theory, the calculation duration is reduced by a factor of 0.5 by every statistical processor. In praxis, 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 heart of the prototype. Results from the statistical processors are uploaded into the visualization by slice, allowing the assessment of the data as soon as parts of the calculations are finished while the rest is still in processing.

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, $Z X + Y$ turns into $Z X$ 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 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. It serves the purpose of *Show the Results* after the *Analyze First* step in Keim's VA-Mantra. The displayed prism shows values above the matrix diagonal.

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 as much as possible. Since the R^2 values are mapped on data point opacity, only large values are highlighted in the prism, guiding the focus to the respective slices. It also creates a very 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 affect to the readability of the visualization.

Rotating with the cube is restricted to 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 does not add value to the visualization. We provide also 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 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 current 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]. [Add Figure reference!](#) The 2D mosaic visualization 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 model displays detailed information about a models coefficients, associated p values and confidence intervals.

6 IMPLEMENTATION

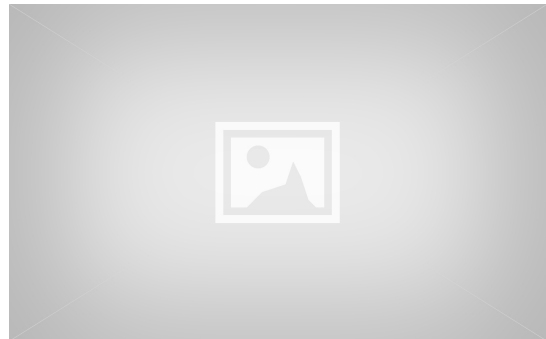


Fig. 4. [Describe the Bootstrap/Angular/D3 Frontend, Node backend, R/OCPU Backend.](#)

Using web technologies for prototyping allows for:

- easy exchange of software, without the need of installing anything other than a modern web browser,
- fast iterative adaption of method changes, without the need of installing updates,
- outsourcing computation-heavy tasks to server-clusters, which do not have to be in local proximity.

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.js¹ abstracts web application into models and views, allowing for a responsive way to combine HTML and Javascript. Angular.js is forcing 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 Bootstrap², which also provides a rich set of UI elements with proper stylings. The 2D mosaic plot is implemented using Data driven Documents (D3.js), which is very popular in information visualization using vector graphics [6]. It provides fast and easy methods for binding data to graphical elements. The 3D plot is created using the WebGL-based Threejs³ 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). We 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.

¹Open Source; Maintained by Google, angularjs.org

²Open Source; Maintained by Twitter, getbootstrap.com

³Open Source; Originally developed by R. Cabello, threejs.org


Back-End. Two server structures serve as back-end. The first one is the web-server, which is written in Javascript using NodeJS⁴, running on Google's V8 Javascript runtime environment. It is hosted on Heroku⁵, a cloud application platform.

The statistical processors yield the  structure. They rely on the statistical programming language R⁶, which is widely adopted in the statistical analysis community, yielding a rich support of fast state-of-the-art statistics algorithms as well as newly published methods. OpenCPU is an R package and provides an API for accessing it via HTTP calls [26]. 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 homepage.

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 foundation 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 extend. Usually, only  features depicting a hypothesis are compiled into a data set to assess them using statistical tools. 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

To assess the ability of a system to discover knowledge is ~~hard to measure~~. Lam et al. [22] propose ~~for this purpose~~ the *Visual Data Analysis and Reasoning (VDAR)* technique, which is focused on the characterization of a system's 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 conducted a study with two participants, who also co-authored this publication. UN is a data scientist who analyzed the hepatic steatosis data set by extracting decision rules. KH is a radiologist (10 years of experience) with specialization in epidemiological research and is also responsible for the SHIP MRI acquisition.

Participants, Setup and Procedure ~~One major advantage of focusing on web-based research was the ability to bridge the geographical gap towards the epidemiological domain experts.~~ We conducted a web-based research by using an online-meeting software, which features voice chat as well as screen sharing. Starting an analysis using these techniques take about 5-10 minutes of setup time.

The sessions started with an initial overview of the system, showcasing its features and functionality. After that 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.

Domain expert for the breast fat data set analysis is KH, a clinician (10 years of experience) with focus on epidemiological research. KH 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 scientist responsible for prior analysis of the data with.

TI, a statistician with focus on epidemiology (TODO 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 Breast Fat Data Set

The breast fat data set was compiled to find associations between parenchyma tissue proportion in the female breast towards 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% [23].

The data comprises 1.186 subjects (368 from SHIP-2, 818 from SHIP-TREND-0 cohort). It contains 231 dimensions, holding information about

- *somatometric features*, e.g., body size, weight or BMI,
- *lifestyle features*, e.g., alcohol/tobacco consume,
- *personal history*, e.g., occupation, marital status,
- *medical history*, e.g., current or prior diseases or laboratory values, such as cholesterol levels,
- *women specific features*, e.g., number of born children, contraception type or hormone replacement therapy and
- *mammography features* derived from the image data.

The latter were derived using a level set based segmentation [17] of MRI image data for each subject. Check this with Mrs. Hegenscheid, probably most of them were segmented by hand!

The data of each cohort was presented as individual SPSS files. All data related to the mammography was 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 expressive, e.g., *chro_09a* is now denoted as *Disease-Osteoporosis*. This avoids the need of defining a separate data dictionary file for the *Regression Cube*, translating the feature names. All male subjects were removed as their data does not contribute to the analysis.

7.3 The Hepatic Steatosis Data Set

We employ a subset of the data set which was used in [25] to identify predictive features with respect to the reversible disorder hepatic steatosis (fatty liver). The binary target feature is derived from the liver fat concentration which was measured with Magnetic Resonance Imaging (MRI). In particular, participants with a liver fat concentration of no more than 10% are mapped to the 'negative' class; values greater than 10% are mapped to the 'positive' class. The data set contains labels for 578 participants for which the MRI data was available to that point. Note that MRI was performed only in SHIP-2, so the target feature is present for just one of three study waves.

Apart from the target feature, the data set contains 199 features extracted from participants' questionnaire answers and medical tests. They are features on socio-demographics (sex, age, etc.), features on consumption behavior (e.g. alcohol and cigarettes), SNPs, features extracted from laboratory data (e.g. sera concentrations), and two features on the results of the liver ultrasound. The names of 85 features have the suffix *s0* meaning that their values have been recorded in SHIP-0 (first study moment), 50 features exhibit the suffix *s1* and 55 features have the suffix *s2*, alongside 10 time-independent SNPs.

In [25], the authors show that women and men exhibit differences in the class distribution of the liver fat concentration. Also, for women

⁴Open Source; Maintained by Joyent Inc, nodejs.org

⁵Owned by Salesforce.com, heroku.com

⁶Open Source; r-project.org

⁷R-based back-end:

github.com/paulklemm/regression-cube-r-package

⁸Front-End and Nodejs Webserver:

github.com/paulklemm/regression-cube-prototype

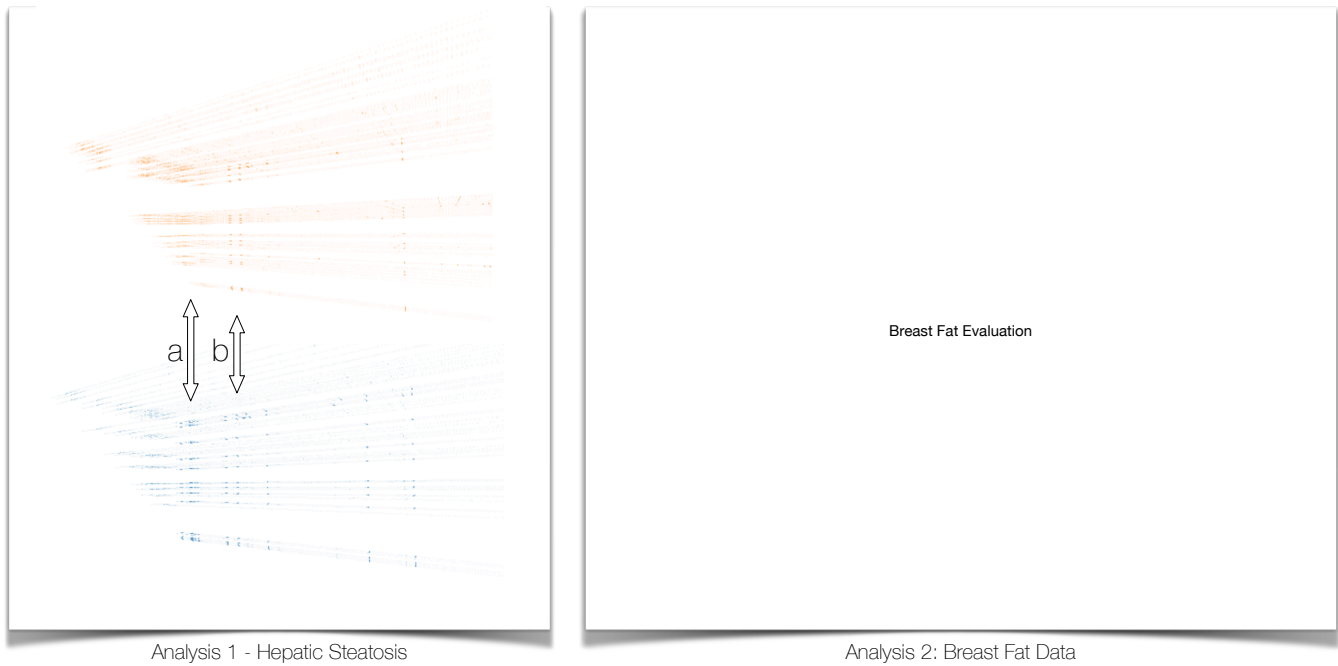


Fig. 5. [Screenshots of Cubes from the Evaluation](#)

they identify that there is a association between age and liver fat concentration. They compute an appropriate cut-off value of 52 years for which the class distribution is most homogeneous within the resulting subsets. Based on this observations, we perform our analysis on three sub-populations, *females*, *males* and *females* older than 52 years.

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

The analysis goal is to find relationships on the breast cancer data using features derived from the mammography analysis. Relationships between the share of parenchyma tissue on the overall breast volume are of high interest [23].

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

We refer to each analysis step regarding to their belonging in the VA-Mantra (Subsec. ??). The analysis goal is reproducing results based [25]. Therefore, *UN* started the **AF** step using the dichotomized MRI fat liver concentration using the formula $mrt_liverfat_s2 \sim X + Y + Z$ for *male* subjects. The **SI** step using the 3D cube representations hotspots at the end of the cube (Fig. ?? a). **ZF** by slicing through the cube using the mouse input to inspect the hotspots using the mosaic plot (**DD**) revealed high correlations for somatometric features, hepatic steatosis indicator-features as well as laboratory values, such as creatinine ([important for kidney](#)) and uric acid ([gout, diabetes](#)) magnitudes. Similar results present for analyzing the *female* groups. *UN* could reproduce most results, some features features exhibit lower correlations though, e.g., creatinine magnitudes. Relationships not described in [25] 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 [25], a complete matching set of correlating features are not expected.

Analysis of non-discretized target feature. Since our method can analyze numerical target features, as opposed to decision trees used by *UN* before, the analysis was conducted again for the not-dichotomized target using the same formula. The 3D-cube representation already showed generally lower R^2 values, but since the analysis is now based on linear regression, the R^2 values can not be directly compared. Correlation hot spots matched with the ones from the dichotomous target,

but were generally lower (R^2 of *XX* for somatometric features as opposed to *XX*). We conclude that the bias introduced by dichotomizing the fat liver content enforces the findings towards liver diseases, while using the numerical features is less expressive.

Interleukin-6 correlation with liver fat. During the analysis, one hotspot was always observed in the *SI* and *ZF* incorporating a high *Interleukin-6* (*IL-6*) correlation with liver fat values (Fig. 5 b). The correlation was high for both the dichotomized and continuous target feature. In the literature, relations between liver cancer [1] and also chronic liver diseases [33]. Strong effects of *IL-6* with hepatic steatosis were also prior described with mice [15]. The finding is subject of further analysis.

7.6 Further Feedback and Lessons Learned

Extracted Hypotheses have to investigated further. [Tills Feedback - Regression analysis also incorporates assumptions about the data,](#)

Overview Visualizations are preferred over Black-box Methods. While the classical epidemiological workflow (see Sec. 2) represents the foundation of the research in this area, explorative analysis based on the data gains importance. Incorporating data mining algorithms with statistically established and overview visualizations allow for reasoning about risk factors using underlying epidemiological data sets. One major advantage we observed is the confidence of an expert in the results he or she actually observe. Their participation in the analysis using human pattern-detection and expert knowledge is for experts preferred over results derived using a black-box method. By observing expected correlations matching the expert knowledge strengthens the confidence in the method and subsequently in the hypothesis generated from unexpected relationships.

Using non-discretized features reduces information bias. [Comment on the lesson learned from the last VAST Paper - "Usage of different categorizations depending on expected outcome." Now we have continuous data.](#)

Attention steering is crucial. [Bring important events to the users attention, otherwise it might be overlooked!](#)

Support for different subject groups. [Include this? Feedback from the evaluation goes here.](#) [Time-aspect critical, interactive analysis requires for fast response.](#) [The method needs to be speeded up.](#)

8 SUMMARY AND CONCLUSION

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 most important features. *Hypothesis-free* analysis employs default regression models. Model expert knowledge using regression formulas allow for *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 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 an analysis platform open to heterogenous data types. We believe in the power of opening up knowledge discovery to allow a heterogenous 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.

REFERENCES

- [1]
- [2] 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. ~~IEEE Computer Society~~, 2011.
- [3] G. Albuquerque, M. Eisemann, T. Löwe, and M. A. Magnor. Hierarchical brushing of high-dimensional data sets using quality metrics. In *VMV-2014 - Vision, Modeling & Visualization*, pages 119–126, 2014.
- [4] P. Angelelli, S. Oeltze, C. Turkay, J. Haasz, E. Hodneland, A. Lundervold, B. Preim, and H. Hauser. Interactive visual analysis of heterogeneous study data. *IEEE Computer Graphics and Applications*, 2014. 
- [5] 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.
- [6] M. Bostock, V. Ogievetsky, and J. Heer. D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.
- [7] 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.
- [8] 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.
- [9] X. Dai and M. Gahegan. Visualization based approach for exploration of health data and risk factors. In *Proc. of the International Conference on GeoComputation. University of Michigan, USA*, volume 31, 2005.
- [10] J. W. Emerson, W. A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H. Wickham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91, 2013.
- [11] R. H. Fletcher, S. W. Fletcher, and G. S. Fletcher. *Clinical epidemiology: the essentials*. Lippincott Williams & Wilkins, 2012.
- [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] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. *Visual analytics: Scope and challenges*. Springer, 2008.
- [19] 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 10th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 85–92, Mar 2015.
- [20] 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.
- [21] 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.
- [22] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.
- [23] 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.
- [24] U. Niemann, T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kuhn. 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)*, So Carlos and Ribeiro Preto, Brazil, June 2015. ~~IEEE~~, accepted 03/2015.
- [25] 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.
- [26] J. Ooms. The **opencpu** system: Towards a universal interface for scientific computing through separation of concerns. *Computing Research Repository - arXiv*, abs/1406.4806, 2014.
- [27] H. Piringer, W. Berger, and J. Krasser. Hypermoval: Interactive visual validation of regression models for real-time simulation. In *Computer Graphics Forum*, volume 29, pages 983–992. Wiley Online Library, 2010.
- [28] 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.
- [29] 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.
- [30] T. Schreck and D. Keim. Visual analysis of social media data. *Computer*, 46(5):68–75, 2013.
- [31] B. Shneiderman, C. Plaisant, and B. W. Hesse. Improving healthcare with interactive visualization. *Computer*, (5):58–66, 2013.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] J. W. Tukey. *Exploratory data analysis*. Reading, Ma, 231:32, 1977.
- [37] 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.
- [38] 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.
- [39] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Information Visualization*, page 1473871614526077, 2014.