3D Regression Heat Map Analysis of Population 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 to define disease-specific risk factors. These data contain information (*features*) about a subject's lifestyle, medical status 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*).
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* w.r.t. a *target feature*. It provides a visualization that reveals insights into the data by highlighting relationships. The *3D Regression Heat Map*, a novel 3D visual encoding, acts as an overview of the whole data set. It shows all combinations of two to three independent features with a specific target disease. Slicing through the *3D Regression Heat Map* allows for the detailed analysis of the underlying relationships. 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 *3D Regression Heat Map* method to a hepatic steatosis data set to reproduce results from a data mining-driven analysis. A qualitative analysis was conducted on a breast density data set. We were able to derive new hypotheses about relations between breast density and breast lesions with breast cancer. With the *3D Regression Heat Map*, 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, Regression Analysis, Heat Map, 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 [11]. 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. The hypotheses are depicted using epidemiological features and are then statistically analyzed.

An important epidemiological tool for deriving such features are *cohort studies*, such as the Study of Health in Pomerania (SHIP) [42]. 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.

Testing features for associations with diseases using regression models is one of the most important epidemiological tools. Using regression analysis 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 visualizations, possible correlations may be missed. Explorative analyses and overview visualizations of the data set as presented in prior work [23], are not tailored to a specific target feature. They 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 w.r.t. a specific disease or hypothesis. For this purpose we provide 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 design based on feedback of epidemiological domain experts to support hypothesis generation w.r.t. a target feature using regression models.
- Incorporation of prior domain knowledge by using freely adjustable regression formulas.
- Metrics selection for analyzing regression models and for details-on-demand representations.
- An open-source web application that can be used with data of different application domains.

2 Epidemiological Background

This section covers the epidemiological workflow, epidemiological data in general and the SHIP in particular.

2.1 Epidemiological Workflow

Epidemiological research is performed by experts from different academic disciplines, such as epidemiologists, 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 with statistical methods. As described by Thew et al. [38], the epidemiological workflow is divided into these different steps:

1. Clinicians/epidemiologists derive hypotheses via clinical observations, experimental studies or literature research.
2. Epidemiologists compile a list of features depicting the hypothesis and include confounding features.
3. Statisticians assess the association of the derived features w.r.t the investigated disease.

Relative risks can be determined if a statistical resilient association of features with a condition is extracted. Derived absolute risks indicate the per-subject chance of developing the disease. Reproducibility of

- *Paul Klemm, 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*
- *Kai Lawonn is with Otto-von-Guericke University Magdeburg, Germany and Delft University of Technology, Delft, Netherlands. E-mail: lawonn@ovgu.de*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

results is an epidemiological key requirement and guides all analysis 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.

We described an Interactive Visual Analysis approach for image-centric cohort study data, which connects to the feature listing step [23]. The methods aim to derive hypotheses through analysis of the data and observations about previously unknown feature correlations. *Hypothesis generation* benefits from 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 epidemiological instruments available to acquire data. We focus on population studies, which impose the largest data sets and yield a highly heterogenous and incomplete information space. Subject data are collected either regarding specific diseases or with the widest range possible. The latter allows the data set to be assessed w.r.t. different diseases. The feature space comprises information about lifestyle, somatometric features, medical parameters, genetic data as well as medical images. These features are gathered through questionnaires, medical examinations and 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 reduces the information space and introduces an information bias, as assumptions are modeled through the discretization.

Medical image data are also analyzed in modern population studies. We incorporate image-derived features, but do not focus on analyzing image data. More discussions of population study data types and characteristics can be found in the works of Preim et al. [34] and Toennies et al. [39].

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 w.r.t. the confounder. *Age* is included as confounder in almost any epidemiological analysis, since most diseases are more likely for older subjects. It also influences the general body condition and thereby almost all features acquired through population studies. Another important confounder is *gender*. Confounders have to be selected by epidemiologists specific to the investigated condition.

The Study of Health in Pomerania (SHIP). The SHIP, located in Northern Germany, aims to characterize health and disease in the widest range possible [42]. Unique for the SHIP is the acquisition of medical image data. 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. Most examinations occur in all stages and are performed according to the same instructions to enable comparisons over the different stages of the study.

2.3 Regression Analysis

Regression analysis is the most important statistical tool when analyzing epidemiological data and is 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*). The *Akaike Information Criterion* (AIC) metric estimates the quality of a model by estimating its information loss of modeling the underlying data and is suited for comparing models [2]. A small AIC value indicates a higher quality 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: $Dependent \sim Independent_1 + ... + Independent_n$. 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. $x|y$)

Due to the different meaning of operators in regression formulas, feature transformations are not available within this notation. The class of the target feature restricts the regression type. Different regression types are available; we focus on the following for describing linear relationships:

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 *continuous type*. Linear regression models can also be described using the *adjusted* R^2, which considers the number of dependent features. It yields lower values, when features with little entropy w.r.t. the target feature are included. The *f-statistic* measures the improvement of the model if independent variables were added.

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 [29], 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.

3 Prior and Related Work

Tukey already stated in 1977 that data are too often analyzed solely using a confirmatory data analysis [40]. 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.

Parameter space analysis using regression models. Sedlmair et al. [35] presented a conceptual taxonomy of parameter space analysis. Based on their input parameter taxonomy, we are visualizing *model parameters* based on *environmental parameters* in a *global-to-local* navigation strategy. We solve a *fitting* task by aiming to find models well suited for describing the input data. The approach of Mühlbacher et al. [28] is closest to ours. They provided a framework for qualitative analyses of relationships and ranking features for numerical target features with regression models. Existing regression models can be validated and compared using 3D views and 2D slice views. Mühlbacher et al. focused on a smaller number of features, which can be assessed in more detail, yielding a plot matrix view, while we cover more features by abstracting the models.

Similarly, Piringer et al. [33] proposed methods for visualizing regression analysis results and properties for developing car engines. Their main goal is to assess the pairwise influence of independent features w.r.t. the target feature using a plot matrix displaying models as contours. 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. Guo et al. [13] presented multi-space visualizations to find linear relationships in the data with focus on extracting groups of best fit. The data space is visualized using a scatter plot matrix. Linear models are calculated by defining dependent and independent features. The model view allows

for assessing different models by color-coding distances to the line-of-fit. The model parameters can then be fine-tuned using line graphs, histograms and model projections. Chan et al. [7] propose the 'Regression Cube', 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 Piringer et al. [33], 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. Piringer et al. [33], Guo et al. [13] and Chan et al. [7] focus on finding and tuning a model for a specific relationship, while we search for models w.r.t. all combinations in the data set. Instead of analyzing one complex model in detail, we process a large amount of models in terms of *different* features.

Visual analysis of epidemiological data. The work of Zhang et al. [12, 43] is closest to ours regarding the application of a visual analysis of population study data. They present *Cohort Analysis via Visual Analytics* (*CAVA*), a framework that distinguishes three major elements of a 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. We incorporate their requirements, which consist of a *flexible* and *iterative analysis*. Angelelli et al. [4] visualize image-derived an non-image data using cube data structures with focus on comparison and knowledge extraction. They use *Pearsson's r* to characterize relationships with target features and employ list views and scatter plots to visualize and rank them. In contrast, we focus on a fast large scale correlation analysis incorporating many features to derive insights.

Steenwijk et al. [36] focus on the *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. 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. Maries et al. [26] proposed GRACE, a framework for visually exploring correlations of features with spatial and non-spatial geriatric data. Sparse Partial Least Squares Regression and Tikhonov Regularization are used to produce predictor subsets and quantify correlation strengths. A multiple view system links spatial with non-spatial data and correlation coefficients are visualized using Kiviat diagrams, correlations are plotted using p-values in the other views. Instead of applying several linked views, we focus on one view of abstracted data.

Generalized pairs plots (GPLOM'S) extend the concept of scatter plot matrices by the pairwise depiction of heterogeneous data using type-combination-dependent visualizations [10, 19]. Dai et al. [9] 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 are visualized by Chui et al. [8] using multi-panel graphs highlighting risk factor differences with age and gender with regard to influenza- and salmonellosis-associated hospitalizations. We incorporate the idea of the arrangement of feature combinations as matrix as well as assessing confounding features.

Statistical analysis. Bertini et al. [5] present an overview of quality metrics describing high-dimensional data. Their refer to their research agenda in our design. Namely, we apply aspects of perceptual tuning with 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. [3] 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. [41] follow a similar approach by using both descriptive metrics for features as well as the features themselves and incorporate them into linked plots.

Niemann et al. [31] 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. Niemann et al. [30] improved 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 further investigate findings presented by Niemann et al. [31].

Prior work. In our prior work, we analyzed the healthy aging process of the lumbar spine. We defined an Interactive Visual Analysis workflow for image-centric population study data [23]. We extended the feature selection step (recall Sec. 2) with an iterative analysis loop incorporating group selection using expert input as well as clustering methods. Information visualizations of population study features were augmented with extracted image data, yielding linked views with both image and non-image information. The 2D heat map [23] 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. The presented techniques were well received by the epidemiological experts, but the explanatory power w.r.t. back pain was limited. It yielded an analysis based on image-derived features, such as extracting and analyzing curvature, torsion and angle of the lumbar spine [22]. We concluded that the model quality is insufficient to characterize back pain.

The difference of the presented approach in comparison with 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 3D Regression Heat Map Analysis of Population Study Data

The *3D Regression Heat Map* is designed to provide an overview visualization to support hypothesis generation. Hence, it is associated with step 1 and 2 of the epidemiological workflow (recall Section 2.1). Relationships observed using such techniques are subject of detailed statistical testing by statisticians with background in epidemiology using statistical processors, such as SPSS.

4.1 Iterative Design Based on Expert Feedback

The *3D Regression Heat Map* design was developed iteratively based on feedback of epidemiologists by using the prototype in joint analysis sessions on their data sets. The idea emerged from analysis sessions of a previous project, which contained a 2D heat map showing pairwise feature correlations based on *Cramér's V* contingency values [23]. It allowed them to reproduce their knowledge about relationships by observing correlations they would expect as well as discovering new correlations. In epidemiology, these relationships are also of interest, but rather w.r.t. their explanatory power on the target feature. This target often indicates the presence of the investigated disease. The domain experts wanted to *model knowledge* about the investigated condition, such as confounding features (e.g., age or gender). For explorative analysis, they preferred an approach which highlights associations w.r.t. various target features to both check for medical soundness of the data as well as detecting unexpected relationships. Additionally, due to the sensitive nature of population study data, the data has to be handled securely. Technical measures to enable a secure transfer and storage are described in Section 5.

Regression analysis is the statistical tool of choice for analyzing relationships in epidemiological data (recall Sect. 2.3). A regression model is based on expert knowledge. There is no rule how to apply models to a given set of features. Thus, they have to be applied with care.

4.2 Regression Heat Map Description Using Regression Formula Notation

Expert knowledge modeling is carried out using *regression formulas*. The formula input influences the type of the chosen regression method as well as the *independent* features describing the target (recall Sect. 2.3).

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 $Cancer \sim X + Y$ yields n^2 (10,000) regression models, describing all combinations of two features describing *Cancer*. This notation is 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 w.r.t. 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, the calculation lasts roughly 14 h.

4.3 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 w.r.t. the target feature. Identifying irrelevant features and excluding them from the feature space considerably reduces computational costs and yields a comprehensible *3D Regression Heat Map* representation. The correlation-based feature selection (CFS) [15] 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 w.r.t. 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 w.r.t. 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 features.

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. The tradeoff involved using the CFS algorithm is the potential removal of interesting features for the domain expert. This problem is discussed in the next subsection as part of the 3D representation of the regression results.

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.

4.4 Abstracting Regression Results

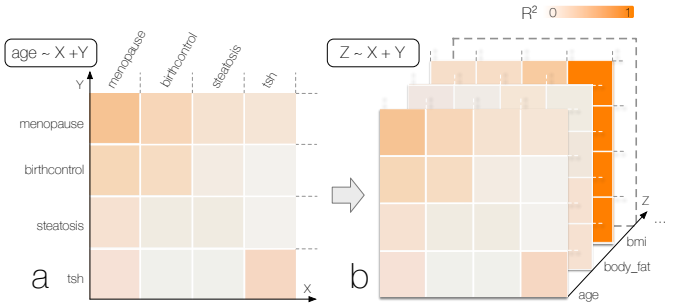

Fig. 1. (a) Overview visualization using a 2D heat map of the formula $Z \sim X + Y$, where Z assumes the feature *age*. The R^2 metrics extracted from the regression model are mapped to color saturation (a saturated color indicates a strong correlation). (b) Now, Z is set to all features n and yields n 2D heat maps. These represent the slices in our *3D Regression Heat Map*. The metric describing the regression model 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 [5]), which is easy to understand. As described in our previous work [23], correlation values scaled between 0 (no correlation) and 1 (perfect correlation) can be encoded with color in a 2D heat map. Regression models are more complex, having many associated describing metrics. For the *3D Regression Heat Map* analysis we are interested in the quality-of-fit of the resulting model. This allows to infer the predictive quality of the independent features included in the model. The R^2, adjusted R^2 and AIC value are metrics allowing for this kind of assessment (recall Sect. 2.3).

2D (slice) view. Since R^2 is scaled between $[0,1]$, it allows for comparison *between* regression models. We can apply the same 2D heat map by translating the R^2 values to *color saturation* (Fig. 1a). This encodes a 2D regression square for dynamic variables X and Y (e.g., $Age \sim X + Y$). Based on expert feedback on early versions of this view, the amount of features used to compare regression models was extended. Therefore, these experts can investigate the heat map with emphasis on specific aspects of the model. *Adjusted* R^2 can be represented in the same way, since they are also scaled between $[0,1]$. *AIC* values have to be normalized in order to map them on color saturation. The resulting scale may be distorted by outliers derived from poor regression models. To tackle this problem, we provide a slider input, which maps the transfer function of the metric to color saturation based on user-selected ranges. Outliers can be cut off to emphasize ranges of interest. Small *AIC* values indicate a good model. Hence, we inverted the transfer function color mapping, assigning low *AIC* features to saturated colors. To include users unfamiliar with these metrics, the *Regression Heat Map* is set per default to show R^2 values.

3D view. Introducing Z creates a 3D heat map (Fig. 1b). The selected metric (per default set to R^2) of each heat map entry (*voxel*) is mapped to opacity to reduce the overlap. Object size is not used to encode information because it would result in a cluttered view. Epidemiologists argued that the visualization of descriptive metrics derived from different regression methods (e.g., $Z \sim X + Y$) is misleading, as they can be compared relatively, but not in precise

numbers. Therefore, we decided to map metrics of different regression methods on distinct colors (i.e., orange for linear regression and blue for logistic regression). Thus, the visualization can be easily extended using other regression types. For *3D Regression Heat Maps* 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. As mentioned previously, the feature reduction using the CFS algorithm potentially removes important features. The z dimension of the visualization contains *all* features of the data set, allowing to assess their influence. The x and y dimensions are restricted to the features extracted from the CFS algorithm.

Our goal is to create an overview visualization for a data set. 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 *3D Regression Heat Map*.

4.5 Analysis Workflow

Our *3D Regression Heat Map* is well suited for different workflow analysis techniques, based on the Visual Analytics (VA) Mantra of Keim et al. [21]:

Analyze first [**1.AF**]. Choosing an initial regression formula triggers the *3D Regression Heat Map* calculation, filtering the dimensions of the dependent feature through the CFS algorithm.

Show the important [**2.SI**]. The 3D visualization acts as an overview over the whole data set. Here, regression models with large regression metric 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 heat map 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. The workflow is highly iterative. Observations in the 2D heat map 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 *3D Regression Heat Map* for a difference view.

Hypothesis-free and hypothesis-based analysis. Early analysis sessions yielded two approaches of analyzing the data. The classic approach is *hypothesis-based*, where the expert already knows the data and potential associations (e.g. reproducing knowledge about hepatic steatosis risk factors based on known risk factors). The *hypothesis-free* analysis allows to derive new insights, such as identifying confounding features or potential targets (e.g. deriving risk factors for breast cancer-associated features). *Hypotheses* about the data are reflected using input formulas. Using the operators, dynamic variables and data set features, many different assumptions can be expressed. To support the *hypothesis-free* analysis, we provide a default formula:

$Z \sim X + Y$. It represents all possible combinations of two independent features w.r.t. 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 exploratory analysis.

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 with the target are analyzed through this *3D Regression Heat Map*.

$Cancer \sim X + Y + Z + feature_1 : feature_2$ 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, this interaction is incorporated for all X, Y and Z values as independent features.

$Cancer \sim X + Y + Z$ subtracted with the regression metric from $Cancer \sim Age$ excludes the confounding effect that age has in view of the target *Cancer* feature. This is achieved through *3D Regression Heat Map* comparison.

3D Regression Heat Map comparison. Comparisons were introduced later in the project. Epidemiologists with focus on statistics pointed out that comparing outcomes of different formulas is suitable for removing the effect of possible confounding features. *3D Regression Heat Maps* can be compared by creating difference views. One formula acts as reference. The absolute difference in the regression metric values with the second formula is calculated. For example, it can be utilized for comparing the influence of a single feature on the complete result (e.g., $Z \sim X + Y$ and $Z \sim X + Y + Income$).

5 System Design

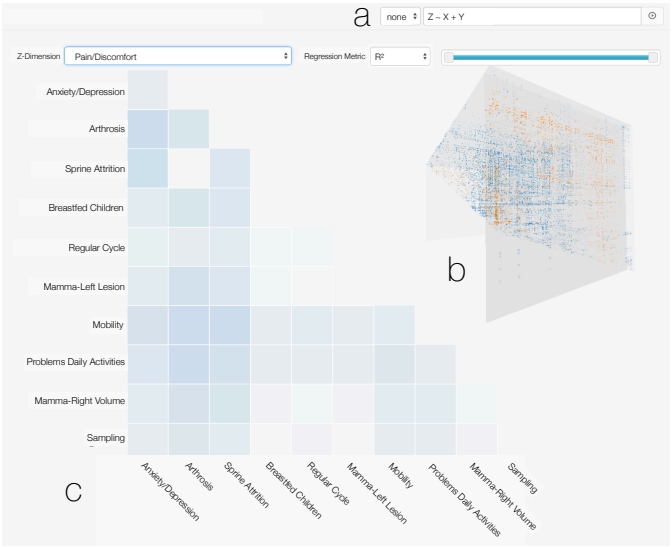

Fig. 2. Breast density data set loaded into our prototype. (a) Using the formula input, the user specifies the dependent feature and calculation rules. (b) 3D heat map showing values above the matrix diagonal as overview. The values of the currently selected slice are mirrored and represented as orange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

We designed our system to be openly 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 with other (statistical) tools. This is, however, out of the scope of this work. Therefore, the system has to be intuitive and comprehensive in order to be adapted by domain experts.

Using web-based technologies offers various advantages w.r.t. the collaboration with epidemiologists. They usually have little time to wrangle software. A web-based approach has no set-up time besides loading up the data set and can be carried out with any computer connected with the web. We can even implement small changes based on feedback of domain experts directly during analysis sessions. By providing a service using a website it has a much larger chance of being tested and potentially adapted by a broad user base. Web technology is based on a client-server architecture. It allows for outsourcing computationally heavy tasks on server clusters and transferring results to the client device. This architecture is also prone to security issues, such as the storage of confidential data, especially in the epidemiological context. Therefore, we have to incorporate technical measures to ensure a secure workflow.

5.1 System Paradigm and Components

Epidemiologists will not adapt complex systems that require substantial training and time. Therefore, the *3D Regression Heat Map* design focuses on a clean appearance, reducing the amount of user interface elements as much as possible. This allows for a fast learning of the system. Our prototype consists of three components:

- The *file upload* section starting the analysis with providing a comma-separated value (CSV) file [**1.AF**].
- The *Regression Heat Map visualization* consisting of the 2D heat map as well as a 3D representation of all regression models with facilities to change the represented regression metric and its range [**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 difference models [**1.AF**, **3.ZF**].

File upload and classification [**1.AF**]. Popular analytics tools, such as WEKA [14], owe part of their success to their support of open file types. To allow other users even outside the epidemiological application domain to access 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. Using a check box, the user can disable the CFS preprocessing step, which is useful for small data sets where the user does not want to reduce the number of features.

Data 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 is carried out 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 *Regression Heat Map* **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 2 by every statistical processor. In practice, data transmission and differences in machine specifications always influence the speed.

Difference heat maps can be generated for each formula added to the system. Using a dropdown menu it can be selected as reference. Since all cells in the heat map are represented using regression metric values, the difference is the absolute difference of regression metric for each cell.

5.2 3D Regression Heat Map Visualization [2.SI].

The visualization and interaction with the *3D Regression Heat Map* is the core of the prototype. Results from the statistical processors are uploaded into the visualization slice by slice. This allows 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 2D heat map matrix. This is due to the symmetry of basic regression operators. 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 represent 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.

Selecting and scaling the descriptive regression metric. The feedback made apparent that other features are of interest for analyzing regression models too. Hence, UI elements for controlling them were introduced. The descriptive metric shown in the 2D/3D view can be selected using a dropdown menu. The default selection is R^2. *AIC* displays model quality. *Adjusted* R^2 values are only available for linear regression. Logistic regression results are represented via R^2 values in this mode. As they are visually distinguished using color, confusions are avoided. The transfer function of the color intensity (2D) and opacity (3D) can be adapted using a slider input. This allows to filter models with desired features, such as only very high R^2 values.

3D prism as data mini-map. In early prototype versions, the 3D prism acted as starting point for the data analysis without the implementation of a separate 2D view. Slices were shown using cutaway planes. This approach was not popular among epidemiologists, because the complexity of the visualization overwhelmed them. The *3D Regression Heat Map* representation was redesigned to act as an overview over the whole data set. It serves as a function similar to a mini-map, guiding the attention to points of interest in the data. It also gives context information about adjacent data values when using the 2D heat map. The distinction between overview and details-on-demand using two different representations was well received with our domain experts. 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 the absolute regression metric values (Fig. 2b). 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 regression metric values and the global mean along the z-axis. This highlights slices with unusually low or high results (Fig. 4). Variables are ordered the same way in the 2D and 3D heat map to preserve the mental model and make them visually analogous.

Tackling the disadvantages of 3D information visualization. 3D information visualizations are criticized for introducing occlusions and interaction problems. These are often not balanced out by the advantages of using the third dimension for visual mapping. We aim to minimize these problems. The regression metric (e.g. R^2) values are mapped on data point opacity, highlighting large values in the prism, which guides the focus to the respective slices. The visualization is sparse, since the majority of the 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 heat maps. Overlapping is still an issue, but greatly reduced in its effect to the visualization readability.

Transformation of the 3D heat map is restricted to the y-axis (horizontal only), preserving the mental map to position individual features. The 3D heat map is always oriented according to the 2D representation, allowing for an easy mental combination of them. Allowing more degrees of freedom was confusing to our users and also did not add value to the visualization.

3D heat map slice selection [**3.ZF**]. In order to *Zoom, Filter and Analyze Further*, the user has to navigate to different slices of interest. We propose two ways to achieve this.

We apply the slicing metaphor from 3D volume data. In medical volume renderings, slicing views are common to view details on a selected plane in the scene. We employ this technique for selecting 3D

heat map slices (e.g., by moving a plane via vertical mouse input while pressing the right mouse button). However, we still display the whole 3D object instead of cutting away information. Early prototypes only provided this method to select a slice of interest, which was inefficient when the user was looking for a specific slice. Hence, an additional method was implemented.

Selecting the slice using a dropdown menu containing the feature names provides fast access to plane selections when the user already knows the slices of interest. The currently selected slice is displayed as a semi-transparent gray plane. Early prototypes rendered the whole *3D Regression Heat Map*, which made it hard to assess the position of the plane. Since the regression metrics are mirrored along the diagonal, the space available from visualizing only the prism generated from the upper half of the heat map diagonal is used to display the 2D heat map of the currently selected plane. The regression metric values are projected on this plane to provide an overlapping-free view. This allows for a easier to identify the current slice.

2D heat map slice visualization [**4.DD**]. The 2D heat map (Fig. 2c) shows all values below the matrix diagonal of the current slice. It creates an optical equivalence with the 3D heat map. 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 heat map.

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, confidence intervals, f-statistics and AIC values. It also contains a scatter plot of the *model residuals*, which shows the difference between the observed data points with the fitted values. Epidemiologists use such plots to validate models w.r.t. the model assumptions, such as homogeneity, normality, and independence [25].

6 Implementation

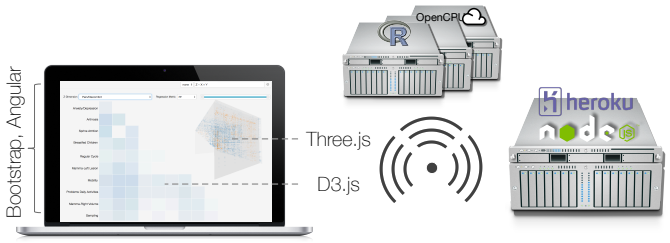

Fig. 3. The front-end (left) is realized with `HTML5/CSS3/Javascript` and different Javascript libraries, such as `Angular.js`, `Three.js` and `D3.js`. The web server (right) is written using `Node.js` and hosted on `Heroku`. `R` and `OpenCPU` constitute the statistical back-end (top) to compute the *3D Regression Heat Maps*. 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 front-end is created using `HTML5`, `CSS3` and `Javascript`. `Angular.js` abstracts web application into models and views, allowing for a responsive way to combine `HTML` and `Javascript`. It is easily expandable by forcing developers to write modularized code. `Twitter Bootstrap` handles the page layout and provides a rich set of user interface elements. The 2D heat map is implemented using the `D3.js` [6] information visualization library. It provides fast and easy methods for binding data to graphical elements. The 3D plot is created using the WebGL-based `Three.js` library.

Two server structures serve as back-end. The web server is written in `Javascript` using `Node.js`, running on Googles V8 Javascript runtime environment. It is hosted on `Heroku`, a cloud application platform. The statistical computations are performed on the second structure. They rely on the statistical programming language `R`. It is widely adopted in the statistical analysis community, yielding a rich support of 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 [32]. This way, any computer which runs `R` can be turned into a statistical processor for our project.

A running instance of the *3D Regression Heat Map* prototype can be found under `regressionheatmap.herokuapp.com`. The source for the prototype is freely available at `Github`.[1,2] Instructions and code to setup running the statistical back-end through a `Ubuntu` server using `OpenCPU` are included in the repository. The front-end can be deployed using `Heroku` by cloning the repository into a Heroku app.

7 Application

In this section, we describe the application of the *3D Regression Heat Map* to two epidemiological data sets. The hepatic steatosis data set was analyzed using data mining algorithms, yielding risk groups, which we now analyze further. We try to reproduce the prior results from the analysis as proof-of-concept of our method. The female breast density data set is the basis for an explorative analysis w.r.t. the influencing parameters of the breast cancer-related parenchyma tissue ratio.

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 knowledge discovery capabilities of a system are difficult to measure. The *Visual Data Analysis and Reasoning* (*VDAR*) technique proposed by Lam et al. [24] 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 based on case studies using thinking-aloud techniques to comprehend the user's reasoning and thought process. 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 took 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 used the system on their own computers. The screen-sharing function was still used to observe the actions of the experts. 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 density data set. She is a radiologist responsible for the SHIP-MRI acquisition and also for the mammography analysis. The hepatic steatosis data set is analyzed by *UN*, a data scientist responsible for prior analysis of the data. The third participant is *TI*, a statistician with focus on epidemiology (8 years of experience), who 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 by Niemann et al. [31] 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 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. The MRI scans for each subject are only available in SHIP-2.

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

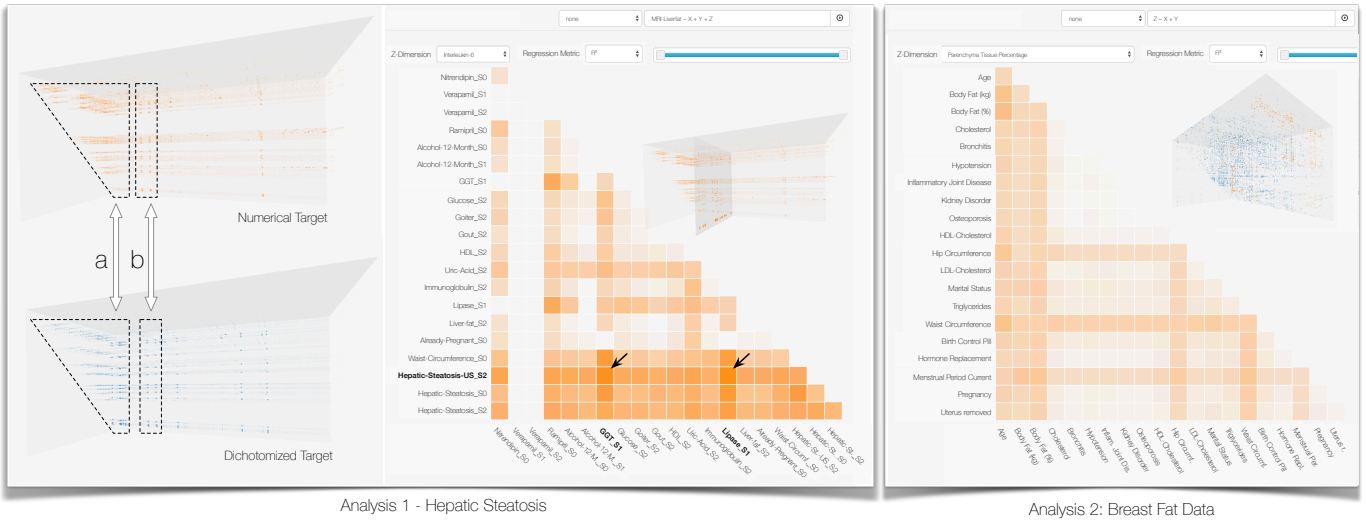

Fig. 4. The analysis of the numerical and dichotomized target feature depicting liver fat values yields similar results (left). In (a), hotspots for somatometrical features with high correlations were found. High correlations were also found for features depicting *hepatic steatosis* (b). A high correlation between *Interleukin-6*, *hepatic steatosis*, *GGT* and *Lipase* (highlighted using arrows) was revealed during the analysis using the 2D heat map. The hypothesis-free analysis of the breast density 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*.

Apart from the target feature, the data set contains 199 features comprising sociodemographic features (e.g., gender, age), consumption behavior (e.g., alcohol and 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). Niemann et al. [31] 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, which is the approximate entry age for the menopause 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 Density Data Set

The breast density data set was compiled to find associations between the parenchyma tissue proportion in the female breast compared to other features in the data. Breast density is denoted as the ratio between parenchyma and cellular connective tissue and 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% [27].

The data comprises 1,186 female subjects. It contains 231 features, holding information about somatometric features (e.g., body size and weight) consumption behavior, personal and medical history (e.g., occupation and prior diseases), women-specific features (e.g., number of born children and contraception type) as well as mammography features (e.g., fat content and parenchyma tissue proportion to volume). The latter were derived from MRI data for each subject, which was manually segmented by radiologists [17, 20].

The data of each cohort were 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. 4.5). The analysis goal was reproducing results with our visual analysis framework that are in accordance to the data mining-based results presented by Niemann et al. [31]. Therefore, *UN* started the [**1.AF**] step using the dichotomized MRI fat liver concentration and the formula $mrt\_liverfat\_s2 \sim X + Y + Z$ for *male* subjects. The [**2.SI**] step using the 3D heat map locates hotspots at the end of the heat map (Fig. 4 left). The Zoom, Filter and Analyze Further Step [**3.ZF**] was realized by slicing through the 3D heat map using the mouse input to inspect the hotspots. Analyzing the 2D heat map [**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* (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, 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 by Niemann et al. [31] 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 by Niemann et al. [31], 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 heat map showed lower R^2 values in general. However, the analysis is now based on linear regression and the R^2 values cannot be compared directly. 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, incorporating a high *Interleukin-6* (*IL-6*) correlation with liver fat values (R^2 of 0.8, see Fig. 4b). The correlation was high for both the dichotomized and continuous target feature. The literature described relations between *IL-6* and liver cancer [16] as well as chronic liver

diseases [37]. For mice, strong effects of *IL-6* with hepatic steatosis were described [18]. The finding is subject to further analysis.

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

The analysis aims to find relationships on the breast density data using mammography analysis features. Relationships between the share of parenchyma tissue on the overall breast volume are of high interest [27]. 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. 4 right). Women with higher *body fat* also have a larger *breast density 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* w.r.t. *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 formula $Parenchyma\_Percentage \sim X + Y + Z$ [**1.AF**]. Using the 3D heat map, *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 on the parenchyma tissue. A surprising finding was the strong correlation 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 using an extensive data set. Lastly, *KH* assessed the influence of contraception-related features, such as the 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 computations 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 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. The overall explanatory power on the numerical feature was lower, supporting the hypothesis that the dichotomized target feature already models knowledge to bias the data w.r.t. 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 the 3D heat map as supporting mini-map visualizations most useful for this purpose, e.g., for highlighting differences rather than displaying absolute values (Fig. 4).

8 Summary and Outlook

We presented a technique for knowledge discovery in population 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 Heat Map* 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 plot for each 3D heat map 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 density data set. The method was well received by our clinical partners, triggering detailed investigations of the findings.

The analysis is limited to three dynamic variables representing the *3D Regression Heat Map* dimensions. Investigating more dynamic variables can be achieved by projecting the high-dimensional space into a three-dimensional representation. This, however, increases the cognitive load and complexity of the analysis substantially and needs to be accompanied by techniques that simplify this approach. Static features can be added using the formula input without increasing the complexity of the visualization.

As a next step, we want to introduce more regression types, which model different kinds of correlations. We also want to extend the 3D heat map to time-dependent data by expanding the difference heat map 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.

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. This work was supported by the DFG Priority Program 1335: Scalable Visual Analytics. Sylvia Glaßer and Kai Lawonn are funded by the Federal Ministry of Education and Research within the Forschungscampus STIMULATE under grant number '13GW0095A'.

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] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19(6):716–723, 1974.

[3] 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.

[4] P. Angelelli, S. Oeltze, J. Haasz, C. Turkay, E. Hodneland, A. Lundervold, A. J. Lundervold, B. Preim, and H. Hauser. Interactive Visual Analysis of Heterogeneous Cohort-Study Data. *IEEE Computer Graphics and Applications*, 34(5):70–82, 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. D^3 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] 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.

[13] Z. Guo, M. O. Ward, and E. A. Rundensteiner. Model Space Visualization for Multivariate Linear Trend Discovery. In *Proc. of IEEE VAST*, pages 75–82, 2009.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] J. M. Landwehr, D. Pregibon, and A. C. Shoemaker. Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association*, 79(385):61–71, 1984.

[26] A. Maries, N. Mays, M. Hunt, K. F. Wong, W. Layton, R. Boudreau, C. Rosano, and G. E. Marai. GRACE: A Visual Comparison Framework for Integrated Spatial and Non-Spatial Geriatric Data. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):2916–2925, 2013.

[27] 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.

[28] T. Mühlbacher and H. Piringer. A Partition-based Framework for Building and Validating Regression Models. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):1962–1971, 2013.

[29] N. J. Nagelkerke. A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78(3):691–692, 1991.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller. Visual Parameter Space Analysis: A Conceptual Framework. *IEEE Trans. on Visualization and Computer Graphics*, 20(12):2161–2170, 2014.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

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