Regression Cube Analysis of Cohort Study Data

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

Abstract—Epidemiological studies comprise heterogeneous data about a subject group (a *cohort*) to define disease-specific risk factors. These data contain information (*features*) about a subjects' lifestyle, medical conditions and also medical image data. 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 [6]. They can also be used to give recommendations regarding a healthy lifestyle and provide information about wide spread diseases. In the standard workflow, physicians translate observations into hypotheses, which are depicted using epidemiological features and then assessed using regression analyses.

An important epidemiological tool for deriving such features are *Cohort studies*, such as the Study of Health in Pomerania (SHIP) [27]. To reduce any selection bias, subjects are invited at random and without a focus on a specific disease. The acquired features range from social and lifestyle factors to prior or current diseases and medications as well as medical parameters, such as blood pressure and also comprises of non-radiating medical image data. e.g. magnetic resonance imaging (MRI). Medical image data quantified using user-defined landmarks, describing for example shape, volume or diameter of a structure.

To assess the statistical resilience of a hypotheses using regression analyses rarely involves more than three features due to the required subject count. Due to missing overview techniques, possibly interesting correlations lie within the data, but are not made apparent. Explorative analyses and first overview visualizations of the data set as presented by Klemm et al. [15] are not custom tailored to a specific target variable and mostly highlights correlations between variables, which are known to the domain expert (e.g. correlation between body size and spine shape). We incorporate the regression analysis, which is familiar to the domain experts into a overview visualizations, which can either be used for an hypothesis-free analysis or a analysis towards a specific disease or hypotheses. This is achieved by providing template regression formulas, which are applied to all potential variable combinations. Since the notation is familiar to epidemiologists, they can rapidly include their domain knowledge into the analysis process. Difference views between regression formulas allow to assess the influences of individual variables in the process.

Our contributions are:

- A overview visualization technique describing feature interactions using target features.

- Visualization techniques, which incorporate overview visualizations of all regression analyses at once as well as details on demand techniques for detailed investigations of feature relationships.

- freely adjustable regression formulas provide a simple, yet powerful way to adjust the regression analysis to specific hypotheses about the data.

- analysis of confounding variables by providing comparison views between different formula results

- The open and web based approach of the system allows for analysis of any data using the presented method.

2 EPIDEMIOLOGICAL BACKGROUND

This section covers the epidemiological workflow and requirements.

2.1 Epidemiological Workflow

Epidemiology unites experts from different academic disciplines, such as physicians, statisticians and medical computer scientists focusing on biometrics and image segmentation. Their goal is derive disease-specific risk factors by assessing epidemiological features using statistical methods. As described by Thew et al. [24] it is divided into these different steps:

- Clinicians make observations in the daily practice which are translated into hypotheses.

- Epidemiologists compile a list of variables depicting the hypothesis and include confounding variables.

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

- *Paul Klemm, Kai Lawonn, Sylvia Glaßer, Uli Niemann, Bernhard Preim are with Otto-von-Guericke University Magdeburg, Germany. E-mail: {klemm,lawonn,niemann,preim}@ovgu.de*
- *Katrin Hegenscheid, Henry Völzke are with Ernst-Moritz-Arndt University Greifswald, Germany. E-mail: {katrin.hegenscheid,voelzke}@uni-greifswald.de*

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

2.1.1 Data Driven Hypothesis-Generation-based Analysis

Klemm et al. describe an approach for a Interactive Visual Analysis approach for image-centric cohort study data, which connects to the variable listing step [15]. Their methods aim to derive hypotheses through analysis of the data and observations about previously unknown feature correlations. Employing *hypotheses generation* requires overview visualizations of feature correlations, which are not supported by standard statistical processors.

2.2 Epidemiological Data

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

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

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

Image Data. Since the Rotterdam study, many modern cohort studies include medical image data. For ethical reasons, the imaging modalities must not include ionizing radiation. The image quality is often inferior to clinical standard, which is a tradeoff between time and cost [20]. These data are hard to analyze as they require segmentation highlighting the structures of interest. This process is prone to inter and intra-observer variability when carried out manually. Automatic or semi-automatic solutions bypass this problem, but are costly and need to be custom tailored to the structure of interest [25]. Image-derived variables are either dichotomous variables indicating a medical finding by a radiologist or continuous, describing the shape of segmented structures (e.g. volume or diameter).

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

2.3 Regression Analysis

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

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

$$Dependent \sim Independent_1 + Independent_2 + ... + Independent_n \tag{1}$$

The most used regression operators comprise:

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

The type of the target feature restricts the regression type.

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

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

2.4 The Study of Health in Pomerania (SHIP)

The `SHIP`, located in Northern Germany aims to aims to characterize health and disease in the widest range possible [27]. It does not focus on a specific disease, making the data set open for many diseases. Unique for the `SHIP` is the acquisition of medical image data per subject. A second cohort, `SHIP-TREND` was introduced in 2012. Data for both cohorts are examined in a in a 5-year time span. New parameters are added in each iteration, extending the range of investigated diseases. For the last acquisition (`SHIP-2` and `SHIP-TREND-0`), MRI scans are included into the cohort [9, 11].

3 PRIOR AND RELATED WORK

Tukey already stated in 1977 that data if too often only analyzed using confirmatory data analysis [26]. 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. Turkey Metrics Paper

The work of Piringer et al. [19], Angelelli et al. [1] and Zhang et al. [29] is closest to ours regarding different aspects. Akin to our approach, Piringer et al. [19] 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 contours 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.

Zhang et al. [29] 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. [1] is similar to ours as they visualize both image-derived an non image data using cube data structures with focus on comparison and knowledge extraction. Their proposed engine represents multiple heterogenous cohort study data sets as 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 *pearssons 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.

In their survey on interactive information visualizations for health record data, Rind el al. [21] 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.

Shneiderman et al. [23] 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. Futhermore, 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.

Schreck and Keim [22] 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

Statistical Analysis

Visual Analysis of Heterogenous Data. Zhang et al. [28] provide a web-based system for analyzing subject groups with linked views and batch-processing capabilities for categorizing new subject entries into the data set. Their definition of a cohort differs from the understanding of the term in an epidemiological context by denoting every parameter-divided subject group as individual cohort. Due to the short paper length, detail is missing on the data types and their algorithms of identifying similar subjects or whether they employ statistical measures. We employ the idea of adding variables via drag and drop into a canvas area.

Generalized Pairs Plots (GPLOMS) are an information visualization technique comparing heterogenous variables pairwise using a plot-matrix grouped by type [5, 10]. They are useful to gain an overview over numerous variables and their distributions. Histograms, bar charts, scatter plots and heat maps are used to visualize variable combinations with regard to their type. The resulting matrix provides an *overview visualization*, but requires a lot of screen space for many variables (127 in our application scenario). We incorporate the idea of adaptive type-dependent visualizations. Dai et al. [4] explored risk factors by incorporating choropleth maps of epidemiological variables (e.g., mortality rates in a region) with parallel coordinates, bar charts and scatter plots with integrated regression lines. Their findings yielded a *Concept Map*, which linked cancer-related associations via graph edges. While their goal to identify possible risk factors using socio-economic and health data is similar to ours, they focus on iteratively refining defined hypotheses and on geographical data. We employ the use of small multiples for incorporating heterogenous data types for comparability. Chui et al. [3] visualized associations in time-dependent epidemiological data using time-series plots highlighting risk factor differences in age and gender. While the work shows how different visualization techniques provide insight into these data sets, it focuses on the time aspect, which is not present for our data.

Prior Work. We visualized lumbar spine variabilities based on a semi-automatic shape detection algorithm of 490 participants of the SHIP-2 cohort [14]. Hierarchical agglomerative clustering divided the population into shape-related groups. As proof of concept, a relation between the size of the segmented shape and the measured size of the subjects was shown. This work focuses on incorporating these derived data as new variables, enabling to include it into the hypothesis validation and generation process. When applying clustering techniques to the non-image data it was found that k-Prototypes and

`DBSCAN` are appropriate, but are strongly dependent on the chosen variables and distance measures [13]. Niemann et al. [18] presented an interactive data mining tool for the assessment of risk factors of hepatic steatosis, the fatty liver disease. Association rules created by data mining methods can be analyzed interactively with their tool and highlight potentially overlooked variables.

4 REGRESSION CUBE ANALYSIS OF COHORT STUDY DATA

Add Data Preprocessing Section. The basic idea of our *Regression Cube* is to provide a overview visualization of large cohort study data sets towards target variables. Overview visualizations of feature relationships as presented by Klemm et al. [15] are often focused on relationships between the visualized features. Correlation metrics, such as the *Pearson product-moment correlation coefficient* or *Cramér's V* contingency values are incorporated to achieve this goal. In epidemiology, these relationships are also of interest, but rather towards their explanatory power towards the target feature. These target features often indicates the presence of the investigated disease. As described in Section 2.3, regression analyses are the statistical tool of choice for analyzing these relationships. A regression model is based on expert knowledge, there is no unified rule on how to apply them to a given set of features, so they have to be applied with care.

4.1 Cube Description using Regression Formula Notation

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

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

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

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

- $Z \sim X + Y$ calculates all combinations of two features towards all possible target features.

- $Cancer \sim X + Y + BodyWeight$ includes the *BodyWeight* feature into all regression models as feature.

- $Cancer \sim X + Y + Z$ calculates all combinations of three features towards the target.

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

4.2 Target-Variable-dependent Dimension Reduction

Uli, Can you please proofread this? The vast majority of features in an epidemiological data have no or a very low relation towards the target feature. Identifying these features und excluding them from the calculation can reduce the number of dimension significantly. The *Correlation based Feature Selection* (CFS) algorithm is very popular in data mining for achieving this task [8]. It uses information entropy to select the features which have the most explanatory power towards

the target feature. At the same time it tries to reduce the correlation between the independent features. When for example the *body weight* has a strong explanatory power towards the target it is likely that *BMI* or *waist circumference* behave similar towards it. They, however, correlate strongly with each other. The CFS algorithm would then select the feature which has the largest explanatory power and discards the other dimensions.

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

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

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

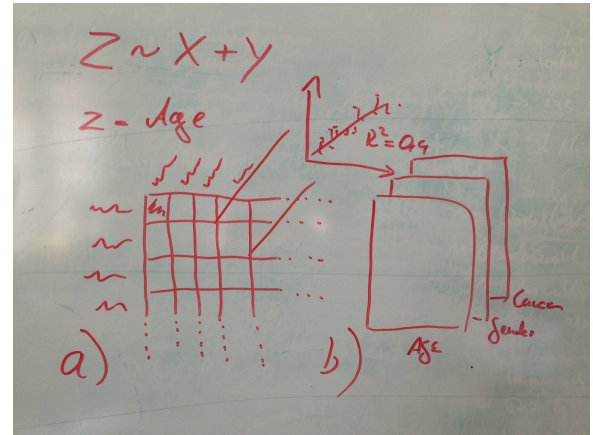### 4.3 Abstracting Regression Results using R^2

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.

The goal of a overview visualization is to provide a comprehensive view on the data, which is easy to understand. As described in [15], correlation values scaled between 0 (no correlation) and 1 (perfect correlation) can be encoded using color on a mosaic plot. Regression models are more complex, having many associated describing metrics. For the *Regression Cube* analysis we are interested in the goodness of fit of the resulting model, which allows to infer about the predictive quality of the independent features included in the model. As described in Section 2.3, the R^2 value is the metric allowing for this kind of assessment.

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

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

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

4.4 Analysis Workflow

I'm not satisfied with the structure in this subsection, please proofread it with focus in the train of thought! The analysis workflow follows
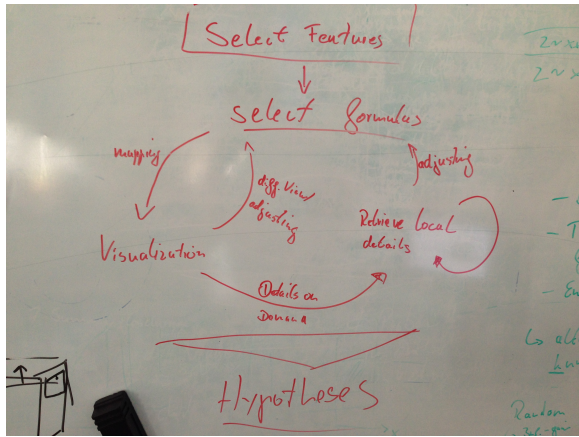

Fig. 2. Workflow of the analysis using *Regression Cubes*. The analysis starts with a set of features given by the data set. Then, the user may specify a formula regarding to specify a hypothesis, or use a predefined formula to start a explorative analysis of the data set. The Regression Cube is then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

the Visual Analytics (VA) Mantra of Keim et al. [12]:

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

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

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

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

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

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. TODO: divide using analysis approaches, e.g. Hypothesis-based & Hypothesis-free with follow up on the formulas used?

Hypothesis-Free Analysis

Hypothesis-Based Analysis

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

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

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

$Cancer \sim X + Y + Z$ subtracted with results from $Cancer \sim Age$ excludes the confounding effect age has towards the target $Cancer$ feature.
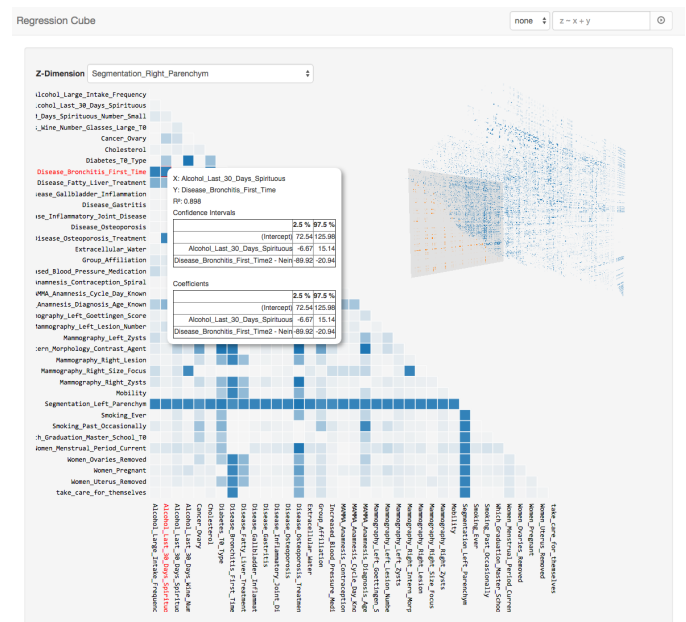
5 System Design

Fig. 3. FIGURE NOT FINAL! 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 a

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 also inherently designed with client-server architecture in mind, making it easy to outsource computation heavy tasks on server-clusters and transferring results to the client device. The design space of 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 ui 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 values (CSV) data set.

- 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.

- The *cube visualization* consisting of the 2D-mosaic view as well as a 3D-representation of the whole cube.

File-Upload and Classification. Popular analytics tools, such as WEKA [7] owe their popularity to their support of open file types. To allow other users even outside the epidemiological application domain access to our tool, we choose to do this using 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 using no character at all, a whitespace, or an empty quotation mark encapsulated string.

Data Security. Security issues are raised by uploading data into a online service, such as our prototype. The use of epidemiological data is preceded by a detailed description on the analysis purpose and has to be approved by ethics committees. Preventive steps have to be taken to restrict access to unauthorized subjects. We decided against a user account system because it reduces the ease of use largely without having a direct advantage for the user. Instead 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 the closing a session.

Formula Editor. 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 (sometimes hard to memorizable) variable names.

Formula Validation and Calculation. The formula is checked for validity directly on input, indicating invalid formulas with a red halo around the text input, which turns to green for valid input. This prevents processing errors on the statistical processor backend. 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 backend 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. If the user selects a reference formula, the absolute difference of R^2 values between the cube described by the reference and the currently active formula are calculated and shown.

5.2 Regression Cube Visualization.

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 mosaic plot matrix diagonal. 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.

3D-Prism as Data Mini-Map. The 3D-cube representation acts 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 occlusion and interaction problems, which often do not balance out the advantages of using the third dimension for visual mapping. We aim to minimize these advantages 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 of the position of the individual features. The cube is always oriented according to the 2D-representation, allowing for a 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. 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 current 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 a overlapping-free view on the data points, which makes it easier to identify the current slice.

2D Mosaic Plot Slice Visualization. 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.

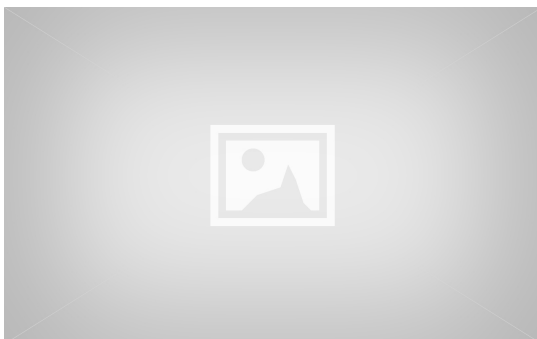Fig. 4. Screenshots of Cubes from the Evaluation

6 IMPLEMENTATION

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

Relying on web technologies paid off in previous collaboration with our clinical exerts, since it 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.js[1] 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 Bootstrap[2], 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 [2]. It provides fast and easy methods for binding data to graphical elements. The 3D plot is created using the WebGL-based Threejs[3] library. We experimented with different ways for achieving the cube representation, including volume rendering, cube primitives for each data point and shader-based solutions. Proper open source volume rendering methods are not available for the web, the effort of writing one from scratch was not worth the counter value. 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.

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

The statistical processors yield the second structure. They rely on the statistical programming language R[6]. It is widely adopted in the statistical analysis community, yielding a rich support of fast state-of-the-art statistics algorithms as well newly published methods. OpenCPU is a R package and provides a API for accessing it via HTTP calls. This way, any computer, which runs R can be turned into a statistical processor for our project. The backend 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 backend 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.

---

[1] Open Source; Maintained by Google, angularjs.org
[2] Open Source; Maintained by Twitter, getbootstrap.com
[3] Open Source; Originally developed by R. Cabello, threejs.org
[4] Open Source; Maintained by Joyent Inc, nodejs.org
[5] Owned by Salesforce.com, heroku.com
[6] Open Source; r-project.org
[7] R-based back-end:
github.com/paulklemm/regression-cube-r-package
[8] Front-End and Nodejs Webserver:
github.com/paulklemm/regression-cube-prototype

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 a 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 a few features depicting a hypotheses 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 a system to discover knowledge discover knowledge is hard to measure. Lam et al. [16] propose for this purpose the *Visual Data Analysis and Reasoning* (*VDAR*) technique, which is focused on the characterization of a systems ability to generate hypotheses and explore the data in order to extract information. *VDAR* can be carried out using case studies using thinking-aloud techniques to comprehend the reasoning and thought process of the user. We conducted a study using 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.

Setup and Procedure. One major advantage of focusing on web based research was the possibility to bridge the geographical gap towards the epidemiological domain experts. We conducted a web based research by using a online-meeting software, which features voice chat as well as screen sharing. Starting a 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.

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% [17].

The data comprises 1.186 subjects (368 from SHIP-2, 818 from SHIP-TREND-0 cohort). It contains 230 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 [11] 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 a 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

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

7.3.1 Data Preprocessing

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

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

7.6 Further Feedback and Lessons Learned

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

8 SUMMARY AND CONCLUSION

We presented a technique for knowledge discovery in cohort study data sets with user-defined target features. *Regression Cube* are based on eponymous regression models, which allow to model domain knowledge about the data (e.g. influencing features on a disease) using formulas. Alternatively, *Regression Cubes* can be calculated using default formulas to assess the explanatory power towards each feature in the data set. These two approaches allow both for a *hypothesis-free* and *hypothesis-based* explorative analysis. Correlation based feature selection reduces the amount of calculations using the target feature by focusing on the important regression models. The prototype described in this work was developed using state of the art web technologies, free services and open source libraries. All code associated to this project is also open source. Two case studies revealed that the presented approach is able to reproduce knowledge extracted using decision tree based data mining methods on a hepatic steatosis data set as well as producing new hypotheses by deriving insight into influencing factors on breast fat tissue in a explorative analysis.

As future work, we want to introduce more regression models to the data set, which model different correlation types. We also want to extend the cube to time-dependent data by expanding the difference-cube approach.

The method was very well received with our project partners, allowing them for the first time to retrieve a overview visualization of features, which does not highlights their pairwise correlations, but rather their explanatory power towards a target feature. The observed relationship are now subject of detailed statistical analyses. Our provided system not restricted to the epidemiological domain, using the provided instance, everybody can upload their own data or even set up their own *Regression Cube* server cluster. 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.

REFERENCES

[1] P. Angelelli, S. Oeltze, C. Turkay, J. Haasz, E. Hodneland, A. Lundervold, B. Preim, and H. Hauser. Interactive visual analysis of heterogeneous cohort study data. *IEEE Computer Graphics and Applications*, 2014. in print.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] P. Klemm, L. Frauenstein, D. Perlich, K. Hegenscheid, H. Völzke, and B. Preim. Clustering Socio-demographic and Medical Attribute Data in Cohort Studies. In *Bildverarbeitung für die Medizin (BVM)*, pages 180–185, 2014.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] U. Niemann, H. Völzke, J.-P. Kühn, and M. Spiliopoulou. Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis. *Expert Systems with Applications*, 2014.

[19] 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.

[20] 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.

[21] 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.

[22] T. Schreck and D. Keim. Visual analysis of social media data. *Computer*, 46(5):68–75, 2013.

[23] B. Shneiderman, C. Plaisant, and B. W. Hesse. Improving healthcare with interactive visualization. *Computer*, (5):58–66, 2013.

[24] 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.

[25] 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.

[26] J. W. Tukey. Exploratory data analysis. *Reading, Ma*, 231:32, 1977.

[27] 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.

[28] Z. Zhang, D. Gotz, and A. Perer. Interactive visual patient cohort analysis. In *Proc. of IEEE VisWeek Workshop on Visual Analytics in Health Care*, 2012.

[29] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Information Visualization*, page 1473871614526077, 2014.