Regression Cube Analysis of Cohort Study Data

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

Abstract—Problem. Epidemiological studies comprise of heterogenous 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 images of the whole body. These features are analyzed using statistical regression analyses towards features indicating a disease (the *target feature*). These analyses usually include two to five features (*independent features*) towards the disease of interest. New Solution. We propose a new analyses approach of epidemiological data sets by incorporating all dimensions in a exhaustive regression-based analyses. It takes all possible combinations of *independent features* towards a *target feature* into account and provides a visualization, which allows for insights into the data by showing relationships possibly unknown to the domain expert. A 3D-visualization of all possible combinations of two to three target variables acts as overview over the whole data set. We call it the *Regression Cube*. Slicing through the regression cube allows for the detailed analysis of features towards the target disease. The regression formula can be adjusted by the user to describe disease-specific hypotheses. Influences of features can be assessed using a difference view, comparing different calculation results. Validation. We applied our *Regression Cube* method to a hepatic steatosis data set to reproduce results from a data-mining driven analysis. We conducted a qualitative analysis with three domain experts on a breast fat data set to derive insights into the correlation between breast lesions and non-image variables. Results. TODO. Implications. Epidemiological data sets can for the first time be visually overviewed using a regression-based analysis of all features towards a disease.

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

1 Introduction

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

An important epidemiological tool for deriving such features are *Cohort studies*, such as the Study of Health in Pomerania (SHIP) [10]. 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. [**?**] 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:

- An 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

2.2 Epidemiological Data

2.3 The Study of Health in Pomerania (SHIP)

After the pioneering Rotterdam study (started in 1990), several MR imaging study initiatives were initiated. They slightly differ in clinical focus, acquired data and epidemiological research questions. Starting in 1997 with a cohort of 4,308 subjects, the SHIP, located in Northern Germany, aims to characterize health and disease in the widest range possible [10]. Data are collected without focus on a group of diseases. This allows to query the data regarding many diseases and conditions. Subjects were examined in a 5-year time span, continuously adding new parameters including MRI scans in the last iteration [5].

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

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

3 Prior and Related Work

Visual Analysis of Heterogenous Data. Zhang et al. [11] provide a web-based system for analyzing subject groups with linked views and batch-processing capabilities for categorizing new subject entries into the data set. Their definition of a cohort differs from the understanding of the term in an epidemiological context by denoting every parameter-divided subject group as individual cohort. Due to the short paper length, detail is missing on the data types and their algorithms of identifying similar subjects or whether they employ statistical measures. We employ the idea of adding variables via drag and drop into a canvas area.

Generalized Pairs Plots (GPLOMS) are an information visualization technique comparing heterogenous variables pairwise using a plot-matrix grouped by type [3, 6]. They are useful to gain an overview over numerous variables and their distributions. Histograms, bar charts, scatter plots and heat maps are used to visualize variable combinations with regard to their type. The resulting matrix provides an *overview visualization*, but requires a lot of screen space for many variables (127 in our application scenario). We incorporate the idea of adaptive type-dependent visualizations. Dai et al. [2] explored risk factors by incorporating choropleth maps of epidemiological variables (e.g., mortality rates in a region) with parallel coordinates, bar charts and scatter plots with integrated regression lines. Their findings yielded a *Concept Map*, which linked cancer-related associations via graph edges. While their goal to identify possible risk factors using socio-economic and health data is similar to ours, they focus on iteratively refining defined hypotheses and on geographical data. We employ the use of small multiples for incorporating heterogenous data types for comparability. Chui et al. [1] visualized associations in time-dependent epidemiological data using time-series plots highlighting risk factor differences in age and gender. While the work shows how different visualization techniques provide insight into these data sets, it focuses on the time aspect, which is not present for our data.

Prior Work. We visualized lumbar spine variabilities based on a semi-automatic shape detection algorithm of 490 participants of the SHIP-2 cohort [8]. 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 [7]. Niemann et al. [9] 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 Cubes

Target Group: Physicians/Epidemiologists with basic statistical background to understand regression formulas and Interaction

4.1 Analysis Workflow

Different Formulas work as different steps in the analysis

5 System Design and Implementation

5.1 Design and Visualization Techniques

System Layout.

5.2 Implementation

Use VAST'14 as Guideline

6 Application

6.1 The Breast Fat Data Set

6.1.1 Data Preprocessing

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

6.2 Participants, Setup and Procedure

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

Procedure. Analyses here ...

6.3 Further Feedback and Lessons Learned

7 Summary and Conclusion

Acknowledgments

Omitted due to blind review

REFERENCES

[1] K. K. Chui, J. B. Wenger, S. A. Cohen, and E. N. Naumova. Visual analytics for epidemiologists: understanding the interactions between age, time, and disease with multi-panel graphs. *PloS one*, 6(2), 2011.

[2] X. Dai and M. Gahegan. Visualization based approach for exploration of health data and risk factors. In *Proc. of the International Conference on GeoComputation. University of Michigan, USA*, volume 31, 2005.

[3] J. W. Emerson, W. A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H. Wickham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91, 2013.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.