INTERACTIVE VISUAL ANALYSIS OF POPULATION STUDY DATA

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von M. SC. PAUL KLEMM

geb. am 12.05.1986, in Schmölln

Gutachterinnen/Gutachter
Prof. Dr. Bernhard Preim
Prof. Dr. Daniel Weiskopf
Prof. Dr. Silvia Miksch

Magdeburg, den 09. Dezember 2015

Für Anja.

DANKSAGUNG

An erster Stelle möchte ich mich herzlich bei meinem Doktorvater Bernhard Preim bedanken. Er war es, der in mir das Potential gesehen hat, die hier dokumentierte Aufgabe zu meistern. Er hat mir die Möglichkeit gegeben, das zu beweisen. Diesen Prozess hat er mit viel Hingabe und beispielloser Fürsorge begleitet. Bernhard hat immer ein offenes Ohr für Probleme und weist den Weg. Gleichzeitig geht er sicher, dass man diesen auch selber geht und sagt einem Unterstützung zu, wenn man holpriges Gelände betritt. Ich wusste stets, dass Bernhard mir den Rücken frei hält! Ich betrachte mich als äußerst glücklich und geehrt, einen Doktorvater zu haben, der sich mit solcher Hingabe für seine Mitarbeiter einsetzt.

Bernhard, die Zusammenarbeit mit dir hat mich sehr zum Positiven geprägt. Es war eine tolle Zeit!

Dann gibt es so viele, denen mein Dank gilt. Ich bin sehr geehrt, mit Weltklasse-Wissenschaftlern gemeinsam Arbeiten publiziert haben zu dürfen.

Besonders danken möchte ich Katrin Hegenscheid und Henry Völzke, die mir mit großer Geduld und Aufgeschlossenheit geholfen haben, das mir zunächst unbekannte Feld der Epidemiologie zu erschließen.

Großer Dank gilt Silvia Miksch und Daniel Weiskopf für das Schreiben der Gutachten für die vorliegende Arbeit.

Ich möchte der gesamten Arbeitsgruppe Visualisierung für die tolle Arbeitsatmosphäre und die gemeinsame Zeit danken. Mein besonderer Dank hier gilt Sylvia Glaßer und Kai Lawonn. Danke, dass ihr mir stets mit Rat und Tat zur Seite standet. Ihr habt hieran einen großen Anteil. Wir waren ein tolles Team!

Ich möchte den Mitarbeitern des Institutes für Simulation und Graphik für die wirklich außergewöhnliche Arbeit danken, die mein Leben so viel leichter gemacht hat. Petra Schumann, danke dass du unermüdlich Ordnung in mein Reisegelderchaos gebracht hast. Stefanie Quade, danke für das geduldige und gründliche Korrigieren meiner Texte. Das mit der Ein- und Mehrzahl habe ich jetzt hoffentlich verstanden. Thomas Rosenburg und Heiko Dorwarth, danke für das Umsetzen so vieler kurioser Ideen und dass ihr immer im Zweifelsfall einen Server für mich herbei gezaubert habt.

Petra Specht, es tut immer gut zu wissen, dass die gute Seele des Institutes auch Berge verrückt, wenn es für die Mitarbeiter notwendig ist. Danke!

Danke an meine Familie für die uneingeschränkte Liebe und Unterstützung über all die Jahre. Danke, dass ihr an mich glaubt und zu mir steht. Ich bin sehr glücklich, eine so tolle Familie zu haben. Je älter ich werde, desto klarer wird mir, wie wenigen Menschen ein solches Glück zuteil ist. Ebenso großer Dank gilt den Fincken, die mich so herzlich in ihrer Mitte aufgenommen haben.

Auch wenn du auf keiner der Arbeiten als Koautor stehst, kennst du deinen Anteil daran. Deswegen ist dir diese Arbeit gewidmet. Du wirst bestimmt sagen, dass es nicht stimmt, aber ohne dich wäre das hier nicht möglich. Danke!

EHRENERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; verwendete fremde und eigene Quellen sind als solche kentlich gemacht. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert,
- fremde Forschungsergebnisse verzerrt wiedergegeben.

Mir ist bekannt, dass Verstöße gegen das Urheberrecht Unterlassungs- und Schadensersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen kann. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, den 09. Dezember 2015

Paul Klemm

ABSTRACT

Epidemiological population studies impose information about a set of subjects to characterize disease-specific risk factors. Population studies comprise heterogeneous variables (*features*) describing the medical condition as well as demographic and lifestyle factors. The data are analyzed using a priori defined hypotheses to find statistically significant correlations between features (*associations*). Modern population studies incorporate medical image data. The statistically driven epidemiological workflow only allows to determine *associations* between image-derived metrics, such as distances extracted from landmarks of the segmentation model.

This thesis proposes visual analysis techniques for both *explorative* and *confirmative* analyses of population study data. Methods for analyzing image-centric population study data with focus on assessing influences of organ shape are proposed. To account for epidemiological key requirements such as reproducibility and statistical resilience of results, the epidemiological workflow is analyzed and divided into different steps. Based on this analysis, an Interactive Visual Analysis (*IVA*) approach is proposed that enables epidemiologists to examine both image-based as well as non-image data, e.g., sociodemographic features and attributes derived from the image data. The new framework enables hypothesis validation and generation by incorporating human pattern recognition as well as data mining methods. Using all reliable information from the image segmentation linked to non-image features aims to unveil *associations* by applying an iterative analysis approach.

Additionally, methods for the explorative analysis of large-scale population study data to derive new hypotheses about the data are proposed. For medical image data, this is achieved by creating shape-based clusters of subjects, which can then be related to other non-image features. For explorative analyses of non-image features, novel techniques to derive overview visualizations for data sets with many features are proposed, which can be customized by including expert knowledge.

ZUSAMMENFASSUNG

Epidemiologische Bevölkerungsstudien erheben Informationen über eine Menge von Probanden, um krankheitsspezifische Risikofaktoren zu beurteilen. Bevölkerungsstudien beinhalten heterogene Variablen (*Features*), welche den medizinischen Zustand sowie demographische Faktoren und Informationen zu den Lebensumständen eines Probanden erfassen. Die Daten werden durch a priori festgelegte Hypothesen mit dem Ziel analysiert, statistisch signifikante Korrelationen zwischen Features (*Assoziationen*) zu finden. Moderne Bevölkerungsstudien beinhalten ebenfalls häufig medizinische Bilddaten. Der durch den Einsatz von statistischen Mitteln gekennzeichnete epidemiologische Arbeitsablauf erlaubt es lediglich, Assoziationen zwischen abgeleiteten Bildmetriken, wie z.B. Distanzen zwischen bestimmten Landmarken eines Segmentierungsmodells, zu bestimmen.

Diese Arbeit stellt visuelle Analysemethoden sowohl für die explorative als auch die hypothesengesteuerte konfirmative Analyse von Bevölkerungsstudiendaten vor. Sie schlägt Methoden vor, mit denen bildbezogene Bevölkerungsstudien mit Bezug auf die Form bestimmter Organe und Gewebetypen untersucht werden können. Die für die Epidemiologen wichtigen Anforderungen, wie das Erlangen von reproduzierbaren und statistisch belastbaren Ergebnissen, werden durch das Identifizieren einzelner Arbeitsschritte der epidemiologischen Analyse definiert.

Basierend auf diesen Anforderungen wird ein interaktiver visueller Analyseansatz (*IVA*) vorgestellt, der es Epidemiologen ermöglicht, sowohl bildbasierte- als auch Nichtbilddaten, wie soziodemografische Features, gemeinsam zu analysieren. Das entstehende System erlaubt sowohl die Validierung als auch die Generierung von Hypothesen, indem es die menschliche Mustererkennung sowie Methoden der automatisierten Datenauswertung miteinander verknüpft. Das Verbinden von verlässlichen Aspekten der Bild-Segmentierungsmasken mit Nichtbild-Features erlaubt es, neue Assoziationen in einem iterativen Analyseansatz herauszuarbeiten.

Zusätzlich werden Methoden für die explorative Analyse von komplexen Bevölkerungsstudiendaten vorgestellt, um neue Hypothesen zu generieren. Dies wird für medizinische Bilddaten durch das Bilden von formbasierten Gruppen erreicht, die dann in Bezug mit Nichtbild-Features gebracht werden können. Die auf Nichtbild-Features basierende explorative Analyse wird durch neue Überblicksvisualisierungstechniken ermög-licht, die durch das Einbinden von Expertenwissen an Problemstellungen angepasst werden können.

CONTENTS

Part I

PRELIMINARIES

1 INTRODUCTION

The analysis of population study data follows a strict hypothesis-driven pipeline, which was employed over many years using reliable statistical methods. Hypotheses are formulated based on research results or clinical observations, translated into hypotheses, which are then statistically evaluated. Population studies increase in size over the years to cover a wide range of diseases and hypotheses. This wide scope of features provides new opportunities, leaving the standard hypothesis-driven pipeline. By employing visual analysis techniques, the data complexity can be used as advantage to conduct explorative analyses to *generate new hypotheses*. Confirmative analyses may be conducted with a wider range for features in mind, to assess their influence w.r.t. a specific disease or condition. Both approaches require new analysis techniques combining proper overview visualizations showing results of analytics algorithms, which derive potentially overlooked relationships in the data.

Analyzing medical image data as part of large-scale population studies adds a new dimension of complexity to the problem. The data has to be annotated and quantified in order to analyze it with standard statistical methods. Here, visual analysis can include shape variance visualizations to allow for comparisons between subject groups without removing information by summarizing it into abstracted metrics, such as volume or diameter. Additionally, explorative analysis of medical image data can be enabled by clustering the annotated models. This provides experts with completely new ways of analyzing their data. Diseases can be related to shape-specific differences between subjects. Subjects of shape-based clusters may share interesting similarities compared to the overall population in the study. This thesis provides an overview of the applicability of existing methods w.r.t. population study analysis. The main goal is to provide explorative and confirmative analysis techniques for data sets with or without medical image data.

The work presented in this thesis was supported by the priority program "*Scalable Visual Analytics*" (DFG SPP 1335) of the German Research Council. The project "*Visual Analytics in Public Health*" aims to provide flexible analysis methods for the data derived in the "*Study of Health in Pomerania*" (SHIP) [270]. The project also involves the Image Processing group of Klaus Tönnies at the University of Magdeburg as well as the Institute of Community Medicine at the University of Greifswald. The analyses methods are required to scale between different research questions of different subsets of the same data pool, which may include heterogenous data types. We try to answer the question how visualizations change as the number of research questions increases. Additionally, do the visualizations change as the number of investigated subjects increases? Special emphasis was put to include medical image data into a visual analysis workflow. Medical data is acquired in large-scale population studies, but they are still hard to assess as part of explorative or confirmative analysis sessions. The development of the methods in this thesis was carried out in a thorough cooperation with Henry Völzke, leader of the SHIP and specialist in internal medicine, as well as PD Dr. Katrin Hegenscheid, radiologist and responsible for the SHIP Magnetic Resonance Imaging (MRI) data acquisition. Both experts are affiliated with the Ernst Moritz Arndt University Greifswald.

This thesis tackles the following challenges:

- Derive hypotheses about complex diseases using a *explorative* data-driven Visual Analysis approach on large-scale population study data.
- Combine medical image data and non-image information to validate and generate hypotheses based on shape variance of organs and tissues.
- Establish a Visual Analysis workflow for epidemiological data, which combines medical visualization, information visualization and data mining methods.

Therefore, this thesis is organized as follows:

- Chapter 2 covers basics in epidemiology to provide the necessary background for the methods developed in the later chapters. It is intended as a brief introduction into the field, which provides the necessary vocabulary and an overview of the applied workflows and analysis pipelines as well as different study types.
- Chapter 3 provides the current state-of-the-art in visualization of statistical results, as well as Visual Analytics and Interactive Visual Analysis and provides possible application areas in population study data analyses.
- Part ii starts with an extension to the population study data analysis pipeline using Visual Analytics and Interactive Visual Analysis methods. The purpose is to categorize the methods proposed in this thesis. The categories are either support of explorative or confirmative analyses and the use of image data.
- Chapter 4 provides methods for joint analyses of medical image data with non-image features, such as lifestyle factors or results of medical examinations. It covers aspects of deriving data structures suitable for shape comparisons as well as information visualization of non-image features with augmented image data. The visualizations presented in this section usually comprise *physical views* representing spatial data. Explorative analysis is enabled by applying clustering methods on both image and non-image data.
- Chapter 5 focuses on analysis methods without *spatial views*. It provides Visual Analytics tools for providing overview visualizations, which include domain knowledge by encoding it using regression models. It also proposes the Decision Tree Quality Plot, which assesses the predictive quality of a set of features w.r.t. the whole data set.
- Part iii and Chapter 6 conclude the thesis by giving a summary of the contributions and discussing and ranking future work in the field.

Most sections in Part ii are based on conference and journal publications. Therefore, each section will contain a detailed description of the contributions per author to clarify the own impact and contributions.

2 EPIDEMIOLOGICAL BACKGROUND

"*Epidemiologists are detectives who research the causes and consequences of illness and disease.*" – Career description on innerbody.com

Epidemiology assesses the spread, causes and effects of health-related conditions. It aims to characterize health and disease by determining risk factors. These risk factors can then be used to determine optimal treatment, develop preventive medical checkups and to give recommendations for a healthy lifestyle. They can also be used to extract high-risk groups for diseases. Features defining these groups act as diagnostic markers.

This chapter provides an overview of the most important terms in epidemiology as well as a summary of the involved experts, their workflow and the data basis.

2.1 IMPORTANT TERMS

Since the focus of epidemiology is on characterizing health and disease conditions, the field developed metrics to assess these terms.

Prevalence and **incidence** are two metrics depicting how often a certain disease (or *clinical events* in a wider definition) occurs in a specified population. *Clinical Events* include diagnoses of severe diseases, for example cerebral strokes or heart diseases. The *prevalence* denotes the percentage of people suffering from a disease at a given point in time. More precisely, *point prevalence* indicates prevalence at one time point and *period prevalence* states prevalence over a period of time. The latter is harder to interpret. Hence, prevalence is often synonymous for *point prevalence*. It is depicted as ratio between a healthy and diseased population size in percentage. For example, the prevalence of a disease affecting 50 people out of a population of 1300 would be ≈ 3.85.

The *incidence* represents how many people *newly* get diagnosed with a disease in a certain interval, usually one year. High prevalence is usually associated with high economic costs. Population-based studies analyze diseases with a high prevalence, such as widespread diseases, i.e., diabetes or back pain. A rare disease, such as amyotrophic lateral sclerosis, may have a prevalence of 5 from 100,000 [296]. Thus, even in a large population-based study *probably no individual suffers from this disease*. Dividing the incidence by a time frame and the number of subjects in the group yields the **absolute risk** (also called **incidence rate**). It is used to determine the risk per subject group of developing a disease and makes their risks comparable. As an example, a study related to risk for cardiac diseases may investigate angina pectoris, myocardial infarction, atrial fibrillation depending on attributes, such as age and gender. According to Preim et al. [296] "**relative risk** (RR) characterizes the increased risk of an individual being exposed to a certain risk factor, e.g., smoking, excessive weight, or alcohol abuse. It is based on a comparison with a control group not exposed to that risk factor. A value of $RR < 1$ represents a factor that protects, e.g., moderate physical activity." Insightful observations are often combined effects of several parameters. A certain factor may be protective for some people and is involved with an increased risk for others. The combined risk may be significantly smaller or larger than could be expected from individual factors. Another important metric is the **odds ratio** (OR) [23], which is a measure of effect size, depicting

Table 1: Fictional example of the relationship between smoking status and having a heart attack. The **Relative Risk** is calculated as: $RR = \frac{140/(140+1780)}{77/(77+7620)} \approx 7.28$. This indicates a 7.28 times higher chance of smoking subjects to suffer from a heart attack. The **Odds Ratio** is denoted as: $OR = \frac{140*7620}{77*1780} \approx 7.7$. This indicates that subjects who smoke are 7.7 times more likely to have a heart attack than non smoking subjects [23].

	smoking	not smoking
heart attack	140	77
no heart attack	1780	7620

the association strength between two binary features. A fictional example of relative risks and odds ratios can be seen in Table 1.

Statistical correlations are prone to **confounding**, meaning that the association of two epidemiological variables is influenced by a third variable, which needs to be isolated (see Fig. 1). As stated in the VAST'14 publication [293], "a famous example is the association between shoe size and mortality, where it can be observed that people with larger shoe size have a smaller life expectation. The shoe size is actually associated with gender, where women have smaller feet and also a longer life expectation. *Gender* therefore acts as confounder for this analysis." *Age* is included as

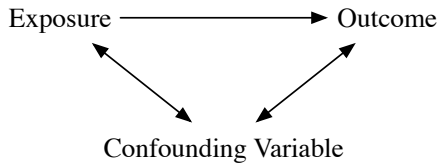


Figure 1: Exposure (e.g., smoking) as well as outcome (e.g., heart attack) are associated with a confounder (e.g., age).

confounder in almost any epidemiological analysis, since most diseases are more likely for older people. It also influences the general body condition and thereby almost all features acquired through cohort studies. Confounders have to be selected by epidemiologists specific to the investigated condition [74]. One possibility to accomplish this are directed acyclic graphs, which allow for displaying causal structures in an epidemiological data set [228, 241]. Each variable is represented as node in the graph, influences are encoded via edges. A path is denoted as causal, when it contains only directed edges from the exposition feature to the outcome. The possibilities of handling confounders are twofold:

i The confounding effect is already considered in the design phase of the study, which includes *randomization* of the subjects (e.g., creating comparable groups w.r.t. known and unknown confounders), or *restriction* of specific subgroups (e.g., only including males or females to eliminate gender confounder). Note that *restriction* is very limited in large-scale population studies, which are queried towards many different diseases.

ii The second method includes controlling at the statistical analysis stage, e.g., by creating groups according to the confounding feature.

Details are described in Section 2.6.

2.2 STUDY TYPES & DESIGN

Epidemiological data are derived from various study types. The study type is chosen depending on the underlying question. Four basic study designs are defined in epidemiology [197]:

- *Incidence studies* aim to acquire exposures and outcomes for all population members. These studies are called **cohort studies** or **follow-up studies**. They include data acquisition at several time points.
- *Incidence case-control studies* aim to study the causes of a specific disease over a period of time. Information about exposed and not exposed subjects are derived and shown using odds ratios. In comparison to the other study types, case-control studies are less expensive due to their focus.
- *Prevalence studies* aim to analyze the prevalence of a disease at a specific point in time. Causations are harder to assess due to the missing time frame. Differences in disease progressions are difficult to determine.
- *Prevalence case-control studies* are, similar to incidence case-control studies, less complex than standard prevalence studies, due to focusing on a specific outcome.

This work is focused on analyzing single time points of cohort studies. A group of subjects (the cohort) is observed over time, usually in fixed intervals. The subject number decreases between each time step as subjects decease or stay away from the next acquisition dates. These data impose the most potential for interactive visual analysis techniques, since they have a vast variety of features and are acquired with a broad spectrum of diseases (outcomes) in mind. Deriving data-driven insight into epidemiological questions is arguably most promising for cohort study data due to the wide scope of acquired data.

Designing epidemiological studies requires professional planing and a thorough understanding of the investigated disease or condition. Choosing the proper study type is, as described above, the crucial first step. Further design strategies are then motivated by excluding different error types:

- *Random errors* can occur when the biological diversity of a population is not representative. For example, conducting a study in a rural area may lead to populations with above-average genetic similarities due to high relationship degrees. Random errors can also occur due to measurement errors in the data collection. Random errors are *not* systematic.
- *Sample size* errors occur when the number of subjects are not sufficient to derive statistical significant conclusions [158]. The sample size depends on the prevalence of the investigated disease in large-scale population studies. Case-control studies require fewer subjects, since one group is defined by diseased patients.
- *Systematic errors* (**bias**) are systematic deviations from the truth. Causes comprise *distortion by selection*, where selected subjects differ from non-selected (e.g., subjects participating in a study because they want to have a check up due to malaise). Other causes comprise *information bias*, where measurements (e.g., determination of a disease) are inaccurate. This can be caused by different monitoring stations (e.g., multiple imaging stations with different calibrations).

2.3 EPIDEMIOLOGICAL EXPERTS

Epidemiology is an umbrella for an interdisciplinary collaboration of experts. Physicians are involved in formulating epidemiological questions as results of day-to-day practice or scientific research. Physicians focusing on epidemiology usually have an above-average understanding of statistics. Most physicians conduct epidemiological research next to their activity as medical doctors. This is accompanied by a very sparse time window for this occupation, requiring efficient yet flexible methods to accomplish the work.

Designing epidemiological studies also includes the collaboration of physicians with different specializations, depending on the investigated conditions. Incorporating genetic information requires consultation with geneticists, deriving medical image data includes radiologists, and so forth.

Statisticians are required at many stages of epidemiological analyses. They are essential for designing studies to be as bias-free as possible. Data storage and analyses are carried out by statisticians. They are responsible for statistically validating epidemiological hypotheses to derive final conclusions whether the data supports a certain assumption or not.

These analyses, however, require statistically evaluable epidemiological variables. Most measurements already yield numerical or categorical variables, which can be evaluated in such a way, e.g., gender, BMI or education level. Others need to be extracted using a post-processing step based on the measured data, such as analysis of medical image data. These data can be extracted by radiologists, but this is expensive and prone to inter- and intra-observer variability. Another approach is consulting computer scientists for writing algorithms to extract the information of interest.

Successful epidemiological research was always driven by an interdisciplinary collaboration of experts to derive insight into the difference between health and disease. The collaboration is guided by a strict analysis workflow, which is described in the following section.

2.4 EPIDEMIOLOGICAL ANALYSIS WORKFLOW

In this section, the process of designing and conducting an epidemiological study is neglected. The focus of this work lies on extending the analysis workflow for epidemiological data.

The workflow herein described was presented in the VAST'14 paper [293]. Epidemiologists follow a workflow mainly driven by statistic tools to validate hypotheses about disease-specific risk factors. Following Thew et al. [254], the workflow can be characterized as follows:

1. A hypothesis is derived from observations made by physicians in their daily routine.

2. A set of features depicting conditions affected by the hypothesis is compiled accordingly. Often, for specific diseases and associated hypotheses, dedicated epidemiological studies are designed. Large-scale population studies can be used when the prevalence of the investigated condition is high enough in the population, since subjects are invited without the focus on a specific disease.

3. Confounding features are identified and taken into account (for example using stratification). This is a complex step involving both medical expert knowledge as well as statistical analysis. Confounders have to be compiled specific to the investigated disease, since each condition has individual influencing factors, which, if not taken into account, will introduce a bias into the analysis, rendering the analysis inaccurate.

Figure 2: The classical epidemiological analysis starts with observations of clinicians in their day-to-day practice. These observations are translated into hypotheses, which are then depicted using a feature list derived from a epidemiological study. Statistical analyses determine whether the features support the hypothesis or not.

4. Statistical methods, such as regression analyses, assess the association of selected features with the investigated disease. This last step is usually carried out by statisticians and is the final step of the analysis. The analysis can also be inconclusive, e.g., by having a too small sample size.

The workflow is depicted in Figure 2. Reproducibility of results is an epidemiological key requirement. It is difficult to achieve, since many physicians are involved when thousands of test persons are examined and interviewed. Thus, both intra- and inter-observer variability needs to be low for all aspects of a population study examination. Longitudinal studies require the acquired attributes to be comparable for evaluation. If the data acquisition process changes, an information bias is introduced to the data, hampering inference in and between acquisition cycles. This is also a *methodological key requirement* when analyzing the data. All conclusions derived through the data have to be reproducible and statistically valid. If a result does not meet the requirements, the conclusion will not be accepted in the epidemiological community [296].

The chosen hypothesis directly influences the analysis of the underlying features. Which information is extracted from the medical image data also dependents on the underlying hypothesis. If, for example, a hypothesis includes the analysis of the average volume of the spine vertebrae, different measurements have to be conducted compared to an analysis including the overall curvature of the spine canal.

2.5 EPIDEMIOLOGICAL DATA

Epidemiological data are highly heterogenous and incomplete. Information about medical history and examinations, genetic conditions, geographical data, questionnaire results and image data yield a complex data space for each subject. For ethical, legal or medical reasons, some features cannot be gathered for each study participant. An obvious example are women-specific questions about menstrual status or number of born children. Follow-up examinations or questions about conditions like medications taken after a diagnosed disease also yield features only available for a small number of subjects.

Indicators for medical conditions as well as questions about a subject's lifestyle are also often *dichotomous*–they have two manifestations (*Yes* or *No*). Dichotomous data can also be derived by aggregating features to yield only two manifestations (e.g., subjects younger or older than 50 years). Medical examinations mostly comprise categorical (e.g., levels of back pain) and continuous values (e.g., age or body size). The distributions of features are also heterogenous. Features indicating rare conditions are sparse.

DATA ACQUISITION TECHNIQUES AND DATA TYPES Data acquisition for population study data is usually carried out by inviting participants to a clinic or center of the study. The acquisition process duration depends on

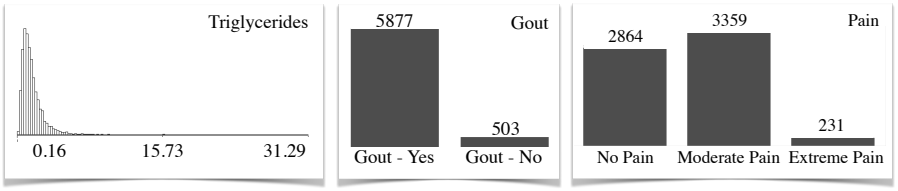


Figure 3: Representation of three features of different type. (left) Numerical feature *Triglyceride Levels* represented using a histogram. (middle) Dichotomous feature *Gout Disease*. (right) Ordinal feature *Pain Levels*.

the number of derived features as well as the accuracy of the examinations. The data acquisition is carried out using different techniques:

- *Expert-guided questionnaires* are suitable to assess lifestyle factors, current medications, nutrition and sporting features. They are also used for a psychological assessment as well as for judging pain levels. The experts are trained to derive standardized answers, which is not always possible, e.g., due to different pain tolerances.
- *Laboratory analyses* allow to extract features about the genetic condition of a subject as well as information from body fluids, such as blood or urine. These features are acquired using standardized methods also used in the clinical day-to-day practice.
- *Medical examinations* are incorporated to diagnose diseases and conditions. They are also used to extract medical parameters, such as heart rate or blood pressure. Just like laboratory analyses, these features are standardized and can easily be compared.
- *Medical image extractions* are carried out to allow for comparisons of inner-body structures. These data are particularly hard to analyze and to compare, as they need a prior segmentation of the structure of interest. Also, acquisition protocols often differ between departments, requiring a strict standardization and calibration of the incorporated machines.

The derived data consist of different data types. These statistical features can usually be categorized as follows:

- *Numerical data* are extracted from metric features, such as body weight, blood fat levels or triglycerides (Fig. 3, left). Numerical features comprise of different distributions. The distributions have an impact on the statistical methods later used in the statistical analysis. Many statistical methods, such as Pearson's R, expect normally distributed features.
- *Nominal data* are categorical data with no inherent ordering of the manifestations. These data are often used to describe lifestyle factors, such as marital status or field of occupation. A special kind of nominal data are **dichotomous** (binary) features, which only assume two possible values. These are very common in epidemiological data, since they are used to describe presence or absence of clinical event. Often, dichotomous features lead to follow-up questions. For example, *"Are you diagnosed with gout disease"* (Fig. 3, middle) may lead to the follow-up question *"Which treatment do you receive for gout disease?"*.
- *Ordinal data* are categorical data with an inherent ordering of the manifestations. Examples are income or pain level (Fig. 3, right). These data are also often incorporated to depict data about lifestyle or medical history. They are also used to depict consumption behavior, such as the alcohol intake or amount of consumed meat.

Another important data type, which is harder to analyze than the aforementioned, are *spatial data*. These data may be extracted either by questions yielding geographical data, such as home or work address. Geographical data are hard to assess using standard statistical methods. Therefore, this data is often neglected in the analysis. This work does not focus on analyzing geographic data, but rather focuses on another spatial data type, *medical image data*.

IMAGE ACQUISITION Imaging techniques involving ionizing radiation for the subject are not suitable for ethical reasons. Therefore, Magnetic Resonance Imaging (MRI) is the main method for collecting population study imaging data. The image quality is a trade-off between accuracy and affordability [208]. This often yields image resolutions inferior to those of clinical day-to-day practice, which makes their analysis more challenging. The equipment used to gather medical image data is kept, if possible, on the initial software and hardware version during a longitudinal study to ensure comparability in and between acquisition cycles.

IMAGE ANALYSIS Decisions have to be made on how image data are *compared* and *quantified*. The use of segmentation masks enables a morphometric analysis using derived metrics, e.g. volume, largest diameter or aspect ratio. As stated in the VAST'14 publication [293], "reliable and efficient segmentation techniques for these data are not available in general, epidemiologists are forced to measure the data by hand, which is a very tedious work with respect to the number of necessary landmarks and the number of subjects. Information derived by landmarks, such as top and bottom point of a vertebra, are by far not as expressive and versatile as segmentation masks describing its whole shape. They are also prone to a high inter-observer variability and are hard to reproduce. This gains even more importance when analyzing multiple time steps. Morphometric information from landmarks comprises thickness, diameter or length of a structure as well as grey value distribution in an area (used for determining the type of tissue)." Fully- or semi-automatic annotation techniques, however, show already promising results for different organs and structures, e.g. in MRI scans of the liver [80], kidney [81] and spine [213]. The methods are, however, custom-tailored to the provided image data and consist of detailed assumptions and domain knowledge about the target structure shape, its variance, as well as expected intensity values. Therefore, they will likely have to be adapted to work for other structures, data resolutions and sequence settings.

DATA SOURCES Epidemiological recommendations are only as good as the data they are based on. Hence, multiple large projects have started worldwide to gather a substantial amount of health data. The *Rotterdam study* [109], which started in 1990, employed also non-invasive imaging data, primarily ultrasound and MRI data. Petersen et al. [202] report on six studies involving cardiac MRI from at least 1,000 individuals in population-based studies. These high-dimensional data allow for answerering analysis questions, such as "*How does the shape of the spine change as a consequence of age, life style and diseases?*" The *National Cohort* in Germany aims to gather data for 200,000 subjects in acquisition cycles of 4-5 years [46]. Other important studies are the *National Lung Screening Trial*, which analyzed approximately 54,000 subjects in a two year time span using either a *low-dose* helical computed tomography (CT) or a single-view chest radiography [253].

Policies exist to decide who gets access to a subset of the data, strictly anonymized, where the collection of data should make it impossible to infer the person's identity. Scientists have to write proposals to get access to the data. Each requested study feature has to be reasoned in detail. Ethics

committees evaluate these proposals. Requesting large sets of features for explorative analyses will most likely be an unusual request for the committees. Therefore, scientists have to make sure that the use of explorative analysis approaches is comprehensible to the scientists involved in the decision making.

The current trend of open science and open data sets yields also freely available data sets. The UK Biobank [289] study in the United Kingdom, with information about genetic predispositions as well as lifestyle and medical features, made its dataset available in March 2012.

The Global Health Data Exchange created by the Institute for Health Metrics and Evaluation at the University of Washington is a catalog for population and health data and provides a web service for the data sets.[1]

MISSING DATA Data are maybe missing, since epidemiological data are incomplete. Subjects may decline answering certain questions, e.g., regarding their alcohol consumption behavior. Other features can only be acquired for a specific subject group, such as women-specific questions. Some features are follow-up questions, such as reason of retirement, or treatment of a specific disease. Imputations in epidemiological data are largely avoided due to the risk of introducing a bias into the data. Other features may be logically imputed. One feature could, for example, capture whether the subject is or was a smoker. A follow-up question regarding the number of years smoking will then only assume values for subjects who answered the prior question positively, leading to a sparse feature. Since the other subjects did not smoke, the feature for them could be imputed with `0 years`.

2.6 STATISTICAL ANALYSIS IN EPIDEMIOLOGY

Statistical analysis is the essential step in epidemiology, which translates the data and hypothesis input into medical knowledge. These conclusions can have a wide range of results with huge impact, e.g., a new risk score for a specific disease. Other epidemiological results may even disprove common medical knowledge, such as empirical information about the volume of the liver. The major impact requires the statistical analysis of epidemiological data to be strict and precise. There is no singular method suitable for all tasks. Choosing the proper statistical test for a specific task is strongly dependent on the underlying hypothesis and the expected outcome. These steps are usually carried out either by statisticians or physicians with a strong background in statistics.

STATISTICAL HYPOTHESIS TESTING Statistical hypothesis testing (synonymous with *confirmatory data analysis*) expects the underlying medical hypothesis to be modeled using a set of random variables [130]. These variables can then be used to validate the hypothesis as a statistical hypothesis. Afterwards, a statistical test determines a test statistic, which is a numerical summary of the data set. The test result is statistically significant if the hypothesis is unlikely to occur solely based on a sampling error. The respective *significance level* is denoted using α and is usually set to 5% or 1%. An analysis involves formulating the hypothesis itself and the alternative null hypothesis, which usually states no associations between observed variables. The statistical test can then be depicted using a *p-value*. It comprises of a continuous statistic ranging between 0 (null hypothesis rejected) and 1 (provided data fails to doubt the null hypothesis). It is also important to understand that the p-values do not indicate the statistical probability of the null hypothesis to be true. The test data provides *not enough evidence to*

1 ghdx.healthdata.org

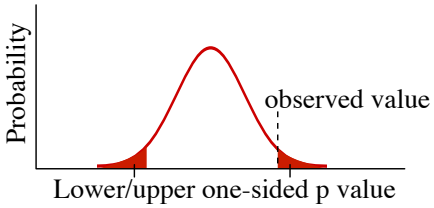


Figure 4: Plot of the probability density of an outcome. A low p-value denotes a very low chance that a random sample was drawn at the edges of the distributions.

doubt the null hypothesis. The p-value can be one- or two-sided [12]. A one-sided p-value states the probability of deriving a result as least as extreme as the observed one direction of the distribution. Therefore, one-sided p-values are divided into lower one-sided and upper one-sided, stating the probability of observing a result either lower or larger than the relative risk (Fig. 4). Two-sided p-values are the sum of the upper and lower p-values and are more conservative than one-sided p-values. The null hypothesis can then be rejected when the p-value is below the specified significance level α [217]. If the null hypothesis is rejected because the α is set too high (statistical test observes significant difference, but in reality there is none), statisticians speak of a *type I* error. The p-value can therefore be seen as equivalent notation of the probability of a *type I* error to happen. With an increasing number of tests carried out on the same data, the probability of encountering a *type I* error increases. Carrying out 10 tests with a α of 5% yields $1-(1-\alpha)^{10}$, a 40% chance of a *type I* error [268]. One way to avoid this problem is increasing the α with each step ($\frac{\alpha}{\mathrm{numberOfTest}}$), which is called Family-Wise-Error-Rate (FWER) [237].

The opposite case, where the test statistic indicates no difference whereas a difference does in fact exist, is indicated as *type II* error. *Type II* errors can be avoided by increasing the statistical power of a test, e.g., by increasing the number of study participants [217]. In good statistical practice, results are not reported solely based on p-values, but rather highlights the complete analysis pipeline, from the design phase, generating and tidying the data to creating the test statistics with information on the underlying method [149]. Effect sizes are also important to quantitatively describe the strength of a statistical phenomenon [129].

Which statistical test is chosen depends strongly on the underlying data type. If the test involves categorical data, usually the chi-squared goodness-of-fit test is used. For numerical features, t-tests are usually applied.

REGRESSION MODELS This paragraph is based on the Regression Analysis subsection of the VAST'15 publication [295]. Regression analysis is the most important statistical tool when analyzing epidemiological data. A regression analysis assesses the influence of one or more (*independent*) features to one target (*dependent*) feature. The regression model yields a function describing the target feature by weighting the independent features. Different metrics, such as the weightings itself and associated p-values, describe the resulting function (the *model*). R^2 values describe the quality of 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:

$$\mathrm{Dependent} \sim \mathrm{Independent}_1 + ... + \mathrm{Independent}_n \tag{1}$$

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

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

The class of the target feature restricts the regression type. Different regression types are available. The focus in this thesis lies on the following [196]:

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

Logistic Regression for Dichotomous Target. Logistic regression implies a dichotomous target variable. The target is described by fitting a logistic function. Logistic models, as opposed to linear models, do not allow for extracting an R^2 quality-of-fit value. Therefore, pseudo-R^2 values are extracted, such as the *Nagelkerke* R^2, which mimics the behavior of the R^2 [182]. *Nagelkerke* R^2 compares the relationship between the likelihood function describing completely independent variables with the actual correlation of the variables. It is intended to agree with the R^2 when both metrics can be calculated. The two metrics cannot be compared directly, as *Nagelkerke* R^2 only mimics the R^2 of linear regression models.

HANDLING CONFOUNDING FEATURES As described in Section 2.1, the best way to handle confounding features is to avoid confounding in the study design phase. Often, this is not possible. Either the data set needs to be queried towards different diseases with different confounders, or some confounders become apparent *during* the analysis phase. If a confounder needs to be considered during the *analysis*, various methods are available:

- *Stratification* is the process of dividing the subjects into different groups. If, for example, the confounding effect of smoking needs to be removed when analyzing the effect between lung cancer and atmospheric pollution, the subjects may be stratified according to their smoking status. Then, the analysis is carried out for each group ("*strata*") [148]. Since these *stratas* can comprise different amounts of subject sizes, simple weighted sums will be more precise on groups with larger subject counts. Therefore, weighting schemes, such as the Mantel-Haenszel method are used to consider these differences in the *strata* size to assess the effect influence [131]. *Stratification* requires more subjects in a study for each considered confounder.
- *Multivariate analysis* is the alternative approach when a higher amount of confounders needs to be considered (usually more than 1-2). This approach incorporates the analysis in statistical regression models, such as logistic regression, or multiple linear regression. In this case, the confounders are considered as covariates of an adjusted analysis. The confounding features are added into the regression formulas to be considered properly.

More advanced and newer methods of handling confounders include finding confounding features using causal graphs based on directed acyclic graphs [85] or marginal structural models [219].

2.6.1 *Statistical Processors and Data Wrangling*

Most statistical processors applied in epidemiology are enterprise tools with proprietary data formats. For basic calculations, physicians often employ `Microsoft Excel`[2], because they know the application and it is often pre-installed on their desktop computers. More specialized analyses are carried out using statistical processors, such as `SPSS`[3] or `STATA`.[4] These tools already include many convenient functions for creating statistical visualizations to communicate the results [174]. Each of these tools has their own proprietary format. This makes it hard to export the data into an open format. Most processors only allow to export the data as character separated values (CSV). In most cases, however, the values themselve are represented using IDs, such as different answer possibilities to a question. For example, the question "*Did you experience back pain in the past three months?*" may have three outcomes: *1 - Yes*, *2 - No*, and *3 - unknown*. The structure translating the IDs to their values is called a **data dictionary**. Both, `SPSS` and `STATA` support data dictionaries in their own data formats, but the export into open formats is not standardized and in some cases not even possible. This makes it hard to process data of these formats. Hence, data wrangling is a challenge when analyzing epidemiological data. Graphic data representation is largely incorporated to present results rather than deriving insights into the data.

The situation improves with increasing popularity of open scientific platforms.

Languages, such as `Julia` or `Python`, compete with proprietary commercial projects, such as MATLAB. In statistics, `R` is a popular and free alternative to the established commercial solutions. However, it lacks the sophisticated user interfaces of `SPSS` or `STATA` and users are required to write commands in order to conduct statistical analyses. This refrains many medical experts without command line experience and is one explanation why the rather expensive commercial solutions are still around. The included features and methods of open solutions, such as `R`, can compete with the commercial competitors. New methods for statistical calculations are usually implemented in `R` as proof of concept.

DATA WRANGLING Converting data into a proper format is the first step when analyzing data from population studies. Large population studies usually incorporate quality control steps, which make sure that the data is in a proper format for further analyses. Since the data acquisition is performed manually, errors may happen. Data wrangling allows to export the data into a proper structure and to find errors. To strengthen their academic impact, large population studies apply quality control protocols to minimize the risk of making mistakes. Often, these controlling measures are put in place *after* the data was acquired. Alternative approaches incorporate continuous monitoring using regression analyses to identify problems, such as defective equipment or systematic bias of a data acquisition expert [94]. The same quality control standards require the data to be converted into a proper format, assign error codes if necessary and follow the predefined data dictionary. Hence, the *data munging* step, which converts raw data into a usable format, is usually not necessary for epidemiological studies. Tools, such as OpenRefine, are nevertheless useful to detect errors that may have been undiscovered in the quality control [91]. This also potentially detects errors occurring by converting data into different formats.

2 Owned by Microsoft; `https://products.office.com/excel`
3 Owned by IBM; `http://ibm.com/software/analytics/spss/`
4 Owned by Stata Corp.; `http://www.stata.com/`

2.7 THE STUDY OF HEALTH IN POMERANIA (SHIP)

This section is based on subsection 2.3 of the VAST'14 publication [293]. After the pioneering Rotterdam study (starting in 1990), several MR imaging study initiatives have evolved. They slightly differ in clinical focus, acquired data and epidemiological research questions. Starting in 1997 with a cohort consisting of 4,308 subjects, the `SHIP`, located in Northern Germany, aims to characterize health and disease in the widest range possible [270]. Data is collected without focus on a group of diseases. This allows the data set to be queried regarding many different diseases and conditions. Subjects were examined in a 5-year time span, continuously adding new parameters including MRI scans in the last iteration [100]. The MRI protocol features a rich number of sequences. A second cohort, `SHIP-Trend`, was established in 2008. The protocols for examining the subjects between `SHIP` and `SHIP-Trend` remained the same, making them comparable. The overall examination time for each person attending the study is two days. A Brazilian cohort named SHIP Brazil is currently established in an area of Pomeranian emigrants. With the standards of the SHIP, the new cohort will allow to analyze differences of risk groups and factors between continents.

3 STATE OF THE ART

This chapter introduces the technical foundations necessary for the methods proposed in the later chapters. It starts with basic visualization and analysis techniques used in complex analysis systems, which are described afterwards. These systems, however, are often not applied to population study data. Therefore, the focus lies on putting them into the context of population study data analyses to show which parts can be applied to solve problems in the application domain of this thesis.

3.1 FOUNDATION FOR VISUAL ANALYSIS SYSTEMS

Interactive visual analysis systems combine visualization, data mining techniques as well as interaction design. In this chapter, techniques are presented, which are incorporated in these systems.

3.1.1 *Visualization of Statistical and Continuous Data*

Epidemiological results are usually communicated using statistical standard diagrams or tables. Graphical representations of single features are usually bar charts (categorical data) or line diagrams (continuous data). The visualization quality is hard to assess and often subject to personal preferences of the domain expert. Tufte [259] proposed two quality measures for information visualizations. The data-ink ratio is defined as:

$$\text{data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to draw graphic}} \tag{2}$$

It measures how much visual information does not represent data. A well-designed graphic should avoid non-data ink as much as possible to steer the user's focus on the data. The goal is a data-ink ratio close to 1, where ink is only used to represent data. Examples for bad data-ink ratios are plots with lots of additional visual clutter to fit it into a specific design. When a line chart displays the development of the oil prices, displaying barrels in the visualization does not add any information value. Tufte refers to this as "chart junk". The second measure is data density, which is defined as:

$$\text{data density} = \frac{\text{number of entries in data matrix}}{\text{area of data graphic}} \tag{3}$$

This metric is solely focused on the space used by a visualization. Small visualizations showing many data elements are preferred. These two metrics are guidelines in choosing visualizations showing the same data types. It can also be used as orientation for creating new visualizations as well as simplifying them. A large pie chart, for example, may have a data-ink ratio of 1, but a very low data density, since the amount of encoded information is very low.

BASIC VISUALIZATIONS OF EPIDEMIOLOGICAL DATA **Histograms and bar charts** are the most basic way of a graphical representation of the distribution of numerical data. Introduced by Karl Pearson, the histogram is created by making equidistant range bins, counting the number values falling into that bin and mapping the count on the height of a bar (see Fig. 3 left) [199]. Therefore, each bin represents a range of data. The larger

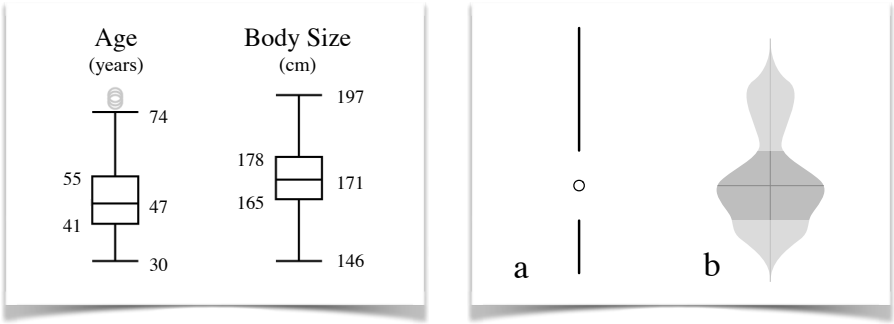


Figure 5: Box plot representation of two variables *age* and *body size* (left). Age includes three outliers, denoted using circles. Two examples of enhanced box plots (right). (a) Tufte's redesigned box plot. (b) A *violin plot* shows additional information about the distribution of the variable by mapping the distribution of the plot diameter.

the bin size, the more variance information gets lost in this plot. Too small bin sizes yield a very flat height profile, which makes it hard to spot trends or major differences. **Bar charts** follow the same paradigm for categorical data. Since there are no bins, each bar in a bar chart represents the count of a distinct value (see Fig. 3 middle, page 10). Due to the similar encoding between bar charts and histograms, bar charts usually leave whitespace between each bar, to distinguish them from histograms.

Box plots are used to visualize a continuous variable and include measures about distribution and the dispersion degree [138]. The box itself is defined by the lower and upper quartile of the distribution. The mean is denoted as line inside of the box. The box plot is best used if the data is normally distributed. An asymmetric box plot with many outliers indicates that the variable follows a different distribution.

The box plot can be altered to be simplified or encode more information. Tufte's redesign of the box plot focuses, for example, on decreasing the ink-to-data ratio and reduces it to two lines to define the space below the lower quartile and above the upper quartiles [259]. The median is denoted using a circle in between the whitespace spanned between the lines. The plot can be seen in Figure 5, right (a). This simplified drawing method allows for better comparison of bar charts sharing the same dimension. The *violin plot* is an extension of the standard box plot, incorporating a kernel density plot showing the probability density (Fig. 5 right b) [108]. Basically, it is a box plot overlaid with a kernel density estimation. A marker denotes the median; the lower and upper quartile ranges are also represented in the same way as in a box plot. Often, multiple box plots next to each other are used to compare variables for different subject group to assess differences. These collections of bar charts are called **forest plots**. They usually also denote different statistical measures for comparison, such as statistical power [150].

Scatter plots are used to depict bivariate relationships between numerical variables. Often, a line of fit is also drawn to depict the course of the distribution (a *regression line*). Multiple groups can be denoted by dying the data points or encoding them using different graphical primitives (e.g., one group as □, the other as △). **Scatter plot matrices** are used to relate a set of variables in bivariate plots by displaying them in a matrix, where each row and column represents one variable. This plot shares the same problem of most plot matrices to get very confusing and cluttered with many variables.

Scatter plots are prone to overplotting for a large number of subjects (Fig. 6). Hence, box plots are used more frequently to assess numerical features.

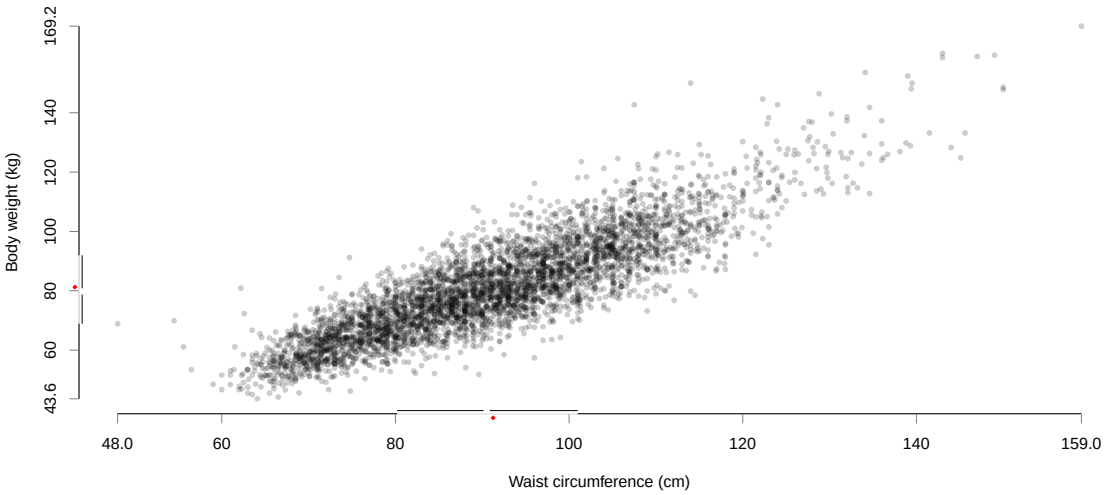


Figure 6: Scatter plot of body weight (kg) and waist circumference (cm) for 4,406 subjects of the `SHIP-Trend-0` cohort. Each data point is rendered with a low opacity to counteract overplotting. Hotspots appear opaque, while outliers are nearly transparent. The axes show information in a box plot using Tufte's design shown in Fig. 5. The red square on the axis denotes the variable mean. The plot was generated using R with adapted code provided by Murdoch [181] and Piwek [204].

Line charts visualize time-dependent variables. Similar to a scatter plot, they show a bivariate variable relationship, whereas one variable is usually the time. The values of the variables are connected via lines to highlight the course. Similar to scatter plots, distinct groups can be mapped to color or point shape. A special version of line charts popular in epidemiology are **Kaplan-Meier plots**, which assess the probability of a subject *not* to be affected by an event (Fig. 7). The plot is often used to depict survival rates over time. The crosses in the curve marks an event (e.g., a patient dies). The dashed lines indicate the confidence intervals of the curve, as the statistical power sinks with each missing patient from the group. The more events occur (patients pass away), the larger is the confidence interval indicated by the dashed lines.

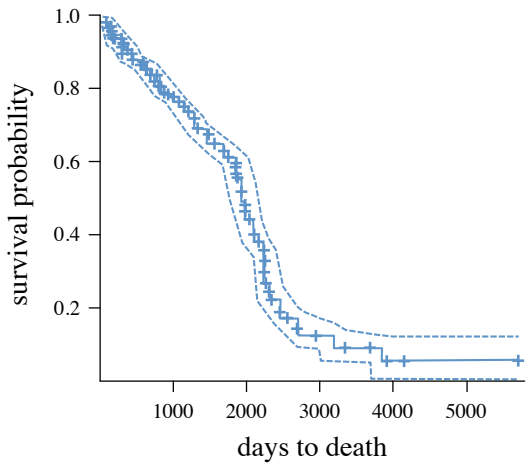


Figure 7: Kaplan-Meier plot depicting the survival probability for a population. Adapted from Preim et al. [296].

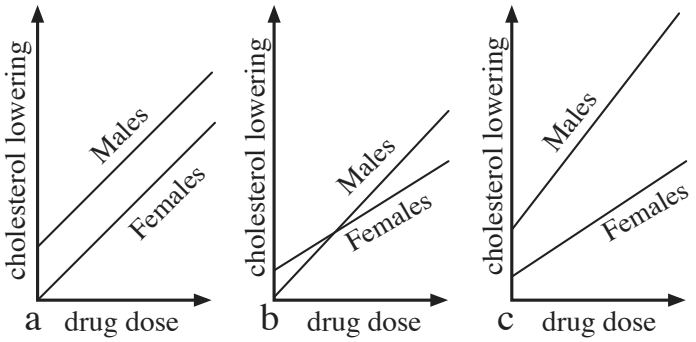


Figure 8: Relationship between the dosage of the drug and their influence on the cholesterol level on patients w.r.t. gender with an interaction plot. (a) Association with gender. (b) Interaction with gender. (c) Association *and* interaction with gender.

A special version of line charts are **range plots**, where each data point consists of range data [174]. The data range is mapped by replacing the data point with a line representing its range.

Pie Charts, introduced by Wiliam Playfair, display variable shares. This chart, however, should only be used with a small number of variable manifestations or shares. A rule of thumb is a maximum of seven subdivisions [58]. Otherwise, the chart tends to be cluttered.

Recent publications urge scientists to use visualizations that show the data, such as scatter plots, box plots and histograms. Results are most frequently communicated using bar and line graphs, which do not allow readers to evaluate the data [275]. Visualizations should be used to facilitate a complete representation of the data.

VISUALIZATION OF ASSOCIATIONS AND INTERACTION TERMS An *association* is also called correlation, meaning that the values of a variable are linked to a second variable. Associations can be calculated using Pearson's R for continuous variables and using Chi-Square tests for categorical data. For a mixture of continuous and categorical data, a one-way analysis of variance (ANOVA) or logistic regression can be conducted. An *interaction* is defined as relationship of at least three variables, where the influence of two variables on a third variable is not additive, meaning that the third variable influences the relationship [51]. In other words, the effect on the third variable is not constant, it varies depending on the value of the third variable. An example would be the relationship between the dosage of the drug and their influence on the cholesterol level on patients w.r.t. gender (Fig. 8). Here, the relationship between the variables is depicted as correlation coefficient using a line. Interaction effects between three or more numerical features are harder to assess, as they cannot be displayed as simple lines. An approach for displaying this data is drawing 3D surface plots, as seen in Fig. 9, or 2D contour plots, where linear prediction is mapped onto color [143].

VISUALIZING MULTIDIMENSIONAL CONTINUOUS DATA Visualization of multidimensional data aims at emphasizing variable relationships. *Parallel coordinates* [118] are a popular tool for visualizing multidimensional continuous data. Each variable is represented using a vertical bar with attached scale ticks. Standard parallel coordinates display each subject as edge connecting the bars and assuming the values of the represented data item. This allows for comparing the relationship of adjacent dimensions. Reordering the variables can be supported by using drag and drop of their respec-

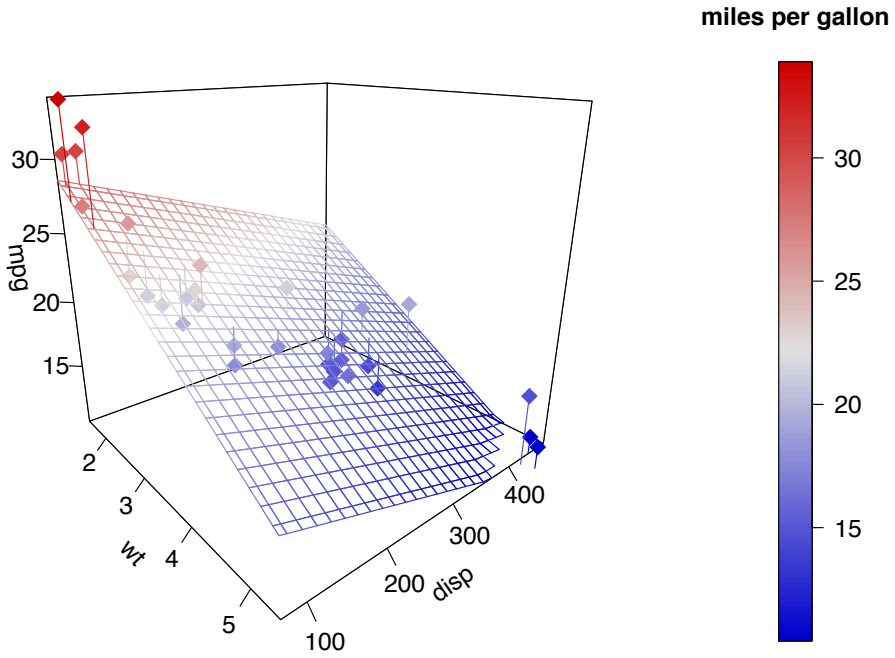


Figure 9: Result of a multiple linear regression of three continuous variables displayed using a plane. The underlying data set comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles [103]. *Wt* measures the weight in lb/1000, *mpg* the miles per gallon and *disp* the displacement (in^3). Miles per gallon (*mpg*) is also color-coded. The plot was created using the `Plot3D` R package from Soetaert [244].

tive axes. Brushing variable ranges usually sets a high transparency value for not-brushed items, allowing to assess the brushed items in context of the whole data set. Parallel coordinates become very cluttered with increasing number of represented items. One strategy for minimizing this effect involves animation by sequentially drawing the items [288]. They are often color-coded by ID to allow a better distinction. Reduced opacity allows to assess the relationship for plots with many items. Clutter can also be reduced by bundling similar curves, similar to edge bundling for graph visualizations [288, 102]. Another solution to cluttered views is replacing the edges with a density plot, which highly increases the plot's readability [102].

Plot matrices allow, similar to parallel coordinates, for pairwise assessment of variables. Examples can be found later in the thesis in Figure 53 on page 124. They are based on Tufte's idea of *small multiples* [259], where a series of similar graphics that use the same axis and scales, allow for fast and easy comparison [78]. The comparability is a trade-off between available space and rendering resolution. Scatter plot matrices arrange variables as x- and y-position, displaying all pairwise combinations in a matrix. In contrast to parallel coordinates, plot matrices allow to compare all bivariate variable combinations, instead of only adjacent ones. Often, additional information, such as the correlation coefficient are also included into the plots. Brushing and linking in this plot is efficient in observing the behavior of a selection in other views. Plot matrices require much space and are therefore only suitable for a small amount of variables. Plot matrices are not restricted to scatter plots.

Generalized pairs plots for example, allow a paired analysis of numerical and categorical data [65]. The visualizations are determined by the variable type combination. Since the plots are mirrored along the matrix diagonal, generalized pairs plots allow for different visualizations above the matrix diagonal to highlight different aspects of the data. For example, the combi-

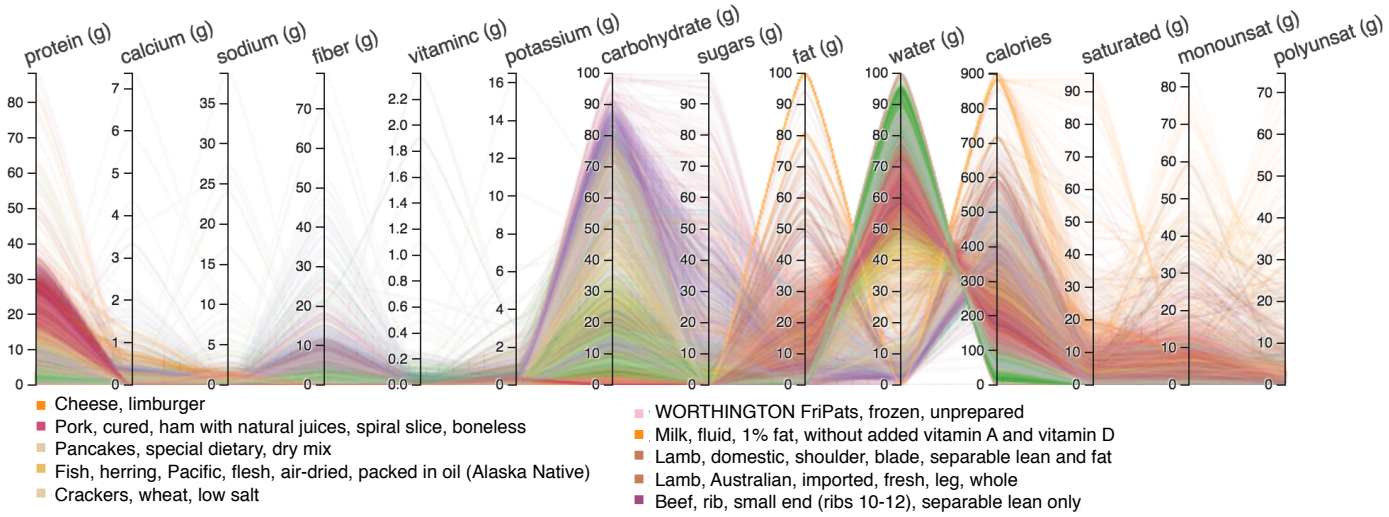


Figure 10: Parallel coordinates plot for 7.637 food products categorized by color and displayed using selected contents. Screenshot taken from web-implementation of Chang [37]. The use of the image was kindly granted by Kai Chang.

nation of categorical and numerical variable can be displayed as box plots below the diagonal and as line plots above the diagonal to check for Gaussian distribution. Generalized pairs plots are incorporated in Section 5.3 and an example can be found in Figure 53 on page 124.

The screen space requirements of parallel coordinates and plot matrices strongly limit their application in the epidemiological application domain if the variable space is not reduced by either an expert selection of features based on a hypothesis or by automatic dimension reduction algorithms.

3.1.2 *Spatial Analysis of Health Data*

A map depicting deaths from cholera was the first graphic in modern epidemiology (Fig. 11). In 1854, John Snow, a famous physician at that time, collected information about deaths from cholera to prove that it is a waterborne disease, contrary to the assumption at that time, that it is airborne. Using this technique, he spotted a cluster of diseased people in Broad Street, around a pump. With this observation at hand, he convinced the local representatives to put down that pump, which led to an immediate halt of the cholera outbreak in this area. Referring to these events, John Snow is now credited as father of modern epidemiology. Since then, spatial analyses have always been an integral part of epidemiology, especially for analyzing the spread and contraction type of diseases. A famous example are the Google Flu trends, where Google uses aggregated search data about health information to estimate flu activity by using the associated spatial location of the users [47]. They built a flu surveillance system, which monitors the flu activity in real time and compared the behavior with usual flu seasons to spot large outbreaks. This approach, however, only works because of Google's popularity and hosting of the most popular internet search engine. Hence, they compensate the data poor quality with a huge amount of data points available. Analyzing a pandemic yields unstructured data mostly acquired from different data sources, such as hospitals and social networks. The importance of this analysis is reflected as the VAST'10 created a challenge to analyze a pandemic data set-based on internet search terms w.r.t. origin of the disease, disease development and spotting of genetic mutations of the disease.

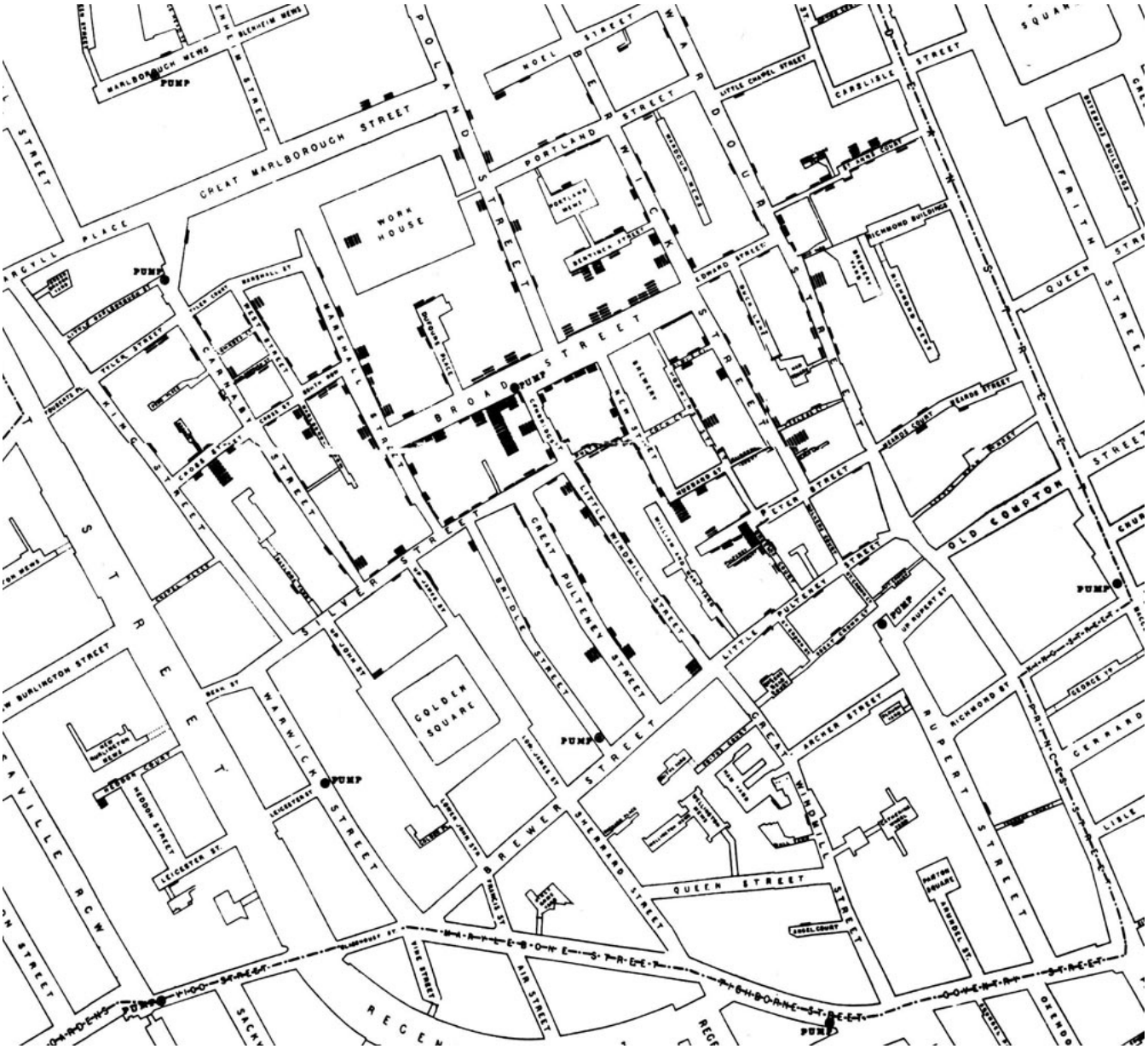


Figure 11: Example of a spatial visualization in epidemiology. John Snow's so-called *ghost map* marks the birth of modern epidemiology. Each point on the map represents a patient diseased from cholera. John Snow observed a cluster around a pump in Broad Street (marked with a ○). Shutting it down led to an immediate stop in the cholera spread in this area. John Snow indirectly proved that cholera disease is waterborne. The image is in the public domain due to its age.

Elliott et al. [63] define three different types of studies:

- **Disease Mapping**: Disease maps are a visual representation of geographic data to provide an overview of the spatial distribution of a specific disease. These visualizations can identify hot spots with high risk and are most prominently used in analyzing pandemic data to steer resources to the regions that need the most help.
- **Geographic correlation studies**: Closer to the goal of large-scale population studies are geographical correlation studies that aim to spot different health outcomes based on other non-spatial data, such as socio-demographic features (e.g., nutrition, sporting activities or income) or medical features.
- **Clustering, disease clusters and surveillance**: Clustering in this context can yield spatial groups of subjects with similar characteristics w.r.t. a disease. Beale et al. [17] point out that small subject groups for specific areas, especially when further divided, e.g., by diagnostic categories or gender, potentially leave not enough samples for the clustering algorithm to work.

Elliott et al. [63] also specified future challenges in spatial epidemiology:

- **Data availability and quality**, which is focused on epidemiological analysis outside of the quality control standards of population studies. The acquired data has to contain accurate health information to allow medically reliable conclusions. The data quality can differ between time points or geographic locations, which may yield false conclusions.
- **Data protection and confidentiality** aim to protect the privacy of subjects when analyzing public health data. For example, when acquiring data from different data sources, the risk of identifying a person by triangulating the data increases. To protect privacy, most population studies do not provide the exact address per subject, but rather the region they live in.
- **Exposure assessment and mapping** describe the problem that spatial data has to be abstracted to aggregate similarities of subject, e.g., the size of tiles, which are used to divide a map. If chosen poorly, the size can highly impact the outcome by being too large or too small.

The first two challenges are easily transferable to all epidemiological studies. The quality assurance departments of large-scale population studies underline the importance to produce bias-free and comparable results and identify error sources. The exposure assessment and mapping step can be transferred to the medical image data domain, where similar problems occur when raw image data is abstracted to segmentation masks and metrics, such as volumes or diameters. Choosing different measures highlights different aspects of the data, which may lead to inconsistent results. Hence, it should be clearly communicated, which information is actually encoded in a visualization to reduce the risk of deriving false conclusions.

VISUALIZATION TECHNIQUES The two major methods for visualizing spatial data in epidemiology are **choropleth** and **isopleth maps** [216]. **Choropleth maps** are divided using geographic information, which is not directly related to the acquired data, such as postal code areas, city districts or states. The areas are color-coded according to total occurrences of an event, such as disease prevalences in the specific regions. Since the areas usually consist of different population densities, the data needs to be normalized w.r.t. area

size and population to avoid a biased visualization. It is the most common geographical visualization in epidemiology, even though it is not without drawbacks. Since the areas are not created data-driven, changes along the borders may implicate false conclusions. This underlines the last challenge identified by Elliott et al., as described above. Additionally, the color of small areas may be harder to perceive than the color of larger ones. Different subject counts in the areas may also bias the data. *Choropleth maps* are often falsely denoted as *heat maps*, since they often use transfer function mapping values on a scale from red to yellow to green.

Isopleth maps are familiar to most users as representations of temperature or rainfall information in weather forecasts. The map is overlaid with color-coded areas spanned by contour lines (lines of equal value, "*isopleths*"). These contour lines are calculated for data over a certain area, such as population density. Ranges with similar values are coded with the same color. The data is interpolated to cover space without data. This requires a *large amount* of *evenly distributed* data to be truthful.

Less prominent are **proportional symbol maps**, which plot information about specific areas on the map on geometrical primitives. For example, the population of cities mapped on the diameter of a circle positioned on the location of each city. John Snow's ghost map is a **dot map**, where the occurrence of an event is mapped on a geometrical primitive, such as a dot or a line at the occurrence location. These maps are used best for spotting spatial patterns [16]. Appropriately sizing the dots is crucial to avoid cluttered maps for large data sets. *Dasymmetric maps* are a fusion between choropleth and isopleth maps. The areas reflect the number of residents and therefore better display the population distribution [243].

Integrated methods for analyzing spatial data in a visual analysis framework are presented by Robinson [220]. They combine multiple views, such as scatter plots, parallel coordinates and line charts. Using brushing and linking, variable ranges can be selected in the views and updated in other views. Carr et al. [34] link small multiples of choropleth maps with box plots line chart visualizations of non-spatial data, called **linked micromap plots**. Small multiples are popularized by Tufte [260] and describe a series of similar graphs or plots with same scales and axes, which facilitate easy comparability. The second visualization proposed by Carr et al., **conditioned choropleth maps**, is used to display data w.r.t. two continuous explanatory variables for hypothesis generation. They divide the continuous variables into three ranges and display the combination of each range using a choropleth map, yielding a matrix of nine maps.

3.1.3 *Shape-Variance Analysis*

Shape variance analysis aims to visually compare structure and form of an object. In medical applications, it is often used to compare representations of tissue by incorporating different modalities, such as CT, MRI or fMRI. In epidemiology, shape variance analysis is focused on displaying population variances or groups with shared characteristics. For example, medical experts might be interested in the mean shape of the male and female liver. Epidemiologists want to derive shape descriptors that are quantitative measures of shape differences. This is, however, not part of the epidemiological day-to-day practice, as both annotation and analysis methods are not generally available. Major questions regarding the shape of organs in epidemiology are:

- *Which different shape groups exist within the population or specific groups* (e.g., divided by gender or smoking status)*?* This is referred to as the analysis of *variation* [28]. This information is of high interest to eval-

uate existing clinical knowledge. For example, a variance analysis of the liver could extract liver shape groups, which can then be compared to the existing textbook knowledge, which then needs to be updated based on the new data [179]. Verifying this knowledge is one major reason for population studies to include medical imaging, because it also allows to assess regional differences in these averages (e.g., European and East Asian livers).

- *Where are the major shape differences/similarities between two groups located?* This is referred to as *comparative analysis* [28]. The focus of this analysis is to spot structural differences between groups, e.g., analysis of the spine for subjects with or without back pain. Observed structural differences are translated into hypotheses regarding the structural change, which can then be statistically evaluated based on suitable image-derived metrics.

Building upon the works of Pagendarm and Post [192], Busking defined a comparative visualization pipeline [28], which builds upon the standard visualization pipeline. The *domain matching* step converts data in a common representation to be comparable. The differences are determined in the *comparison filtering* step, which are then visually represented in the *comparison mapping* step. The final *composition* step usually involves the rendering step to produce the final image.

Miller presented principles of computational anatomy in a survey [173]:

1. Construct (automatically) a representation using points, curves, surfaces and sub-volumes from the image data,
2. compare these representations and
3. statistically verify the anatomical shape and structure to infer knowledge about the disease and underlying structural responses.

SHAPE ANALYSIS METHODS This thesis employs shape variance visualizations based on *Statistical Deformation Models* (SDMs) and *Statistical Shape Models* (SSMs) that capture different shape aspects.

SDMs represent information about a collection of 2D and 3D images by incorporating *deformation fields*, which are usually created by a nonlinear registration of different models. This yields a data structure, where each voxel is associated with a distribution function, capturing information about the voxel variance. This information can be directly extracted from registering image data without the need of a prior segmentation [226]. The ITK-based `elastix` toolbox [133] is a popular solution for creating these data by providing a rich set of registration methods.

SSMs capture shape information of object models by extracting surface grids [207]. Therefore, information is captured about the structure boundary. Deformation information inside of the structure are not covered, this requires an SDM. Due to the missing volume deformation information, SSMs are easier to visualize compared to SDMs. For medical image data, this requires a prior segmentation of the structure of interest. To create SSMs, the correspondence problem between points of different meshes has to be solved. One method to do this is the Growing and Adaptive Meshes (GAMEs) algorithm, which creates a set of meshes with corresponding mesh points for each provided object instance [70].

The decision whether to use SSMs or SDMs depends on the organ or tissue of interest as well as the underlying hypotheses. SSMs allow for comparing surface structures and are well-suited for displaying shape variances of tissue with low elasticity, such as bones. If segmentation masks are available, SSMs can be compared using the GAMEs algorithm or similar approaches.

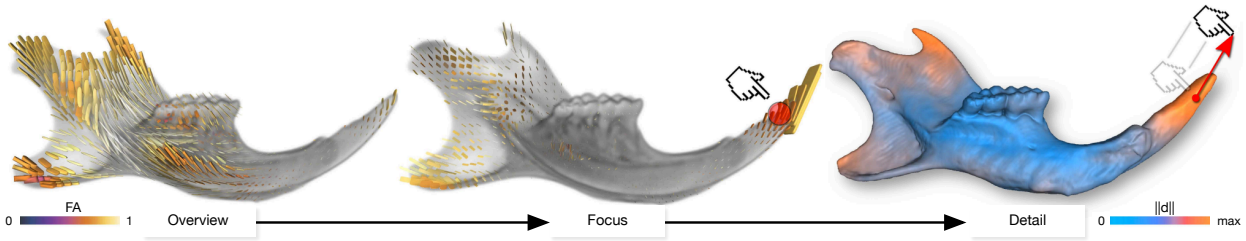


Figure 12: Three abstraction level visualization of anatomic covariation of mouse mandibles proposed by Hermann et al. [104]. The overview shows information about non-trivial covariation patterns together with the shape (left). A user-defined point displays the underlying covariation patterns corresponding to it (middle). By dragging the point, directional dependencies of covariation can be uncovered (right). The image is courtesy of and kindly provided by Max Hermann.

SDMs allow for comparison of a whole tissue structure and are not restricted to the surface. This allows to highlight different densities or tissues inside of a structure. Hence, SDMs comprise of a high information density, which leads to more complex visualizations compared to SSMs. SDMs should be applied when the composition of a structure is of interest, not its shape.

STATISTICAL SHAPE MODEL METHODS AND APPLICATIONS Ferrarini et al. [71] applied the GAMEs algorithm to an epidemiological dataset, yielding SSMs. The model is then used to visualize inter-subject differences by mapping the differences along the mesh normal to color. They found differences in the hippocampus and thalamus for Alzheimer's disease patients compared to healthy subjects, which matches with textbook knowledge. Chou et al. follow a similar approach for Alzheimer's disease by plotting p-values in ventricle surfaces to map disease-associated values directly on a 3D tissue representation [40]. To create the SSMs, they first registered all subjects into one space. Then, surface meshes were mapped into the subject scans using fluid registrations, which yield the different mesh models. The differences were then encoded using color.

Styner et al. [251] present a similar approach for analyzing statistical shape models by converting segmentations of the brain into spherical harmonic descriptions which are then converted into a triangulated surface. Similar to Ferrarini et al. [71], they map differences to color, but also support the visualization with an arrow glyph indicating the direction of the difference. Hermann et al. [104] employ SSMs to identify local deformation changes by investigating shape-related differences on rodent mandibles (Fig. 12). User-specified regions of interest are mapped to associated anatomic covariation using tensor visualization. Specifying a change in shape of a certain area by the user shows how other parts of the model would react using covariance tensors. Covariance tensors indicate how a model is deformed according to user-specified deformations. The user specifies a deformation of interest and showing corresponding changes in the shape using covariance tensors. This method enables rapid hypothesis validation and is able to reproduce textbook knowledge about rodent mandibles.

Lamecker et al. [142] use animation as a different method for comparing SSMs of the liver and the pelvic bone, but unfortunately they do not go into detail on trade-offs of the approach. Busking et al. [30] present methods for comparing two shapes as SSMs, which in an epidemiological context could for example be the average of two distinct groups. They map differences on different visual parameters, e.g., by intersecting surfaces overlaid by contours to highlight intersections between the objects. Busking et al. solve the occlusion problem by rendering the larger object semi-transparent

with opaque glyphs on it, which cast shadows on the smaller object. The conducted user study favors the difference mapping using glyphs over a different approach, where occluded space between objects was filled with fog. The latter approach, however, seemed to be too abstract for the users.

STATISTICAL DEFORMATION MODEL METHODS AND THEIR APPLICATIONS Caban et al. [31] investigated the suitability of four different variance visualizations of SDMs. Likelihood volumes map the density function of the distribution to color. The deformation grid shows the density function directly, but results in a cluttered view for multiple layers and is therefore suitable for 2D views only. Line-based glyphs generate a variation glyph as line for each voxel to display variance. This technique was found useful when combining it with the selection of a region of interest. The user study conducted by Caban et al. favored spherical glyphs, where the variance is mapped on a sphere. The sphere mapping was preferred by users because it introduces less clutter to the visualization than the other methods and reduces the amount of occlusion. Klein et al. [132] applied `elastix` to obtain a B-spline registration of whole brain volumes to cluster on the resulting dissimilarity matrix retrieved from the SDMs. They could divide healthy patients from those with dementia disease. Rueckert et al. [225] utilized SDMs to construct an average anatomy of a structure, which also includes the variability across a population.

The distribution of shapes in a SSM space derived from a PCA is plotted by Busking et al. [29] in a 2D-projected plane of the space. This projection is used to visualize the range of different variation modes. Interpolated views can be created by the user in a separate view as well as comparisons in a contour view. Interpolation is carried out by mesh morphing. The distance to the mean shape is color-coded. Differences between structures are highlighted using color mapping of the difference to the mean shape, but are rather hard to recognize due to small renderings of each subject in the shape space. Via mesh morphing interpolated views can be created by the user in a separate view as well as comparisons in a contour view. The contour view itself allows for rapidly parsing through a single 2D slice of all subjects.

Hermann et al. [105] visualize SDMs using 3D image warping in GPU raycasters. They employ the term *shape ensemble analysis*. They employ group mean shape visualizations that allow to display differences between groups as well as likelihood volumes for variability overviews. Detailed variation is shown using streamline visualizations and requires the user to specify regions of interest.

3.1.4 *Set/Categorical Data Visualization*

A set is a collection of unique objects, denoted as set elements [32]. The set elements are not ordered within a set. Sets can contain overlapping set elements. One example of sets in clinical studies are subjects divided into *healthy* and *pathological* groups to assess the differences. Clustering algorithms usually create sets by grouping the subjects. In public health data, sets can also be derived through different feature manifestations, such as subjects with a specific condition or within a pathological range of features. Different analysis moments can also be viewed as sets. The basic way for representing set data in epidemiology are *contingency tables* (Table 2). By defining categorical variables as rows and columns, the count of the subjects of each possible combination is printed in the table. It is incorporated in most epidemiological scientific papers, because it is a truthful way of presenting the data. Its lack of visual encoding makes it hard to depict trends or hot spots based on *contingency tables*. Hence, they are suitable for commu-

Table 2: Contingency table of *body size* related to *gender* and *back pain*.

Gender	Male		Female	
Back Pain	Yes	No	Yes	No
Body Size (cm)				
139 - 153.5	0	0	149	101
153.5 - 170	286	262	1609	960
170 - 186.5	1341	1123	435	245
186.5 - 203	137	78	0	1

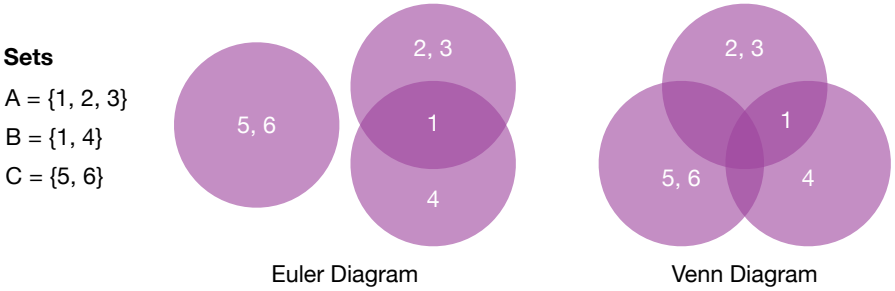


Figure 13: Example of a Venn and Euler diagram for a collection of three sets.

nicating epidemiological results or for cross-checking information observed in graphical plots. Alsallakh et al. [8] and Freiler et al. [75] point out that set-based data is usually not seen as elementary data type in the InfoVis literature and community. Hence, the understanding of their characteristics and visualization challenges is not as well known as for other types, such as graphs or hierarchies [8].

VENN AND EULER DIAGRAMS Venn and Euler diagrams are the most popular visualizations for set-based data. They map each set on a circle or oval as seen in Fig. 13 [15]. Venn diagrams belong to the Euler diagram class, but show *all* possible set intersections, even though the data does not contain such intersections. Hence, the visualization gets cluttered for large group numbers. Both diagram types can be drawn with different restriction aesthetics, such as smooth curves, elliptical and polygonal shape or region shading. Most modern set visualizations are based on Euler diagrams [8]. By mapping the cardinality (the number of elements in a set) on the area of the visual representation, Euler diagrams become *area-proportional*. These diagrams, however, are hard to align and are often not well comprehensible. Area-proportional circle-based Euler diagrams are hard to create, usually elliptical representations have to be incorporated. Enhancing Euler and Venn diagrams with glyphs allow for a different way to display the cardinality as well as additional information mapped on the glyph [8].

Apart from their intuitive representation of sets, Venn and Euler diagrams get cluttered fast for complex sets with many variables. Therefore, their use in the epidemiological application domain is limited. **Overlays** can contribute additional information to these diagrams, such as plotting elements on a geospatial reference [170] on a map or sorting them along a time line [45]. In this case, polygonal shapes have to be used to visualize each set. Overlays, however, are limited w.r.t the number of elements they can represent. The layout may also lead to unwanted artifacts, such as overlaps or crossings [8].

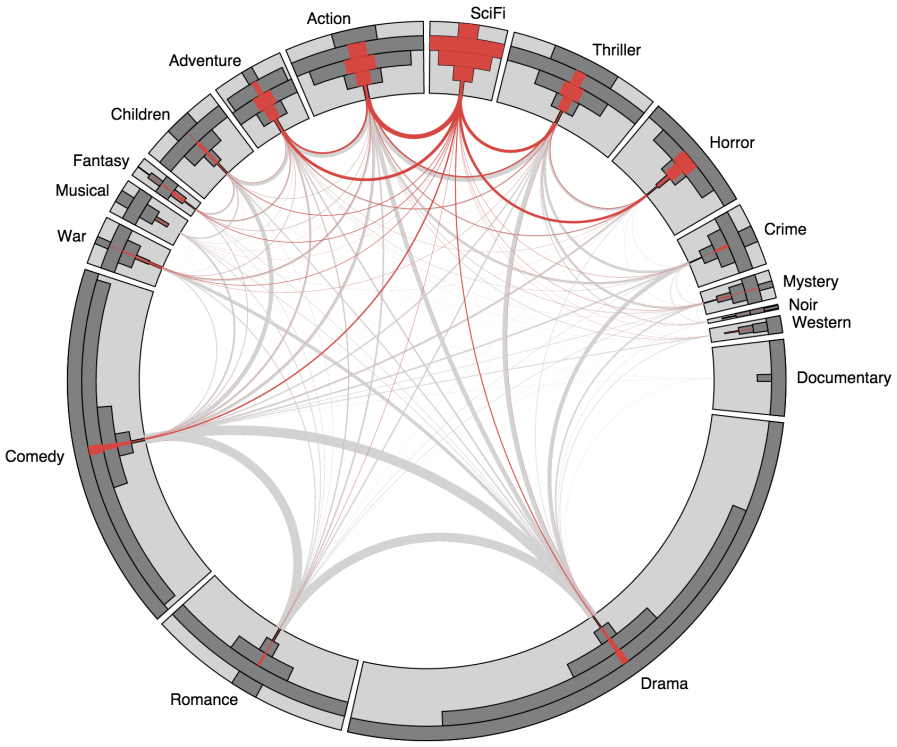


Figure 14: Radial sets visualization of a movie database designed to show overlaps in genres. The bars in each set represent the number of genres of a movie. The selected science fiction category shows that they mostly share the action, thriller or horror genre [145].

NODE-LINK DIAGRAMS Node-link diagrams as well as **matrix-based** techniques are well suited to show *relationships* between either sets or set elements. **Node-link diagrams** are usually graphs, often in circular layouts to preserve the context when displaying different relationships [7, 139]. Each set can be represented as node, shared elements are indicated using an edge. If, for example, each diagnosed disease in an epidemiological data set is viewed as an individual set, the lines would indicate the co-occurring diseases. This may yield cluttered views due to large interconnected disease patterns. Applying fisheye views and color-blending can be used to reduce the clutter [231]. The cardinality of each node is hard to depict in **node-link diagrams**. *Radial sets* render nodes as histograms of the containing elements by degree (Fig. 14) [7]. The degree of an element equals the number of sets that it contains. The histogram shows the number of elements ordered by their degrees of the containing sets. The cardinality is mapped on the size of the node. Selecting a node or a specific cardinality subset highlights the connections. *Circos* is a popular tool for creating a **node-link diagram**, where sets represented using circular nodes are connected using ribbons [139]. It is freely available and can be extended by various visualizations, such as histograms or bar charts. It is particularly popular for visualizing connections in genomic data due to its high data-to-ink and data density ratio.

Node-link diagrams are well suited for showing element relationships as well as highlighting clusters of similar relationships. The diagrams get cluttered for complex relationship patterns, resulting in many edge crossings, which render the edges hard to trace. Complex *Circos* visualizations for example are hard to comprehend without the ability to highlight ribbons of interest. For epidemiological data they are well suited for displaying the co-occurence of indicators, such as simultaneously occurring diseases.

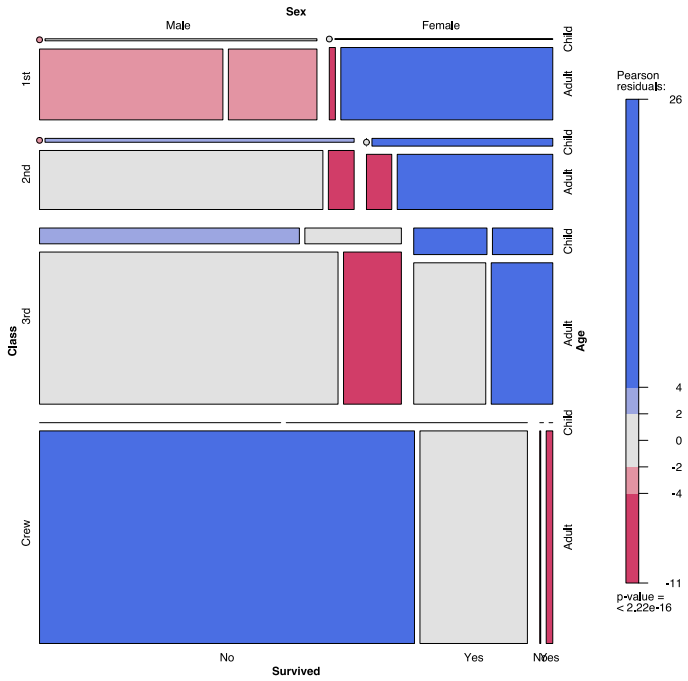


Figure 15: Mosaic plot of the titanic data set [54], displaying three variables: passenger gender, booked class and survival status. Differences in diseased status can be observed in gender as well as in class. The image was created using the `vcd` R package [171].

MATRIX-BASED VISUALIZATIONS Matrix-based visualizations aim to show set memberships using matrix representations [8]. The adjacency matrix introduced by Bertin shows which elements share sets by color coding the respective entry in a resulting *heat map* [20]. *Heat maps* plot each entry of a matrix to color using a transfer function, which is usually defined by the global minimum and maximum value. *Heat maps* are usually employed to visually highlight values to spot potentially interesting combinations. **Matrix-based** visualization approaches of sets share the disadvantages with Venn diagrams by displaying *all* possible intersections. Hence, the adjacency matrix is mostly sparse, which wastes space in the visualization.

Matrix-based visualizations usually scale very well and are therefore suited for large epidemiological data. The complexity of the depicted relationships, however, is limited. Spotting interesting relationships often depends on the matrix sorting.

AGGREGATION-BASED TECHNIQUES Aggregation-based techniques incorporate quantitative representations for sets [8]. They are suited for large sets, where the rendering of individual elements would yield a cluttered view. Interactive bar charts allow for comparison of the elements of a selected set w.r.t all elements in the data [7]. The bars in the radial set plot shown in Figure 14 utilize this technique to display the shares of science fiction movies in other genres. Therefore, *radial sets* are also categorized as **aggregation-based** technique. Set'o'gram stack bar charts represent the subject counts based on their degrees [75]. This technique is incorporated in the *radial sets* as representation of the nodes. Combining this technique with interactive bar charts allows for a selection of granular subsets [8]. *Mosaic plots* allow to display two or more multiple categorical variables to detect relationships [110]. They follow the same paradigm as bar charts, by assigning each bin that represents all possible manifestation combinations of the

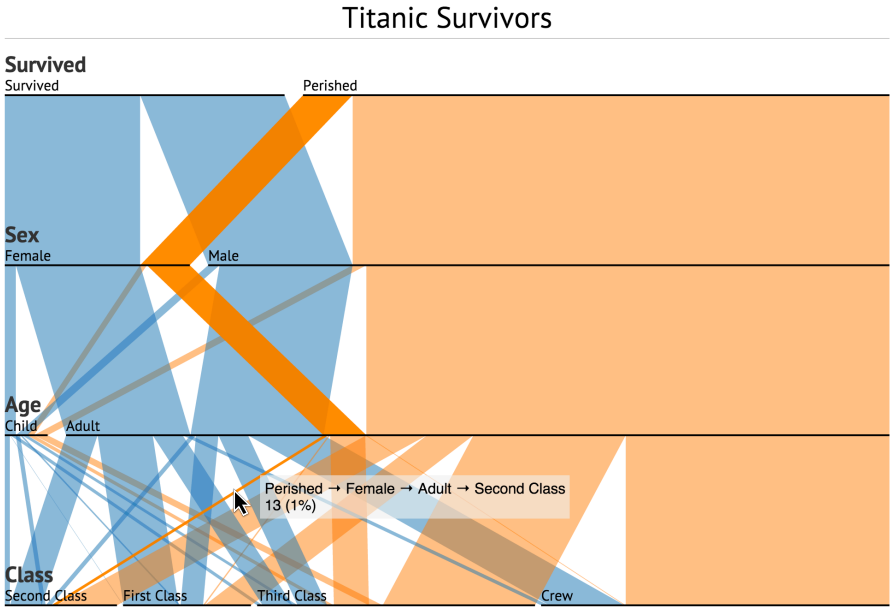


Figure 16: Parallel Sets visualization of the titanic data set shown in Figure 15. Perished female adults of the second class are highlighted and represent 1% of the whole population. The plot gets cluttered with an increasing number of categorical variables, as more combination possibilities are introduced. These increase the visual complexity of the plot. The plot was created using the web-based implementation of parallel sets by Jason Davies [53].

variables a size representing its count. Other than the scatter plot for numerical data, the mosaic plot allows for the comparison of more than two variables. Depending on the number of variables and their manifestation count, the mosaic plot can get visually cluttered. It is recommended not to use more than three to four variables. Mosaic plots do not allow to display confidence intervals [110]. *Double-Decker plots* employ mosaic plots to show the overlap of a selected set in all mosaics [111]. Similar to *interactive bar charts*, the share of the selected set is denoted using color proportional to its cardinality. *Mosaic matrices* show the pairwise combination of variables using *mosaic plots* in a plot matrix [77]. Each panel in the matrix shows the bivariate association between the variables. It is hard to assess the combination of associations of more than two variables using a *mosaic matrix*. This is the strength of *parallel coordinates* and *parallel sets*. *Parallel coordinates* can display categorical data. The resulting plots, however, suffer from overplotting, as most data entries pass through a small amount of points. Hence, parallel coordinates should not be used for displaying categorical data. *Parallel sets* adopt the idea of *parallel coordinates* for categorical data (Fig. 16) [18, 136]. Each variable is represented using a set of adjacent lines. Each manifestation is represented by a line; the percentage of subjects exhibiting the manifestation is mapped to the line width. Lines between adjacent variables represent co-occurring manifestations. Lines are rendered semi-transparent to avoid occlusion. Most *parallel sets* employ highlighting using lines, which renders the selected manifestation opaque. As seen in Figure 16, parallel sets get cluttered with an increasing number of variables. The clutter also depends on the number of variable manifestations and their interconnectivity with the manifestations of other variables. More than four variables usually yield a cluttered view [8]. Hence, they are only applicable to explorative population study analysis systems for a restricted number of categorical features.

3.1.5 *Data Mining*

Data mining is the automatic extraction of valuable information from raw data [128]. The tasks are divided into predictive (classification) and descriptive (pattern mining) [193]. Predictive methods, such as classification or calculation of regression models, aim to categorize new values, as they are introduced into the model. Usual methods for this approach comprise decision trees, support vector machines or neural networks. Descriptive tasks aim to detect patterns, correlations, clusters and outliers. This is the focus of clustering algorithms.

REQUIREMENTS Most clustering techniques expect categorical variables as input [184, 186]. Numerical variables, such as *age* or *body size*, may be discretized into bins of equal size or using quartiles/quintiles. Therefore, results derived using data mining methods have to be analyzed with care to avoid any bias introduced by the discretization. Derived correlations have to be statistically evaluated in order to be of epidemiological value. Data mining methods act as data-driven analysis of epidemiological data and enhance the classical hypothesis-driven approach. An additional challenge in data mining is the parametrization of the incorporated algorithms. Many methods expect input information, such as the number of clusters or sensitivity values to avoid reaching local minima. Strongly varying results depending on small changes in the input parameters weaken the confidence of medical experts in the results of the clustering algorithm. Good default values, which produce results that are robust to parameter changes are desirable.

TECHNIQUES Data mining in epidemiology mostly aims to find separations between subject groups with different outcomes, i.e., different diseases. Hielscher et al. [107] and Niemann et al. [186] study how the similarity among cohort members contributes to improving the separation between members with and without the outcome. Niemann et al. [186] present an interactive data mining tool for the assessment of risk factors of hepatic steatosis, the fatty liver disease. Classification rules derived using data mining methods can be analyzed interactively with their tool and highlight potentially overlooked variables. Subjects were divided into males and females and afterwards analyzed using *decision trees*. The latter is a classification method, which identifies predictive feature ranges w.r.t. the outcome variable and constructs a decision tree based on variable thresholds to separate the classes. The hepatic steatosis indication variable was derived from ultrasound images from an experienced radiologist. The decision tree yields features that separate females older than 52 years (the average entry age of the menopause) as well as males.

Hielscher et al. [107] model subject similarity by finding its *k nearest neighbors* (*kNN*) in the feature space. The *kNN* classifier is trained on a partitioned data set to exclude obvious similarities, such as gender. This way, they identified different separation strategies for males and females. Niemann et al. [187] improved the classification performance by generating features (called *evolution features*) that describe latent temporal information *across* the study waves. This approach allows to correlate feature *differences* between acquisition cycles with disease markers. Experts may define unhealthy ("predictive") feature change ranges, such as an unhealthy increase of body fat percentage and its relationship to heart conditions. The subjects are clustered at each time step. Transitions between clusterings are monitored to highlight cluster splits and cluster merges. One extracted finding was the association between restless legs syndrome (an indicator of sleep disorder) and non-alcoholic fatty liver disease. The features, however, are

still discretized. Hielscher et al. [106] show that the sequence of recordings for some assessments is more informative than the single recording in one acquisition wave. Since new acquisition protocols are added into large population studies with each wave, sequences of recordings are often not available. The analysis complexity is increased, since these additional conditions need to be considered. Niemann et al. [185] also concluded that subspace clustering for epidemiological data is hard to apply, since interesting subspaces are hard to find. Setting parameters is an important issue, as they are mostly not intuitive to the medical expert. Prior knowledge is hard to incorporate as it does not directly translate into parameter settings. Niemann et al. also concluded that the mix of heterogenous data types does not fit the design of the algorithms used, such as the PCA.

Data Mining methods also allow to derive information from complex information sources, which can then be clustered w.r.t. a specific target disease. Roque et al. [222] apply text mining methods to electronic health records to discover disease correlations. They search for keywords from the international classification of diseases provided by the WHO. Hierarchical clustering is incorporated to divide patients into groups. The information has to be treated with care, as clinical data includes many biases introduced by different diagnosis protocols between doctors and different diagnostics equipment. The text mining based on keywords itself is also prone to errors. Large amounts of data minimize these biases, but are often hard to achieve because confidentiality and privacy concerns lead to restricted data access.

3.2 CONCEPTS OF VISUAL ANALYTICS AND INTERACTIVE VISUAL ANALYSIS

This section describes the concepts and techniques of Visual Analytics and Interactive Visual Analysis and highlights differences and similarities of the approaches.

3.2.1 *Visual Analytics*

Visual Analytics (VA) combines data analytics techniques with *interactive data visualization* to derive insights into complex data sets [128]. Therefore, it is an umbrella for a number of scientific disciplines, such as information analytics, geospatial analytics, statistical analytics, knowledge discovery, cognitive and perceptual science, interaction design, and more. VA solutions are integrated systems designed for specific application domains. This eliminates the need of switching between tools and allows for a smooth analysis workflow. Therefore, designers of VA systems need to find out as much as possible about user, task and context to ensure a good suitability for the system. Meyer et al. [172] state that the creation process of interactive visualizations is well understood, but the process of the human reasoning to derive conclusions based on these visualizations is not. The methods have to scale with different complexity levels of both the reasoning and underlying data. To allow for these different levels, Keim et al. [127] define the *Visual Analytics Mantra*, which is divided into four different steps:

1. **Analyze First**: The data are analyzed first using data mining techniques to extract and rank information or group data items. The analysis algorithms at this stage usually rely on an empirically derived parametrization.

2. **Show the Important**: The result of the *analyze first* step is the input for the initial data visualization. It shows information relevant to the user

to enable decision making. Often, overview visualizations provide the user with a mental map for the data. This incorporates highlighting hot spots in the data to steer the user's attention.

3. **Zoom, Filter and Analyze Further**: Overview visualizations often allow for zooming into the data, yielding a detailed view of the data subset. Other interaction facilities allow for brushing data to show information for a subset only, e.g., female subjects. Based on the refined selection, analytics methods can be applied, for example, to derive new information about a data subset or to automatically group them.

4. **Details on Demand**: Detailed information about data subsets, such as summarizing statistics, or values of single data record entries are shown on demand. This may be necessary for viewing a particularly interesting group or individual data records.

These steps are implemented in iterative analysis workflow loops. New insights into the data trigger new questions, which yield new analysis steps with either different methods, different parametrizations of the analysis method or the analysis of a specific data subset. A Visual Analytics system usually provides means for an explorative data analysis *and* for verifying existing hypotheses.

DESIGNING VISUAL ANALYSIS METHODS Designing Visual Analysis Methods requires a thorough requirements analysis. The users of Visual Analysis systems are usually no computer scientists. Hence, the system has to incorporate the application domain terminology. At the same time, trade-offs of the applied analysis algorithms have to be communicated to the user through the visualization to avoid false conclusions. An alternative is the pair analysis approach, where a computer scientist and a domain expert use the visual analysis system and combine their knowledge. The approach will be explained later in Section 3.2.3. A good visual analysis system should:

- exploit the human pattern recognition system by using suitable visual representations and appropriate representations of connections,
- reduce the search for interesting data points by using dense visual representations,
- allow users to perceive a large number of potentially interesting events to classify them,
- provide means for manipulation for both the *data* as well as the parameter space of the *analytics methods*.

Lammarsch et al. [144] describe how the user is included into the Visual Analytics process (Fig. 17). Domain knowledge of the user in combination with observations and research results provides the hypotheses that are validated to formulate models. These models can then be investigated using visual analysis techniques. This yields insights which then again influence the domain knowledge and trigger new analyses. Most VA tools incorporate various visualization techniques using *multiple coordinated views*. This allows to apply different perspectives on the data derived by visualizations, which highlights different characteristics. User interaction with views, such as selection of a data subset (*brushing*), is commonly propagated over all other views (*linking*) [62]. This allows for a quick assessment of trends and patterns by brushing various variable ranges and analyzing the result in views displaying other aspects of the data. Views displaying spatial data, such as geographical or medical image information, usually also incorporate interaction techniques allowing to select the section of data shown (*panning*) as well as its level of detail (*zooming*).

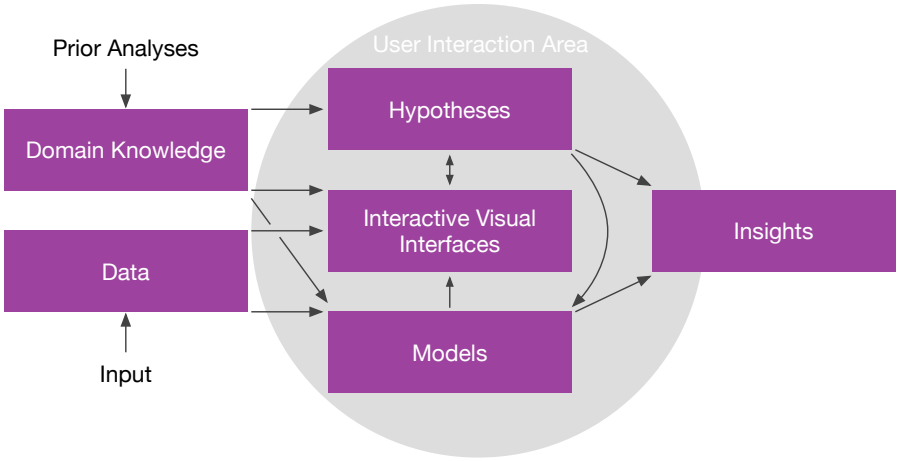


Figure 17: Visual Analysis process inspired by Lammarsch et al. [144] and Keim et al. [128]. It defines a model as a representation of system entities, phenomena, processes and hypotheses as models whose outcomes are not compared with real-world data (validation).

Baldonado et al. provide guidelines for designing systems with multiple views [14]. The design process is described as a trade-off. One view showing much information can potentially confuse the user. Multiple views with less information have to be cognitively connected to be comprehensive. The cognitive workload is determined by

- the system learning curve,
- the amount of information the user has to memorize,
- the complexity of comparing data, and
- the cost of changing contexts.

Additionally, multiple views usually involve more computation time and require more screen space. Baldonado et al. propose rules for designing multiple views [14]. Views have to cover diverse aspects of the data and complement each other. Complex data has to be decomposed to support visual divide-and-conquer approaches, while the number of views should be as low as possible to reduce the cognitive load of creating the connection. Views should be self-evident by including visual clues highlighting the connections as well as consistently mapping data points. Attention management and *guidance* are applied to steer the users' attention to the right view at the right time. An important design lesson of Baldonade et al. is reducing the information depicted in each view as much as possible as well as minimizing the number of views. Reducing the cognitive workload of the users allows them to focus on solving the underlying problems rather than using it all to create the context out of many cluttered visualizations.

A typical VA example are the "cross-filtered views" proposed by Weaver et al. [272]. They incorporate multiple views, which transform the data by mapping elements to visual variables and group elements. The visual representation needs to be brushable in order to link the selection among the views. The representation applied by Weaver et al. ranges from simple lists to maps, small multiple views, histograms and bar charts. The context greatly influences the design of the system, yielding different views for analyzing, e.g., a movie database compared to a baseball game data set. It underlines the necessity of custom building systems with the users, tasks and contexts of the application domain in mind. Weaver et al. also abstract

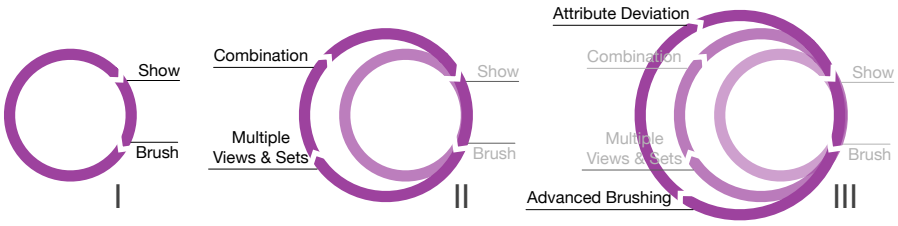


Figure 18: Interactive Visual Analysis interaction levels adapted from Konyah et al. [135]. Brushing and linking (**I**) is combined with a logical combination of brushes from different views (**II**). At the third level (**III**), analytics methods derive new attributes, group subjects, and allow for advanced brushing techniques, which go beyond simple binary range selection.

features by deriving descriptive metrics from them to provide aggregated views. This method is well suited for providing an overview visualization over the data and fit the first and second step of the VA Mantra.

Van den Elzen and van Wijk [266] provide an alternative approach to multiple linked views by combining *small multiple* visualizations with *large singles*. The core idea is to comprehend the analysis workflow by performing splits. Splits are divided into *filtering*, reducing a selection attribute into a defined range; *mapping*, to choose an appropriate visualization for a given parameter combination; and *analytics*, to perform clustering and assess the influence of individual parameters. The result of a split is rendered using small multiples. A mapping split, for example, creates small multiples of the data using different visualizations, such as scatter plots, parallel coordinates, bar charts or radial plots. The user can then choose the preferred visualization, which renders them as a *large single*. This can then be investigated in detail and new splits can be applied. One big advantage is the analysis history, which is created by continuously displaying new plots right of the canvas area. This yields an analysis time line that can be used to comprehend how many steps were taken to derive an insight. It can also be used to return to a prior state of the analysis and move on from there.

The data-ink ratio as well as the data density measures of Tufte (see Sec. 3.1.1) are important guidelines for incorporating clutter-free visualizations that focus on the data. Tufte proposes a set of rules for designing good visualizations [259]: (1) Above all else show data, (2) Maximize the data-ink ratio, (3) Erase non-data-ink, (4) Erase redundant data-ink, (5) Revise and edit. The data density can be increased by shrinking down visual elements without losing legibility or information.

Mackinlay [160] proposed *expressiveness* and *effectiveness* as design principles for visualizations. A visualization is *expressive* when it communicates the data, but the design does not force certain conclusions. Mackinlay requires a visualization to be *effective*–when choosing between two visualizations, the representation which can be perceived faster is chosen.

3.2.2 *Interactive Visual Analysis*

Interactive Visual Analysis (IVA) [273] focuses on the pattern recognition system of the human visual perception. It has a very strong methodological relationship with VA; both disciplines heavily intersect each other. IVA is usually employed for data, which incorporates 3D spatial data, such as medical images or vector fields as well as non-spatial data associated with them. In the visualization community, these data are typically displayed using *scientific visualization* techniques (SciVis), such as volume rendering or flow vizualization. VA is primarily focused on data without 3D spatial data, such as surveillance information. This thesis combines workflows presented

in VA as well as IVA, because herein VA and IVA are seen as different perspectives on the same problem.

IVA incorporates multiple linked views of visualizations, which are combined using brushing and linking. The analysis is iterative, insights into the data yield new questions and hypotheses, which can be analyzed further. The means of interacting with the views is divided into *four different levels* [273, 135]:

1. **Show and Brush** (Fig. 18 **I**). Brushing and linking is the basic IVA level. Relationships between variables represented in different tools can be highlighted well using this interaction technique. At least two linked views are required for this analysis type.

2. **Relational Analysis** (Fig. 18 **II**). Brushing and linking combined with logical operators yield the second IVA level. Multiple brushes can be put in relation using *AND*, *OR* as well as *NOT* statements. This allows for advanced queries in the data, for example in a weather data set only regions with low humidity and warm temperatures.

3. **Complex Analysis** (Fig. 18 **III**). Computational analyses are integrated in the third level, which yield new dimensions, statistical metrics as well as data clusters. By incorporating the newly derived information with the existing data, more complex relationships can be investigated. Advanced brushing techniques, such as angular brushing, can also be incorporated. They show only data points that exhibit a specific correlation described by their angle in a parallel coordinate view [96].

4. **Proprietary Analysis.** An application developed for a data set or an application domain achieves the fourth and highest IVA level. This incorporates the integration of application-specific feature definitions.

In the *IVA* context, data are characterized by a combination of independent variables, such as space and/or time, and dependent variables, like temperature or pressure. Two kinds of views are employed to inspect the data:

- **physical views** [190], e.g., volume rendering, show information in the context of the spatio-temporal observation space [189], while

- **attribute views**, such as scatter plots and parallel coordinates, show relationships between multiple data attributes.

These views are employed using different *IVA patterns* [190]:

- **Local Investigation** is defined as the result of brushing independent variables mapped on a dependent view. This allows to examine characteristics for regions of interest or time frames. For medical image data, derived metrics could for example be displayed in a separate attribute view for all subjects.

- **Feature Localization** can be derived by brushing data points of dependent variables (e.g., temperatures) to see which independent data points in the physical view exhibit these values. This allows to locate entries in the data set with a specified set of features, such as regions of the bone with a very low density, indicating a mineral deficiency.

- **Multivariate Analysis** is conducted by brushing independent variables and observing connections to other independent variables. This employs standard brushing and linking behavior of information visualization views without spatial references. Depending on the employed IVA level, brushing can be used to identify correlations of differing complexities.

IVA TECHNIQUES IVA systems are usually applied when heterogeneous data types have to be analyzed simultaneously. This is usually achieved by connecting different views specialized for either numerical or categorical data, set connections or spatial information. Therefore, basic visualizations become IVA tools by employing multiple view systems as well as brushing and linking facilities. An example are *Set'o'gram*s [75] (see Subsection 3.1.4), which allow for highlighting related blocks by selecting a stacked bar chart. Multiple *Set'o'gram*s allow for linking the selections among these plots, which already enables a *relational analysis*. The visualization is implemented in ComVis [168], a coordinated multiple view system, which also includes other visualizations, such as histograms, scatter plots (enhanced by point size encoding of value frequency) and parallel coordinates. The *Radial Sets* presented in Subsection 3.1.4, which represent relationships between sets using connecting arcs in a graph-based radial visualization, are included in the "Contingency Wheel" visual analysis tool [6]. A bar chart contains the columns of the underlying contingency table data and shows their frequency. The contingency wheel shows the distribution of the table data. The data itself is also displayed using a separate contingency table. The items of a selection can be seen using an element list. Data can be selected in *all* views to highlight their distribution using *all* other views. Additional text-based queries as well as combining selection using union or intersection modifiers allow for *complex analysis* in the sense of the IVA terminology.

Another approach is the simultaneous representation of heterogeneous data types in *one* plot, which includes brushing and linking. Parabox mixes concepts from parallel sets, parallel coordinates and box plots [61]. Each dimension is represented using a vertical bar. Continuous variables are visualized using brushable *box plots*. Analogous to parallel coordinates, lines highlight individual entries. Categorical variables are represented as circles for each manifestation indicating their cardinality. Highlighting a categorical manifestation renders a second set of box plots on top of the now grayed out plots for the whole data set, allowing to investigate the differences. Box plots can be brushed as well. The visualization itself is embedded into a commercial visual analysis tool `Advisor Solutions Data Analyst`.[1] Another example of a visual analysis using a complex single visualization are Generalized Plot Matrices (GPLOMs) [116]. Similar to generalized pairs plots presented in Subsection 3.1.1, they enhance plot matrices to categorical data by pairwise displaying them using a plot matrix grouped by type combination. Continuous variable pairs are displayed using scatter plots, categorical pairs using heat maps and mixed types using bar charts. They are useful to gain an overview of numerous variables and their distributions. Visual elements, such as bars, heat map tiles or scatter plot entries can be highlighted to show their correspondence in the other plots. Textual queries and categorical manifestation selection allow for filtering.

Just like *SPLOM*s, *GPLOM*s require a lot of screen space. An experimental implementation for 54 variables yielded a plot of $20,000 \times 20,000$ pixels. The scatter plot diagnostics ("Scagnostics") proposed by Tukey and Tukey yields a set of measures of scatter plot aspects, such as outlier proportion, data density or scatter plot shape [261]. Wilkinson et al. reduced the computational complexity of Scagnostic measures using Delaunay triangulation to allow their calculation for many plots with numerous data points [278]. They employed Scagnostic measures to order scatter plot matrices. Albuquerque et al. [5] map Scagnostic measures to color in order to steer the user's attention to interesting plots in the matrix. By selecting the appropriate metrics, the user can find plots showing desired behavior. Using brushing, the system allows to create subgroups, which are described using regression models. The

1 Owned by ADVIZOR Solutions, Inc.

technique was applied to a synthetic data set containing 30 dimensions of 900 samples as well as a food composition data set consisting of 18 attributes for 722 samples. In the synthetic data set it could be shown that reordering based on different metrics highlighted well-defined clusters. Similar results are achieved in the food data set, where the *clumpy* Scagnostic measure yields a plot displaying *Manganese* and *Vitamin C*, where two clusters were observed. Brushing the clusters and applying other Scagnostics allowed to assess the clusters further.

The Table Lens [214] adds a focus and context fisheye view to large tables. Numerical variables are depicted by bar charts. Categorical data elements are represented using shaded, colored and positioned points. Selected rows (focal area) show their data using the visual mapping as well as a textual representation. Relationships between variables can be depicted by sorting the tables and observing distribution patterns in other variables. Elements with the same categorical manifestations can be highlighted to observe their behavior when the table is sorted for other variables. New variables can be added using variable mutations defined using a spreadsheet formula input. The strength of the Table Lens is employing the list view, which is already familiar to most users, with simple, yet powerful visualizations that allow for observing relationship patterns.

The problem of solutions incorporating a single view showing complex data relationships is the large visual complexity with increasing number of dimensions as well as the increasing screen space (with exception of Table Lens). Typical IVA applications also employ views for domain and range variables by employing different views.

IVA APPLICATIONS WEAVE (**W**orkbench **E**nvironment for **A**nalysis and **V**isual **E**xploration) was one of the first IVA tools with focus on domain and range variables, proposed in 2000 by Gresh et al. [86]. They link statistical variables into the physical 3D spatial view using color. It contains various visualizations for range features, such as parallel coordinates, histograms and scatter plots. They employ cardiac simulation data and provide a tensor visualization of the blood flow next to 3D renderings of the heart surface. Emphasis was put on the extensibility of the tool to data of other application domains. Advanced brushing as well as analytics methods are not applied. Hence, WEAVE employs the first *IVA* level.

Konyha et al. [135] apply *IVA* methods to assess multiple measurements and simulation runs regarding the same physical entity *families of curves*. The curves are associated with scalar parameters, which yield the *range features*. Emphasis is put on advanced brushing facilities, which allow for the selection of groups with complex rule sets. They employ multiple views with logical brushing, angular brushing as well as similarity brushes. Synthetic data are derived using attribute mutation. The authors point out that the mutation of new variables increases the complexity of the analysis. The new variables have to be comprehensible for the user, which requires a good understanding of the underlying analytics method. Otherwise, the analysis is not effective.

Blaas et al. [22] analyze multi-field medical data, using linked physical and feature space views. These are data from different data sources, which capture different aspects. For medical image data, these information are usually registered to create the context. In this thesis, it is classified as an IVA method. The physical space view is used to display the domain variables. Corresponding range variables are visualized in the feature space view, which is additionally reduced using a PCA. The feature space features that are visualized using scatter plots and histograms are projected into the physical view using color encoding. The system itself uses a grid system

that allows to plug in different views, depending on the current analysis task.

Oeltze et al. [188] analyzed perfusion data using IVA tools. The data was acquired using dynamic contrast-enhanced MRI, where signal intensities over time encode contrast agent accumulation. Plotting per voxel the intensities over time yields so-called time intensity curves. Multiple descriptive parameters are derived from these curves. A PCA is applied to reduce the parameter number and minimize correlations. Oeltze et al. understand features as regions and events, such as regions affected by an ischemic stroke. Thereby, feature localization is achieved by searching the 3D domain representation, which visualizes range features using color coding. Multi variate analysis is carried out by brushing attributes in range views and observing the change in other range views via linked views. Brushing features in the range views and observing the changes in the 3D domain visualization yields local investigation. Via advanced brushing, the system allows the selection of flow with very specific characteristics. Hence, complex hypotheses can be investigated.

The paper is a good example of the strong link between IVA and VA. The workflow presented by Oeltze et al. [188] can also be fitted into the VA mantra. The data is preprocessed in the analyze first step, which yields the representation of the most important features using the domain and range views. These can then be brushed further to derive details. Hence, one can argue that IVA focuses on describing the method complexity using the different levels. Also, it distinguishes between domain and range variables, requiring a physical view and associated information. VA emphasizes the analysis workflow and which methods are suitable in different stages of the analysis cycle.

3.2.3 *Cooperative Visual Analytics and Evaluation*

Most visual analysis systems are designed with one active user in mind. Cooperative Visual Analytics focuses on multi-user interaction, either on one or multiple locations. Emphasis has to be put on inter-user communication and synchronization. By combining different expertises and supporting discussions about the displayed data, the cooperation works best on new data. Social interaction leads to refined strategies in the data evaluation [98]. Cooperative Visual Analysis can be supported through appropriate hardware, such as larger monitors or multiple input devices allowing all experts to interact with the system (e.g., using tabletop systems). Remote sessions are best carried out using web-based applications, allowing for quick and easy setup. Communication can be carried out using existing Voice over IP solutions. To support a steady and easy exchange of ideas and a fast communication loop, the requirements have to be systematically analyzed [98]. Artifacts supporting the information exchange have to be retrieved. The focus lies on an asynchronous cooperation, where an analyst creates results that act as starting point for an exploration by a second analyst.

Shneiderman [240] describes education goals with computing technologies as "*collect* (gather information and acquired resources), *relate* (work in teams), *create* (develop ambitious projects) and *donate* (produce results that are meaningful to someone outside the classroom)". This set of goals is translated into specific tasks to help people to be more creative. Since collaborative work is a highly creative process, these tasks are relevant to the cooperative analytics. Important are "*visualizing* data and process to understand and discover relationships, *consulting* with peers and mentors [...], *thinking* by free associations to make new combinations of ideas, *exploring* solutions [...], *composing* artifacts and performances [...], and *reviewing* and

replaying session histories to support reflection, and *disseminating* results to gain recognition and add to the searchable resources." Isenberg et al. [120] summarize and substantiate these steps by proposing three different levels of collaborative analysis engagements.

- **Viewing** is achieved by presenting static or animated plots and is also suitable for larger audiences. The discussion can only be focused on the displayed information, since the presented information cannot be adapted as a result of new questions or hypotheses.
- **Interacting/Exploring** allows users to select subsets of the data and choose alternative views, usually using a jointly used visualization software. This can be carried out using chat, comments or email, or co-located in one room.
- **Sharing/Creating** is the highest collaborative level, which allows to create, upload and share new datasets and visualizations.

The methods presented in this thesis are evaluated using collaborative analysis using the first two engagement levels **viewing** and **interacting/exploring**. According to Isenberg et al. [120], the analysis is distributed, meaning that the experts are not co-located, which requires means of communications build into the system. This can be achieved by providing screen-sharing facilities as well as means of pointing to interesting areas to avoid communication overhead caused by synchronizing the focus of the experts.

PAIR ANALYTICS Pair analytics is a variant of cooperative analysis [11]. It focuses on a two-user context, where one user has a computer science background and the other one is from the specific application domain. The latter has the high-level control, whereas the computer scientist has deep knowledge of the underlying analysis method and the usage of the system. This allows both experts to focus on their strengths. The domain expert does not have to learn the visual analysis framework interface. The approach requires that the domain expert knows the limits of the visual analysis framework and the computer scientist has good knowledge of the requirements and targets in the application domain. Arias et al. also argue that the role allocation is dynamic; in some cases, the computer scientist could also take the lead by triggering automated analysis steps or showing alternative visualizations [11].

Another major advantage of pair analytics is the communication necessary to capture the mental models of both users. Similar to the think aloud-technique, where a user comments on his or her reasoning, all comments are recorded and evaluated together with the software input logs. However, the think aloud technique is limited, for example when a user focuses on a complex task, which usually leads to little or no commentary. The cooperation can be supported and steered by hard- and software. Facilities to highlight, annotate and save interesting areas are of high importance.

EVALUATION OF VISUAL ANALYSIS SYSTEMS The Joint Action Theory [42] is an established method for structured evaluation of knowledge discovery processes. It discriminates action that show, flag or characterize something, navigate to a position, or confirm an action. They were applied by Arias et al. [11] to evaluate pair analytics sessions. The insights derived through this modality are manifold. The associated effort on evaluating audio/video recordings and interaction protocols, however, is very time-consuming.

Lam et al. [141] summarize seven evaluation scenarios for information visualization systems: They distinguish scenarios for understanding the data analysis:

- Understanding Environments and Work Practices (UWP),
- Evaluating Visual Data Analysis and Reasoning (VDAR),
- Evaluating Communication Through Visualization (CTV),
- Evaluating Collaborative Data Analysis (CDA),

as well as scenarios for understanding visualizations:

- Evaluating User Performance (UP),
- Evaluating User Experience (UE),
- Evaluating Visualization Algorithms (VA).

It should be the goal to evaluate all aspects highlighted by Lam et al. [141], but due to time and personal limitations emphasis has to be put on certain aspects. Hence, the methods are focused on *UWP*, *VDAR* and *UE*.

Understanding the *environment and work practices* (*UWP*) of the application domain is a key aspect to provide methods that are incorporated into the day-to-day practice. It summarizes the requirement analysis by understanding as much as possible about the user, task and context of a domain. As Lam et al. [141] state, "studies that involve the assessment of people's work practices without a specific visualization tool typically have the goal to inform the design of a future visualization tool." The methods for *UWP* are *field observations*, to observe current work practices and how visualizations are already used, *interviews* as well as *laboratory observations*, to allow for a controlled study situation. As stated in Chapter 2, Thew et al. [254] already provide substantial prior work regarding a *UWP* evaluation of the epidemiological application domain.

The methods in this thesis evaluate *visual data analysis and reasoning* (*VDAR*). It assesses the ability of a visualization tool to support reasoning about the data. More precisely, the facilities of a method to provide means of *exploring* the data, supporting *knowledge discovery*, *generating new hypotheses* and leading to *decisions* are analyzed. According to Lam et al. [141], these tasks are particularly hard to standardize and to quantify. For this reason, evaluations are usually *field studies* in the form of *case studies*. Domain experts are observed as they solve evaluation problems using the proposed methods. Data about the evaluation can be derived using the think-aloud technique, where the user is asked to comment on her train of thought. Capturing the evaluation on video allows a detailed assessment of each session.

Lam et al. [141] argue that collecting data over a longer period of time with participants regarding their analysis problems is an additional way to derive *VDAR* insights. Methods comprise logging and self-reporting, e.g., using a diary technique where domain experts compile their experiences in a short text passage on a daily basis. Long-term case studies are *Multi-dimensional In-depth Long-term Case studies* (*MILCs*), where logging, interviews, surveys and observations are combined. This time-consuming evaluation approach often yields a thorough view on the suitability of a method. The evaluations conducted in this thesis are focused on case studies, since the schedules of the domain experts did not allow for intensive longitudinal evaluation strategies.

3.2.4 *Data Mining Visualization*

Ferreira de Oliveira and Levkowitz categorize data mining visualization approaches as follows [55]:

- **Visual data exploration for mining** is mostly carried out using parallel coordinates or scatter plot matrices. The features of a system include filtering, querying and brushing data. These aspects are discussed in detail in Section 3.2.1.
- **Visualization of mining models** conveys the results of mining algorithms. The result of a clustering or classification algorithm can be visually processed and validated by the user. Examples are self-organizing maps [134] to display results of a neural network, dendograms displaying results of a hierarchical clustering or the visualization of a decision tree [67].
- **Visual data mining** combines the prior two approaches into integrated frameworks [279]. This allows to both tune parameter selection and data exploration. In this approach, the visualizations are an integral part of the mining algorithm. Puolamäki et al. [209] provide an overview of visually controllable data mining methods for more detail on this matter.

VISUALIZATION OF ASSOCIATION RULES Detecting association rules is one of the most popular goals of data mining methods [88]. Such analyses often yield a vast amount of association rules. Applying appropriate visualization methods to these data aims to discover the interest rule sets. Association rules are automatically generated insights into variable connections. Visualizations can help to provide a concise view on the results to assess their importance. Certain rules may for example incorporate associations which are already known to the user. This can only be assessed by a medical expert, since it is hard to model complex medical knowledge. Association rules are denoted as if-then rules ($A \rightarrow B$). A denotes the right hand side (RHS) of items in a rule, B the left hand side (LHS) items.

A basic association rule visualization is a scatter plot of two interest measures (typically *support*, which is the proportion of transactions showing the given association in the data, and *confidence*, which can be summarized as probability of finding the RHS rule under the condition of the LHS) as axis dimensions. This allows to display and compare a large amount of rules as points in the scatter plots. Labeling the data points, however, is hard due to overplotting. Graph-based visualizations display items as vertices and association rules as directed edges. Interest measures can be mapped to edge color or arrow width [88]. Graph-based visualizations of association rules become cluttered as the number of rules increases. Yang et al. [282] employ parallel coordinates to display associations. The items are displayed on the y-axis as nominal values. The x-axis shows the position of the item in a rule. Hahsler and Chelluboina [88] propose a matrix visualization of association rules, where relationships are encoded on the data points between data items. The rows denote the RHS of the rules, the columns the LHS. Interest measures are plotted on the point diameter and color. Sekhavat and Hoeber [234] employ linked matrix, graph and detail views to analyze association rules. The matrix acts as overview of all rules. Selected items are represented in the graph view. Details on demand are shown in the detail view. The resulting system was evaluated with data sets containing a varying amount of association rules with a constant number of items (39). The study, containing twelve participants, showed that the visual analysis system scales well with the number of association rules.

VISUALIZATION OF CLUSTERING RESULTS Clustering algorithms are often seen as black boxes by domain experts. Lack in understanding of the functionality of a method weakens the confidence in its results. Hence, clus-

tering result visualization aims to illuminate this black box to make the results comprehensible.

Seo et al. [235] use a multiple coordinated view system to analyze hierarchical clustering results of genome data. The hierarchical clustering approach avoids to set a cluster size prior to the analysis, which is most often not intuitive and already introduces assumptions about the data into the analysis. Hierarchical clustering yields results for any cluster size, but has the trade-off of being complex and hard to analyze. Seo et al. provide a system, which enables an overview using a dendrogram over all the data in order to locate all hot spots. These structures can then be queried, e.g., by restricting the number of clusters. They incorporate scatter plots using coordinated displays to show features and color code cluster affiliation. They incorporate reordering of the scatter plot matrix to show variables first, which are most relevant to the clusters. They also include a gene ontology browser to quickly look up interesting results.

Seo et al. [236] also introduce Graphics, Ranking, and Interaction for Discovery (GRID) principles as guidelines for an analysis approach. Graphical and statistical methods are integrated into a framework, which allows to visualize features using histograms, boxplots and scatter plots as well as detection criteria using 1D or 2D axis-parallel projections. To ensure comprehensive explorative analyses, their first recommendation is analyzing every dimension *first in 1D visualizations*. Then, relationships between variables should be highlighted with *2D plots as well as statistical summaries*. 2D relationships are visualized using heat maps. Observed relationships are then ranked by the user.

Choo et al. [39] present the iVisClassifier. It visualizes results of a linear discriminant analysis (LDA). The LDA is a user-guided dimension reduction method. The dimensions extracted from this method are usually hard to assess. They are hard to comprehend for the user, because it is not clear which information they represent. They incorporate a cluster data set retrieved from an automatic face recognition algorithm. The iVisClassifier employs a visual analytics approach for this by incorporating multiple views by combining a parallel coordinate plot, which contains all dimensions extracted from the LDA together with scatter plots. They incorporate a filtering processes in the different plots to allow for a drill-down to points of interests. The system is restricted to processing numerical data only.

Cao et al. [33] present DICON, an icon-based cluster visualization system displaying statistical cluster information. The cluster size is mapped on the size of the glyph. The glyph shape is defined by the distribution of the underlying data. Categorical features, such as diseases, are mapped to the glyph color. Additionally, they introduce a layout algorithm to align the icons w.r.t. other variables. Judging from the interviews they conducted, the domain experts used the system primarily to define cohorts using the included drag and drop capability. The tool can be used to visualize multidimensional clusters of population study participants using an icon representation for the clusters, which is based on a treemap. Similar clusters appear as similar looking icons.

Turkay et al. [262] incorporate a visual analysis method to analyze structural changes of clusters changing over time. They introduce the temporal cluster view to assess the structural quality of clusters and the type of structural changes, which are called structural signatures and are defined by cluster cohesion and homogeneity. They use silhouette coefficients as clustering structure metric. The proposed interactive visual analysis tool consists of two major linked views. The cluster view displays the cluster quality and structural changes by encoding each cluster as axis in a parallel coordinate view. Rectangles on these axes represent clustering moments, curves between the

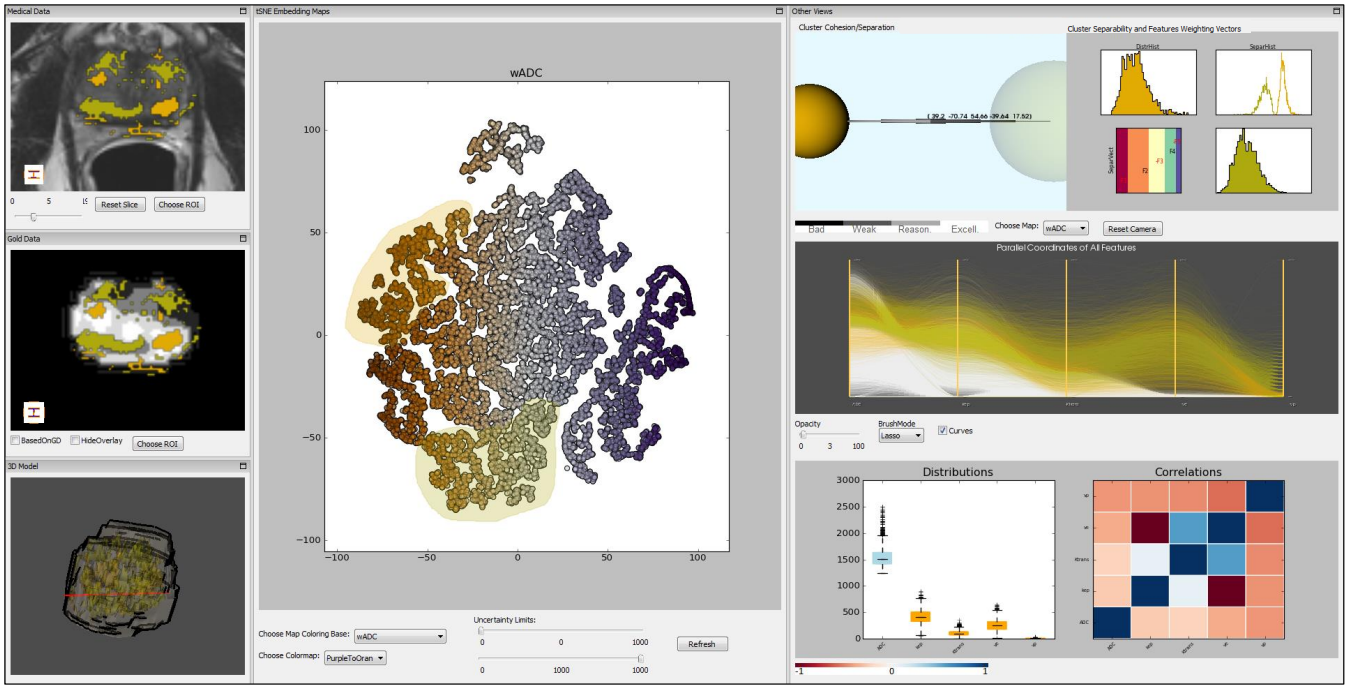


Figure 19: VA system of Raidou et al. [212] for visualizing heterogeneous tumor tissue. It is divided into image space (left) and feature space features (middle and right). The scatter plot in the middle is a 2D projection of the parameter space, which can be used for brushing features of interest. The views on the right can be used to assess details on the clusters using heat maps, parallel coordinates, bar charts and other views. Image is courtesy of and kindly provided by Renata Raidou.

axes represent data items. The temporal signature view visually summarizes statistical properties of the clusters over time to reveal structural changes. It is calculated for a selected group, which must not necessarily belong to a singular cluster and displays the maximum and minimum average distances between the elements as well as the standard deviation, which encodes homogeneity. This view is used to evaluate temporal variations of a single or a group of clusters. The selection is carried out using brushing.

Raidou et al. [212] incorporate a visual analytics approach for heterogeneous tumor tissue (Fig. 19). The goal is to characterize different tissue types based on data derived from medical image data to develop a targeted treatment. The data is divided into image space and feature space variables. The main view is a 2D scatter plot feature space of image-derived tissue characteristics. Detailed analyses of local structures of the feature space can be conducted in separate views. They incorporate a dimension reduction using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to project the N-D feature space derived from the imaging to a 2D space for the scatter plot. The data points can then be brushed in the scatter plot and are linked to their spatial location in the medical image data. Correlations between variables are visualized using Pearson's p-value mapped on a heat map. Parallel coordinates are used to display multiple variables in one plot. A separate cluster view encodes three cluster measures on a glyph using opacity (cohesion), arrow glyph size between spheres (separation) and color (average silhouette coefficient).

PARAMETER SPACE ANALYSIS USING REGRESSION MODELS This paragraph covers work related to the Regression Heat Map Section 5.4. It was previously published in the VAST'15 publication in Section 3 [295]. Sedlmair

et al. [232] present a taxonomy of parameter space analysis. It includes an abstract application domain-independent data flow model, navigation strategies for exploring parameter spaces as well as a characterization of analysis tasks. Using the taxonomy, approaches can be classified to find works with similar problems, derive new design ideas and evaluate ideas. Based on their input parameter taxonomy, some methods proposed in this thesis visualize *model parameters* based on *environmental parameters* in a *global-to-local* navigation strategy (see Section 5.4). There, a *fitting* task by aiming to find models well suited for describing the input data is shown. Mühlbacher et al. [180] provide a framework for qualitative analyses of relationships and ranking features for numerical target features with regression models. Correlations between features with the target yield an ordered matrix plot, where feature combinations are used to depict models of interest. The visualization allows to assess different model complexities. Existing regression models can be validated and compared using 3D views and 2D slice views. Mühlbacher et al. focus 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. [203] propose methods for visualizing regression analysis results and properties for developing car engines. Their main goal is to assess the pairwise influence of independent features w.r.t. the target feature using a plot matrix displaying models as contours. They also incorporate 3D visualizations for each pairwise combination, but mainly because of their popularity with the target domain engineers. Linked views of model deviations allow to select outliers. This limits the method to comparing a few models at once, as the plot matrix becomes complex with increasing feature number. The main difference is their focus on analyzing one complex model in detail, yielding extensive plots. They focus on metric features, while this work processes categorical data as well. Guo et al. [87] present 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. [36] 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. [203], 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. [203], Guo et al. [87] and Chan et al. [36] focus on finding and tuning a model for a specific relationship.

The 3D Regression Heat Map, which is proposed later in this thesis also incorporates regression models. But instead of analyzing one complex model in detail, a large number of models in terms of *different* features is processed in the 3D Regression Heat Map. Ahmadi et al. [2] define the *Sparse Regression Cube*, which partitions sparse high-dimensional data into subspaces, which are then described by their most reliable linear regression model. Their goal is to find most fuel-efficient roads connecting two user-defined landmarks. They focus on an algebraic representation for efficient regression model calculation to find the best fit for a subspace.

DATA MINING TOOLS Numerous tools provide state-of-the-art data mining methods. Weka [90] is an open source framework for preprocessing,

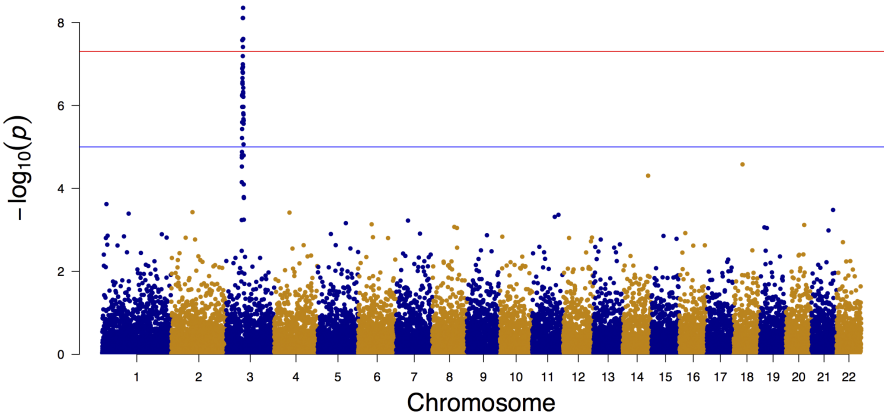


Figure 20: Manhattan plot showing 16,470 SNPs (**S**ingle **N**ucleotide **P**olymorphism, variation of a single base pair of a DNA string) of 22 chromosomes along the x-axis and their association p-values along the log scaled y-axis [265].

visualizing and mining data. Due to its popularity in the data mining community it contains a large collection of various machine learning algorithms. Rudimentary visualization systems allow for displaying data in a plot matrix as well as plotting results from mining algorithms, e.g., with dendograms or decision trees. Rapidminer [112] is a proprietary machine learning environment similar to Weka, but it also allows to define analysis workflows by connecting graphical representations of modules. KNIME [19] is an open source software, which also allows to create workflows that integrate interactive views on the data and on data models. Preferred programming languages with many implemented techniques are R [211] and MATLAB.

Data mining methods are an integral part of Visual Analytics and Interactive Visual Analysis techniques, which are discussed in the following section.

3.3 VISUAL ANALYTICS AND ANALYSIS IN EPIDEMIOLOGICAL AND PUBLIC HEALTH DATA

The concepts described by VA and IVA are well suited for analyzing complex health data, ranging from public health information and biological data to epidemiological study data. Usually, complex relationships of variables are analyzed with focus on a specific disease or condition. The goal is to provide better diagnostic tools by either deriving potential risk factors or indicators. Efficient treatment can be determined by comparing different treatment methods and pathways. The visual exploration of such large information spaces can be superior to an algorithmic analysis if implemented correctly to exploit the human pattern recognition system. A good example are Manhattan plots, which are a scatter plot type incorporated for analyzing genome data (Fig. 20). It associates a *phenotype*, which usually represents a disease, with a set of alleles. The goal is to identify the alleles that are associated with the phenotype. The plot maps the alleles on the x-axis and their associated p-values on the y-axis. A logarithmic scale is usually applied to the y-axis to not distort the plot by alleles with high p-values. The visualization allows for fast detection of alleles with high associations. These associations can then be analyzed in further detail.

Health data analysis is not fully automated, because most analytics methods can only find associations. These associations, however, always have to be assessed by a domain expert, as *correlation does not imply causation*. This underlines the strength of VA/IVA systems. They put the expert in charge

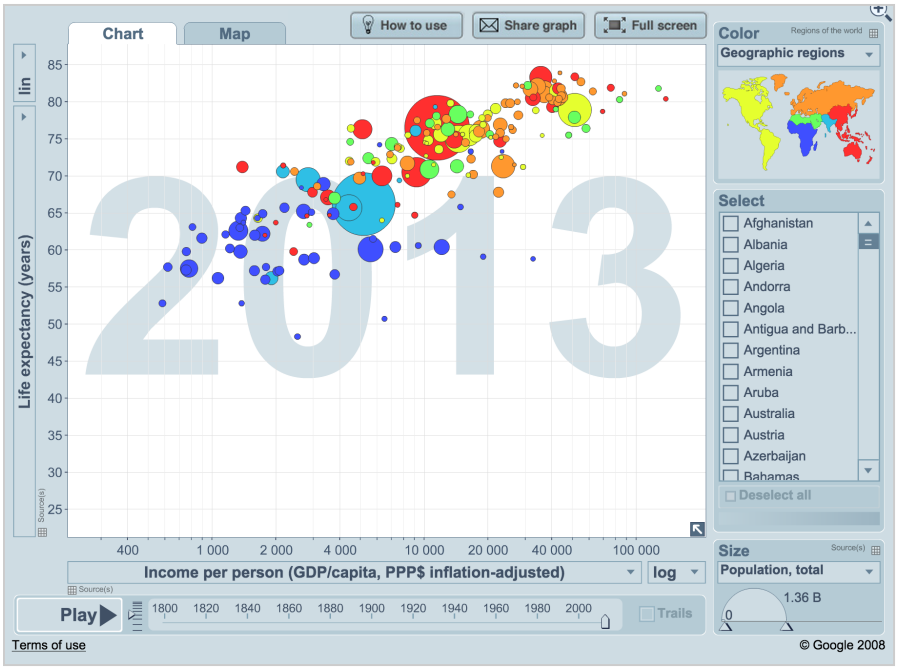


Figure 21: The Gapminder plots the *life expectance in years* against the *income per person* on a logarithmic scale for countries for the year 2013 using a scatter plot. The dot size indicates the total population of the country, whereas the color shows the geographic region. The Gapminder can be used to visualize a total of four *numerical* health variables simultaneously using the x- and y-axis as well as point color and size. The image was kindly provided by Anna Rosling Rönnlund.

of both the analysis methods as well as the evaluation of the results. The expert can be supported in the decision making about the data by exploiting the knowledge, the visual pattern recognition as well as the methodological background.

3.3.1 *Visual Analysis of Biological and Public Health Data*

This section covers two visual analysis systems, that have a major impact on the analysis of public health data. Gapminder was one of the first visual analysis tool that was used to educate a large audience of people using health related issues with visualizations. Caleydo and its open structure is home to a wide variety of scientific analysis methods and is very popular in the visual analysis community.

GAPMINDER The Gapminder proposed by Rosling et al. [223] is probably the most popular visual analysis system of public health data (Fig. 21). Its goal is to *communicate* statistical data to large audiences. It incorporates data sets depicting various variables for countries over time. Data sources are open data sets provided by different organizations, such as the World Bank, the World Health Organisation (WHO), the United Nations (UN), Lancet, Forbes and more. The core of the visualization system is a scatter plot. The variables represented by each axis can be adapted using drop-down menus. Each data point encodes a country. The point size initially shows the population size, but can be adapted to show any other variable in the data set using a drop-down menu. The points are color-coded by continent, but can also encode numerical variables using color scales. Points can be marked, which reduces the opacity of all non-selected data points. Different points in time are shown by *animating the plot*. Therefore, Gapminder incorporates

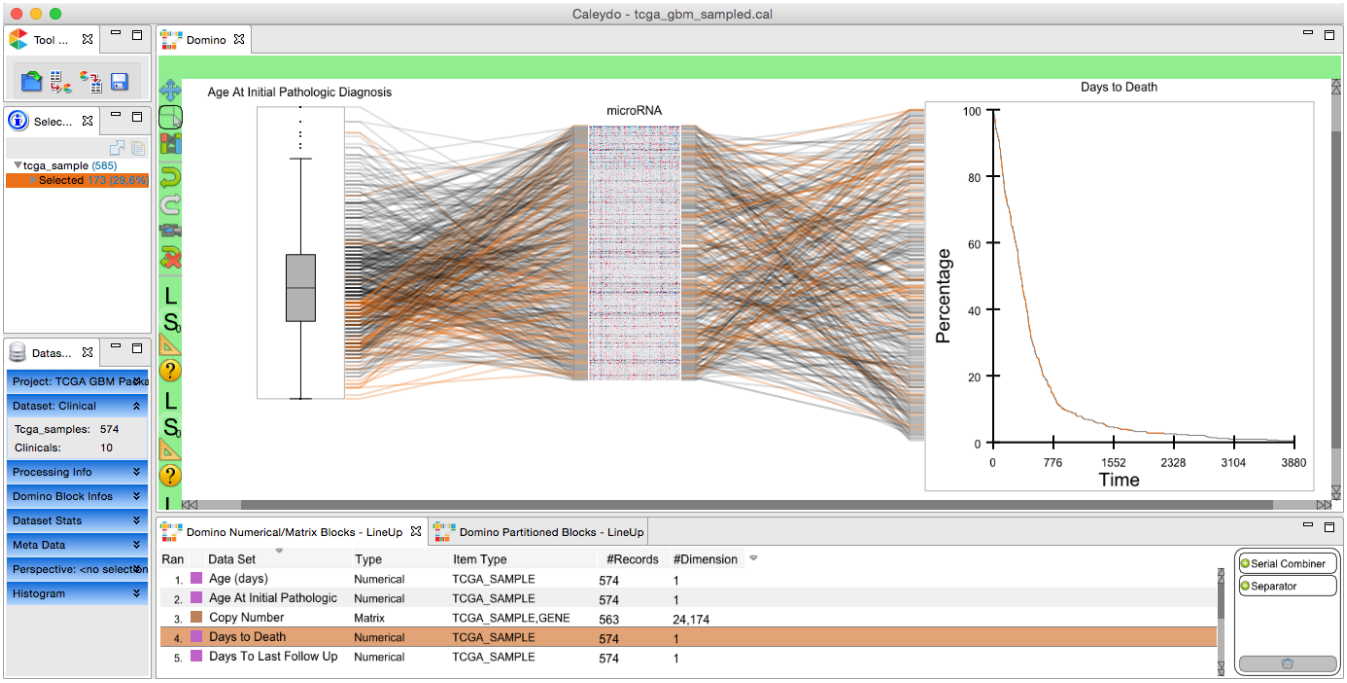


Figure 22: Domino plug-in [84] for Caleydo displaying a glioblastoma multiforme data set. Similar to parallel coordinates, each dimension is connected via lines representing subjects. Each dimension is visualized using an approriate visualization. The age of each subject when glioblastoma multiforme was diagnosed is represented using a box plot. The microRNA, which represents unencoded RNAs playing an important role in the gene regulation, are represented using a heat map. The time to death is encoded using a time plot on the right.

a movie control panel metaphor by providing start/pause and stop buttons, which shows the plot for each point in time for a short duration, creating an animation. The current point in time is depicted as date in the background of the visualization. The animation visually depicts global developments as well as the change in individual countries. The latter can be additionally highlighted by selecting the countries and activating the trail function, which connects the current position of the selected data entry with its prior values.

Besides its good visualization of relationships between country features, the Gapminder owes its popularity to two factors. First, during popular TED conferences in 2006 and 2007, Rosling held energetic talks using the Gapminder to narrate the stories behind the data and popularized the tool. Rosling and his colleagues still give such talks and workshops, which increases the visibility of the tool. The second advantage is the public availability of the Gapminder. It is available as website and can be used without any prior installation. It is very easy to use as it has few but well explained interface elements. Also, the user is not required to manually load any data. Easy and open access further extends the visibility.

The Gapminder is mostly limited to communicate data results. It can be used as analysis tool, but its lack to open data sets provided by the users prevents this. No analytics algorithms are implemented, the analysis is solely visual. Rosling et al. renounce multiple views and therefore do not employ linking techniques. Thus, it does not reach the first *IVA* level.

CALEYDO Caleydo [151, 249] is a good example of a visual analytics framework, which started as software for analyzing genomic data, but is now extended to act as framework for many different techniques suitable for var-

ious data sets. The goal of Caleydo is to combine the visualization of gene expression data with the corresponding gene pathways. Side-by-side relationship views of pathways are possible, but are hard to comprehend, as the cognitive workload of putting them into relation is too high. Caleydo uses visual linking between different views, connecting corresponding data items on selection with lines. Still, the user can only analyze a few pathways simultaneously, as they use up much screen space. This problem is tackled by arranging views in a 3D space, which places each view in a square bucket-like fashion. Gene expression data can be visualized using parallel coordinates and heat maps. Analytics methods can be applied to cluster data. A web browser allows to display details for selected genes. One of Caleydo's strengths lies in its visual linking, where subset selection and details on demand views are propagated among views by connecting the data items using visual cues [151, 249].

Caleydo is built upon the Eclipse framework and acts as foundation for new visualization techniques, which are integrated using a plug-in system. The enRoute [195] extension allows for visual analysis of sub-paths selected in a pathway view. The sub-path is then displayed in context with the data set using a list view containing a small multiple representation using histograms, bar charts, box plots and other plots. The small multiples can be brushed to highlight changes in other views.

StratomeX [152, 250] is a Caleydo plug-in, which displays data sets as columns. Each row represents a data set and data subtypes as individual bricks of the column. Each brick can be analyzed in detail using an expansion view. Pathways can be used to analyze subtypes. By adding multiple data sets and, hence, multiple columns, the distribution between each subset is denoted using ribbons similarly to *parallel sets*. Using StratomeX, Turkay et al. [264] applied their dual analysis approach (see Section 3.2.1) to characterize cancer subtypes. This underlines the flexibility of the open framework, which allows to integrate and combine different methods.

Domino [84] is a Caleydo plug-in for visualizing heterogeneous data and to derive subsets (Fig. 22). Variables can be added via drag and drop to a canvas area. Visualization techniques appropriate for the data type can be selected. The visualizations are visually linked using lines or ribbons. Combinations of features can also be applied, yielding an appropriate visualization such as a scatter plot for two numerical features. The authors show that the versatility of this technique allows users to create many established visualization techniques, such as scatter plot matrices, parallel sets, and even complex systems, such as StratomeX. Using Domino, visualization systems custom-tailored for a specific data set can be easily created. Subsets based on complex rules can be derived using brushing and linking. Domino shows how the Caleydo framework can also be used for data sets other than genomic data.

Most publications in visual analysis and visual analytics of health data implement their own framework. The solutions are isolated applications with little functionality provided to load external data sets of open formats.

3.3.2 *Visual Analysis of Population Study Data*

With their increasing complexity and heterogeneity, population and cohort studies are of high interest for the visual analysis community. Methods for the exploration of cohort data are described first in this section, followed by techniques analyzing event sequence data.

COHORT COMPARISONS Gotz et al. [82] hold a patent since 2014 on *Iterative Refinement of Cohorts Using Visual Exploration and Data Analytics*. It describes a system for creating sub-groups of existing cohorts using visual filters. They describe analytics as a way to modify the cohort as well as expanding it. The patent is based on their CAVA (**C**ohort **A**nalysis via **V**isual **A**nalytics) system [285, 286], which describes a visual analytics system for cohort study data.

CAVA does not solely focus on epidemiology. The data are clinical data sets and therefore the cohort consists only of diseased subjects. Therefore, conclusions of found relationships w.r.t. the general public have to be drawn with care. Three different artifact types, cohorts, views and analytics are used to derive insights into the data. All three artifact types are part of a panel in the system. A list shows the available cohorts for analysis. A cohort can be dragged and dropped to the view panel, which depicts all available visualizations as icons. The demographic view, for example, shows the gender and age distribution of a cohort as well as disease indicators, using histograms, pie charts and mosaic plots. Available views also comprise a table view as well as a flow diagram of patient symptoms over time and small multiples of histograms to compare treatments. Unfortunately, the description of CAVA does not go into detail on the design of those views, how the information is encoded and how they deal with heterogenous variables. The views support brushing and linking, whereas a brush can be applied as filter, which only shows the current selection. This way, the cohort can be refined manually. A history of the selection is kept and can be used to restore prior selections. Cohorts can also be dragged and dropped on entries of the analytics list, which creates a popup window to specify the analytics parameters. CAVA distinguishes two types of analytics. Interactive analytics blocks all user interaction until the interaction is complete, such as BMI calculation. Batch analytics, such as risk stratification, runs asynchronous and allows the use of the system while the calculation is in progress. The analytics then yields further variables and cohorts, which can be analyzed further using the views. The authors do not elaborate in detail on how diseases are modeled or if and how confounders are considered.

COQUITO by Krause et al. [137] is a tool that allows users to apply temporal constraints as queries to a population data set (Fig. 23). The resulting cohorts are displayed using COQUITO, whereas categorical and ordinal features, such as gender or age, are displayed using bar charts. Treemaps show the share of diagnoses and procedures for each subject. Each query reducing the number of subjects is displayed using a graph, where each node represents the result of a query. Circles around each node represent the overall number of subjects contained after applying the query. A search bar allows to locate additional event constraints. The event changes as well as demographic information after each selection are visualized using the implemented plots, allowing to visually comprehend the query results. The main use of the tool is to provide experts with a visual query tool to quickly obtain a specific group of subjects for further analyses.

COMPARE EVENT SEQUENCES The flow diagram of CAVA aims to highlight event patterns. It is derived from the Outflow method of Gotz and Wongsuphasawat [280]. The goal of this visualization is to derive subject groups with interesting temporal event progressions (sequence of symptoms or diagnoses). These events are represented using a directed acyclic graph. This graph is represented using the Outflow graph, which is very similar to parallel sets by mapping sets of symptoms to rectangles, the amount of subjects comprising the set to the rectangle size and the flow of subjects to adjacent sets as ribbons.

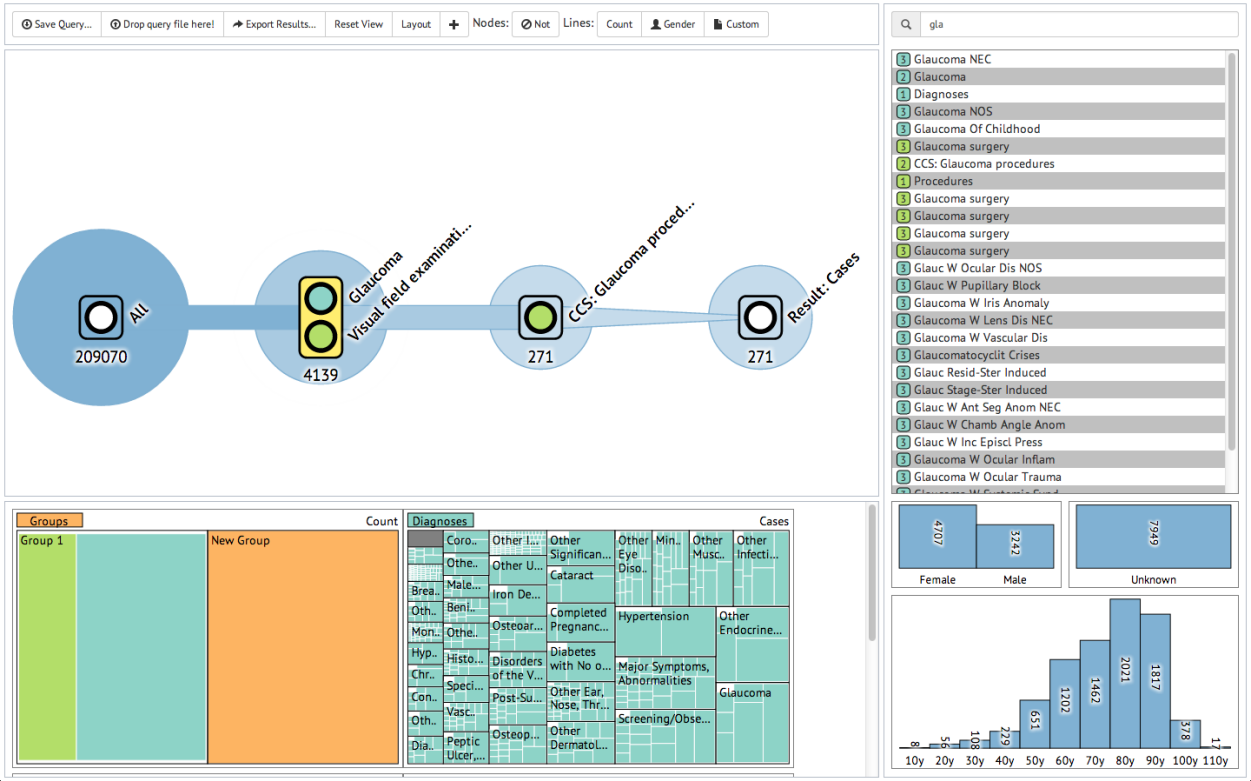


Figure 23: Screenshot of the COQUITO tool of Krause et al. [137]. It allows for a fast selection of desired groups of subjects based on temporal events. The resulting cohorts can be explored using the various plots. The data set is filtered using a search bar of events on the right. Information about gender and age is displayed using bar charts. A treemap displays diagnoses for the currently selected group. The screenshot was kindly provided by Josua Krause.

The CareFlow of Perer and Gotz [201] employs the Outflow visualization for treatment plans instead of diagnoses. Here, the color encodes the health condition of the patients after the treatment. It can therefore be used to identify desirable care plans.

Wang et al. [271] propose Lifelines2 for temporal summaries of the prevalence and temporal ordering of events. Lifelines2 encodes each event as colored triangle along the time line. The user can display temporal summaries of event-associated readings, such as the distribution of creatinine levels among subjects. Temporal summaries can be derived at different granularities (year, month, week, day, etc.). They can also be used for filtering using brushing facilities. The summaries of events can also be used to filter temporal ranges, e.g., by selecting a specific range in a creatinine test result for subjects with a diagnosed disease. Lifelines2 only handles point event data and not interval data with a start and end point [271].

Eventflow by Monroe et al. [177] addresses this problem and extends Lifelines2 to interval data. Monroe et al. aggregate hospital event data into a 2D view, where the x-axis encodes time and the y-axis encodes the number of records. Events are color-coded and mapped to rectangles, which stack for subjects with identical event progressions. The length of the rectangles reflect the event length. The events also include death of patients, hence treatment sequences with a fatal outcome can be identified. This representation can grow complex due to the high number of different treatment sequences and health-related events. In order to analyze specific hypotheses, the user can remove events to reduce the complexity of the visualization.

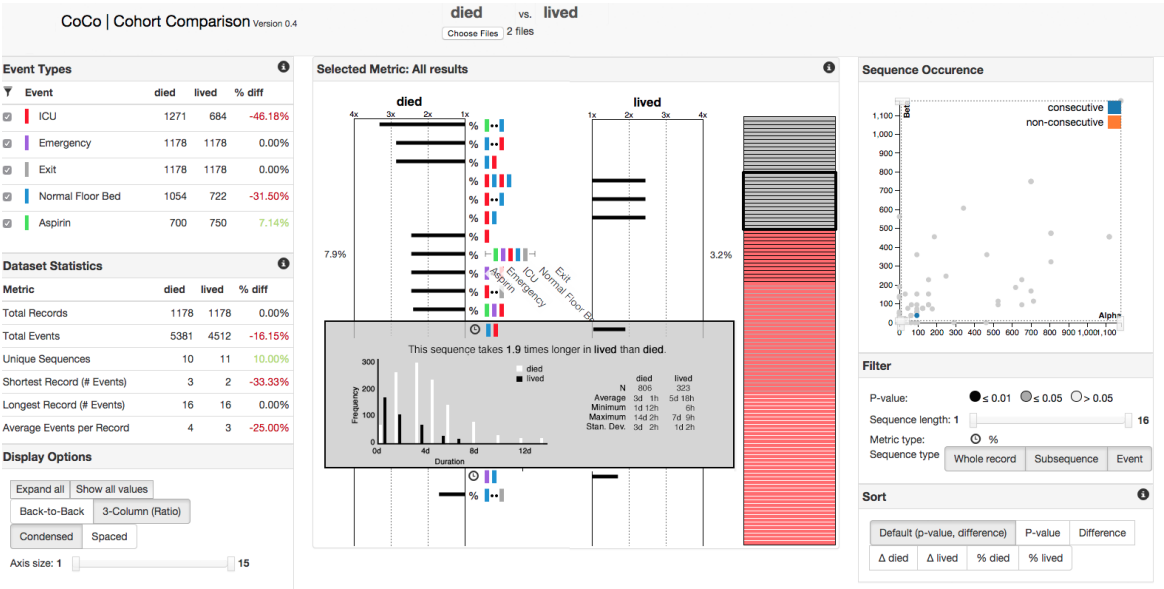


Figure 24: CohortComparison (CoCo) [164, 165] visualizes two data sets of subject pathways in a hospital. One subject group has a fatal outcome. CoCo encodes events using colors and also shows total dataset statistics using a table. The centered panel allows to compare pathways for both data sets and highlights differences in details-on-demand, such as differences in the duration. Reviewed event sequences are encoded gray in a minimap and keep track of the observation space. The user can sort and filter results by event sequence length, event types, sample size, and significance using the right panel. The image was kindly provided by Sana Malik.

Groups with specific sequences can be queried. Event progressions can be replaced in order to simplify the data set by aggregating similar events into an umbrella event. Sequences can be queried in order to show detailed information, such as the exact number of subjects. Uninteresting sequences can be removed to simplify the visualization even further.

MatrixWave by Zhao et al. [287] uses heat maps, which display the pairwise occurrence of sequences. They connect these heat maps in a zick-zack pattern, which allows for tracing the different sequences. The visualization leaves room for all potential event sequences, which requires a lot of space. In most data sets, the observed sequences then require a small subset of elements in these visualizations. They are suitable for very large and complex event sequence data. With CohortComparison (CoCo), Malik et al. [164, 165] aim to provide means of comparisons between subjects with different sequences (Fig. 24). Therefore, various metrics are calculated for subject groups, such as the number of subjects, survival rate, details about the events, such as prevalence, order, co-occurrence as well as total time or duration between events. CoCo includes a selection, filtering and sorting of such metrics, which are then visualized using back-to-back bar charts to pairwise display differences between the groups. Groups can be sorted by p-values or sequence length. Total statistics, such as number of subjects, unique sequences, minimum and maximum size of a sequence are encoded. By querying for specific subgroups, complex groups can be compared to each other to assess the effectiveness of sequences.

Analyzing event data shows high potential for epidemiological data. Epidemiologists can identify pathological pathways leading to a condition. Most epidemiological reasoning rests upon total values and states, such as the additive effect of asbestos exposure and smoking to lung cancer. Event data

analysis may also yield a much higher risk for subjects who transitioned from non-smoker to heavy smoker in a small time frame and are exposed to asbestos. This analysis requires several moments of subject data. The population study data discussed in this work is based on *three* moments, where many features are added through updated acquisition protocols in later moments. This thesis focuses therefore on the latest acquisition moment, which also comprises of medical image data.

3.3.3 *Analysis of Pandemic and Clinical Data*

Analyzing contagious diseases and pandemics imposes different requirements on analysis systems. Marathe and Vullikanti [167] elaborate on the challenges and the future in this field. They suggest the triangulation of various data sources, such as digital social network information, clinical data, census data as well as activity and movement data acquired from surveillance systems to track pandemic sources and characterize infection pathways. The gathered data are then included into graph models, such as Markov chains. The results are information about pandemic risk, vulnerable populations, available interventions, implementation possibilities as well as pitfalls and public understanding [73]. The complex models can then be utilized to simulate various outbreak scenarios and see how the system reacts. The results are used to adapt policies, assess social responses and employ forecasting methods. Marathe and Vullikanti propose easy data availability for querying the data using advanced analysis systems. In reality, these triangulations are hard to conduct, as most institutions and corporations gathering data are bound to legal and ethical confidentiality measures. Providing large networks containing sensitive information imposes the thread of its abuse. Hence, preventive mechanisms have to be incorporated. Unfortunately, Marathe and Vullikanti do not elaborate in detail on how the models are steered and analyzed.

Systems for analyzing such data focus currently on visual exploration of the data. Livnat et al. [156, 155] propose *Epinome*, an epidemiological VA workbench. They focus on identifying characteristics of a pandemic outbreak. The target group are not clinicians, but public health officials. Therefore, the VA system uses different visualizations than CAVA. A geographic information system (GIS) shows features in a spatial context of a choropleth map. Line plots show information about the disease incidence. A contingency table view displays all data entries for detailed analysis of individual records. Similar to Gapminder [223], Epinome incorporates a movie control panel metaphor by providing start/pause and stop buttons, which stepwise show data associated with a specific time interval. The filtering and linking works differently than in most other VA applications. Dragging and dropping a value (e.g., gender = female) into a view shows the filtered value automatically, but does not apply the filter to *all* other views. Dropping a feature in the workspace creates a filter over all views, which is called a *workspace level filter*. This simple yet powerful method allows for assessing specific subgroups in one view, while filtering it in another view containing all subjects and observe changes. They call the approach *loosely coordinated multiple views*. Users rated the system highly useful for the analysis of public health data. The approach of Livnat et al. to provide different filters, "which empower users to explore different hypotheses in adjacent views yet still apply global filtering" [156] is well applicable for a population study scenario, as it adds even more flexibility to the VA than classic brushing and linking. Handling all visual representations of variable manifestations as draggable objects for filtering is a powerful way of adding interactivity without bloating the interface.

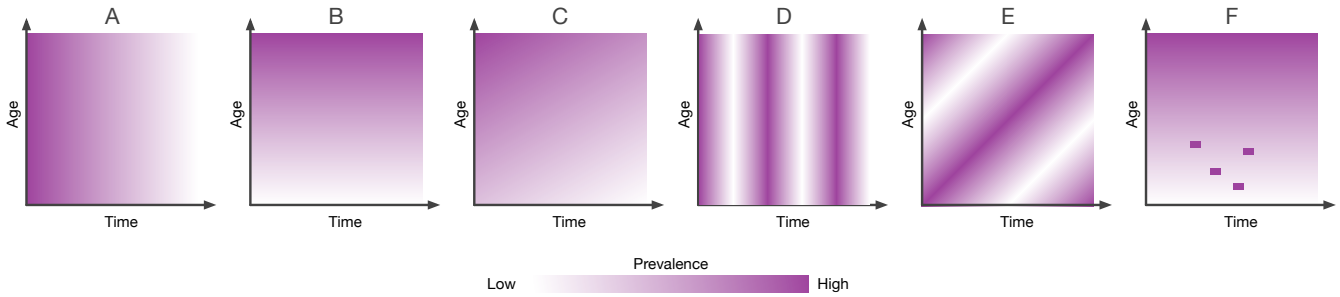


Figure 25: Heat map patterns inspired by Chui et al. [41]. The incidence decreases over time in pattern A and is age-dependent in pattern B. Pattern C is a combination of A and B. Pattern D shows a horizontal striped pattern described as seasonal relationship. Pattern E depicts the age-cohort effect, which indicates that the outcome is seasonal and age-dependent. The last pattern F highlights several hotspots, which can indicate a disease outbreak.

Chui et al. [41] propose an alternative approach by providing a *fixed* number of multiple linked views, which they refer to as *multi-panel graphs*. This term chosen by the authors may lead to confusion with the mathematical construct. It has nothing in common with it but the name. Similar to Livnat et al. [156], Chui et al. focus on public health records from hospitalizations of contagious diseases of several thousands of subjects. The VA system aims to highlight the interactions between age, gender, time and the disease. In epidemiological terms, age and gender are considered a *confounder* (see Section 2.4). The system consists of three visualizations.

- The *outcome pyramid* shows relationships between *age* and *gender* by plotting vertically juxtaposed histograms for males and females along the age axis. The histogram shape allows to infer the distribution, asymmetrical histograms highlight a difference in gender, spikes highlight age-dependent risk groups.
- A *time series plots* is a histogram showing the incidence along the time line, since the prevalence is often highly dependent on subject age.
- To capture this information, Chui et al. propose a so-called *image plot*, which is a heat map. The position of each heat map tile is defined by the age and time, the contrast maps the disease incidence. Different patterns of the heat map were identified by the authors and can be seen in Figure 25.

Using these different patterns, different relationships and influences can be assessed. The authors showed for example that the seasonal pattern for influence can be easily spotted. The multiple views do not support brushing and linking, the domain expert works with static visualizations. The method is well suited for assessing the influence of confounding variables in a data set. *Multi-panel graphs* visualization types can be easily constructed using statistical visualization packages. The approach cannot be customized to specific hypotheses. Missing brushing and linking does not enable drilling down approaches for large data sets.

Multi-panel graphs and the *Epinome* underline various aspects essential for epidemiologists. It is desirable to provide the user with a system that can be adapted based on the current hypothesis. This includes brushing and linking as well as creating new views for feature combinations, which ideally can show various subgroups. Confounding variables have to be identified and considered accordingly to avoid false conclusions. This can be achieved by visualizing feature combinations dependent on a third variable, e.g., by

heat maps. Data acquired from hospitalizations and clinical standard procedures have to be analyzed with care, as they are likely to be produced by different physicians in different clinics, which makes a systematic error (bias) likely. Hence, results need to be cross-checked with other data sources and population studies.

3.3.4 *Combining Medical Image Data With Non-Image Data*

This subsection was published in the VAST'14 paper in Section 3 "Prior and Related Work" [293]. Medical image data is analyzed concurrently with non-image data in multiple view systems, such as WEAVE discussed in Section 3.2.2. Turkay et al. [263] incorporate the idea of deriving descriptive metrics to create deviation plots. Descriptive metrics are calculated for continuous variables derived from MRI scans of a cognitive aging study as well as sociodemographic data. The MRI scans are divided into 45 parts to derive information for specific regions of the brain. These metrics include the mean, standard deviation, median, inter-quartile range, skewness and kurtosis. The authors then incorporate scatter plots of two types. Standard scatter plots allow for comparing two variables, such as age against education level. Each data point represents a subject. Deviation plots display metrics, such as skewness against kurtosis. Hence, each data point in this plot *represents a variable*. The two plots incorporate different backgrounds to make them preattentively distinguishable, yellow for deviation plots, blue for standard scatter plots. By employing brushing and linking with these simple plots, the system becomes very powerful and was successfully applied to analyze epidemiological hypotheses about the data. Turkay et al. call the visualization of data items as well as descriptive metrics for dimensions *dual analysis*.

Steenwijk et al. [246] propose a relational database to organize cohort study data for a visual analysis based on linked views such as parallel coordinates, scatter plots and time plots. Information about medical image data is incorporated via mappers, which extract comparable metrics about the data. Medical image data can be displayed individually for subjects, e.g., for analyzing outliers.

Angelelli et al. [9] focus on the data organization for an interactive visual analysis of heterogeneous cohort study data. The proposed data-cube model facilitates the seamless integration of image-based and non-image data. In a demonstration of the model, brain image data was integrated into the analysis by first segmenting brain regions and tracking neural pathways and then deriving attributes from both, e.g., volume and fractional anisotropy. A multiple coordinated view framework then linked spatial and non-spatial data views.

3.3.5 *Commercial Analysis Systems*

Various commercial Visual Analytics systems, such as Spotfire[2], Qlik View[3] or Tableau[4] are capable of analyzing epidemiological data without the need of writing any code. The systems are focused on business intelligence to allow companies to adapt the commercialization of their products to specific markets. Emphasis is put on highlighting relationships using connected interactive visualizations as well as basic analytics methods. Most recently, these systems are also adapted by journalists to provide readers in digital issues of magazines and newspapers with means to analyze presented data themselves to derive conclusions. Hence, commercial analysis systems are

2 Owned by TIBCO Software Inc., spotfire.tibco.com
3 Owned by QlikTech, qlik.com
4 Owned by Tableau Software, tableau.com

well suited for creating dash boards, which are a set of information visualizations highlighting specific aspects of the data [161]. Creating those dashboards to derive valuable insights, however, requires background knowledge in visualization and statistics to rule out errors introduced by the improper use of visualizations for a specific task.

The main disadvantage of commercial visual analysis systems, besides their cost, is that they are proprietary systems. The code cannot be accessed. APIs for enhancing specific views or adding new views are limited and often require more training than building a system from scratch specifically designed to a task, given the proper set of skills. Enhancing existing visualizations is often not possible. Additionally, the systems are bloated with many functionality, which is not of interest for the domain expert. It distracts them from the main task of analyzing the data. In epidemiology, the target user group for commercial visual analysis systems is not clear. The computer scientists and statisticians usually have the relevant skills to reproduce the provided functionality with tools such as *R* or web-based visualization methods. The clinicians do not have the time to learn the user-friendly but still complex systems to derive the insights they are looking for. Commercial visual analytics systems are well suited for prototyping new ideas for multiple view analyses, but the price tag usually prevents their application. This may, however, change with the introduction of free entry-level versions of the systems, such as Tableau Public.[5]

An alternative approach of commercial systems in epidemiology are personalized medical apps. Regierer et al. [215] show how the virtual patient can be described using anatomical, physiological and molecular models. Virtual patients are the basis of computational analysis and comparison with real patients to allow for a fast and precise diagnosis as well as the optimal treatment plan. The EU project Information and Communication Technology for the Future of Medicine (ITFoM) aims to implement personalized medicine by 2025 with 160 academic and industrial partners [215]. There are, however, already personalized medical solutions available to the patients [252]. 23andMe is a personal genomics service, which provides patients with genome analysis based on a saliva sample. The service is private and will not be ordered and supervised by a doctor. The service was praised by the public media and was even named the invention of the year in the Time magazine in 2008 [92]. Customers can download the raw genetic information from the site. They can also order additional information, such as information about their ancestors and close relatives. Until November 2013, the service also allowed customers to assess inherited traits and genetic disorder risks. As a result of potentially lethal false-positive and false-negative test results, the U.S. Food and Drug Administration (FDA) prohibited 23andMe to sell these tests [10]. The customers were left alone with the results and were not able to precisely assess the risk of their diagnosis. One example are mutations in BRCA genes, which are associated with breast or ovary cancer. The German ethics board urges the EU to prohibit private gene tests [49].

Other services such as `patientsLikeMe.com` [277] rely on self-reporting data to analyze disease progression w.r.t. the underlying treatment and prescribed medications. The basic question they try to answer is: "Who is similar to me and which medical conditions do they have?". The advantages include the data inflow speed as well as a good access to patients for medical scientists. Results, however, still need to be cross-checked with a clinical study [277]. Various biases are introduced in such services. Patients investing time in these services first need to be familiar and comfortable

5 public.tableau.com

with computers and potentially have a higher than average health awareness. Additionally, self-reported responses from medications are inherently subjective. For this reason, these information are assessed expert-guided in large-scale population studies. Issues are raised with subjects experimenting with drugs based on information gathered from services such as `patientsLikeMe.com` [224]. These are often patients who exhausted all treatment options, but do not qualify for clinical trials. Additionally, privacy issues are raised when health data is gathered by cooperations whose sole business model is to monetize this information.

Smartwatches and fitness wearables, such as the Apple iWatch or the Jawbone UP are being widely adopted, yielding data sources for lifestyle factors, such as sporting activities or nutrition as well as medical information, such as the heartbeat or sleep cycles. Most applications employ this information to provide fitness plans and guides for the users. For example, the user can compare their sporting activities to other users. Smart coaches allow the system to give recommendations to the users based on their lifestyle to live a more healthy life. While these data sources show much potential for gathering information about diseases, no application monitoring causes and effects for specific diseases are available for these devices to the date of this thesis.

3.4 BIG DATA IN EPIDEMIOLOGY

Big data is denoted as data that cannot be assessed using standard data processing techniques, because it exceeds the processing capacity of conventional database systems. Often, big data is the result of triangulating measured data from different sources to create extensive data sets. Users are interested in hidden patterns and relationships within the data to explain various phenomenas. The monetization of such data is the foundation of large cooperations, such as Facebook, which provides special target advertising custom-tailored to each user. Online shopping sites, such as Amazon, assess the shopping behavior of their clients to provide better product suggestions. Big data gained so much momentum that the U.S. government announced the Big Data Initiative in 2012, which gave $200 million to associated research projects [274]. In this thesis, population studies in the sense of epidemiological data are *not* seen as big data.

Big data is characterized using the four major aspects:

- **Volume.** Having more data beats having better models is one basic principle of big data analysis. The volume inherently yields one of the largest challenges for computer infrastructures. As the data sets cannot be processed by standard databases, alternative storage and analysis methods have to be applied. Apache Hadoop for example distributes computing problems to multiple connected servers to process the data. Population and cohort study data usually have a vast amount of features from the different assessments described in Section 2.5. The number of subjects, however, is often limited to a couple of thousands due to financial and temporal restrictions. The data sets measure a couple of megabytes. When taking raw sensor information from the medical image data or laboratory equipment into account, the data size rapidly grows into terabyte scale. The use of this raw data for the analysis, however, remains questionable.

- **Velocity.** The information retrieval speed is denoted as velocity. The system has to be capable to react to the new data, process and store it. For example, analyzing geospatial data from smartphones involves

processing multiple position updates per second, depending on the underlying application task. The velocity of population study data is minimal. Cohort study data are usually acquired in cycles of years. Hence, velocity does not apply to this type of data.

- **Variety.** Data is usually complex and unstructured and therefore not ready for processing at arrival. For example, raw feeds from GPS sensor data need to be converted in a proper format to be processed efficiently. Social network relationships may be translated and integrated into graphs. A principle in big data is *when you can, keep everything*, as it potentially contains useful information that can be incorporated in the analysis. In cohort study data, variety is given due to many different data acquisition modalities. The data inflow is carefully monitored by quality control experts to avoid any acquisition bias. This involves manual work, which can be replaced using big data methods.

- **Veracity.** Veracity determines the uncertainty of the data. In other word, data entries or even whole dimensions are described with an uncertainty value. The reasons for this range from either a low quality in the data cleaning step leading to unreliable entries or quantifications, to personal motives of subjects to knowingly or unknowingly lie. This may be due to an ambiguously formulated question or a question regarding personal habits, such as alcohol intake, nutrition or sporting activity.

Efforts are put in conducting epidemiological studies with a large number of participants from multiple data sources [256]. Counter arguments are made that even large-scale cohort studies could lead to false results, which were disproved by randomized trials [38]. Selection bias as well as residual confounding can still be a problem for studies with large attendance. Additionally, it is possible that "extremely large studies may be more likely to find a formally statistical significant difference for a trivial effect that is not really meaningfully different from the null" [119]. To summarize the points why population study data is not yet to be considered as big data:

1. They can be processed by classic statistical approaches and can be stored in standard databases.

2. They comprise a very low velocity.

3. The strictly controlled data acquisition process ensures a very high veracity.

However, visualization techniques applied for big data can, in some cases, be applied to population study data. A scatter plot for over 2,000 subjects may for example suffer from overplotting. Liu et al. [154] propose binned visualizations, where the visualization space is divided into hexagonal or rectangular bins, which are then color-coded depending on the aggregation of the containing subjects (total count, sum, average, min/max). This metaphor can be adapted by employing other existing visualizations that use binning, such as heat maps, choropleth maps, line plots, histograms and bar charts.

Part II

EXTENDING THE EPIDEMIOLOGICAL WORKFLOW WITH INTERACTIVE VISUAL ANALYSIS

Parts of this introduction are based on

> **Paul Klemm**, Steffen Oeltze-Jafra, Kai Lawonn, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Trans. on Visualization and Computer Graphics*, 20(12):1673-1682, 2014.

The epidemiological workflow was characterized together with our clinical experts Katrin Hegenscheid and Henry Völzke. Steffen Oeltze-Jafra, Kai Lawonn and Bernhard Preim provided valuable discussions on how the IVA principles can be adapted to the epidemiological workflow.

This part extends the epidemiological workflow described in Section 2.4 with VA/IVA methods and gives methodological recommendations based on scientific questions on population study data sets.

The epidemiological workflow emphasizes the reproducibility and statistical integrity. Introducing the *IVA* principle to the epidemiological domain aims to compensate the weaknesses of the existing workflow rather than replacing it (Fig. 26). In the current state, the workflow treats the data like a black box. Statistical tests on variables associated to a hypothesis yield a value for deciding whether the data supports the hypothesis. Variables not included in the analysis may potentially support the chosen hypothesis by discriminating the population in the expected way, but are not highlighted. This becomes even more important when the workflow is adapted to the analysis of the medical image data, where domain experts have to identify landmarks tediously to derive measures, such as diameters. This leaves out the majority of information in the image data by abstracting it to single values. Considering all of the available data potentially makes those results more trustworthy and could also identify possible anatomical confounders– arguably an epidemiolgical research result in itself.

IVA tries to illuminate the black box by making the domain experts part of an iterative variable selection process (see Fig. 26 b). Pearce and Merletti [198] pointed out that methods are needed which can cope with this complexity and allow for the search of underlying causes of a certain condition or disease. According to them, "*risk factors for disease do not operate in isolation but occur in a particular population context*". *IVA* also aims to project back into the hypothesis formulation step to amplify hypothesis generation. This has to be handled with care, since *overfitting* of expectations to the data is an imminent danger [263].

While the methods differ depending on the type of the underlying data and the nature of the investigation (detailed further below), the analysis cycle remains the same. The entry point of the cycle, however, depends on the type of the hypothesis and exploration. Figure 27 displays the analysis cycle, which is similar to the *IVA* workflow described in Section 3.2.2. The initial variable listing step, either user-driven or carried out using data mining techniques acts as input for the visualizations. In the sense of Keim's VA Mantra, the visualization can also contain the vast amount of variables as part of an overview visualization to highlight hotspots. The visualization may then lead to various decisions. The user may iterate on the variable selection step, because the expected behavior is not seen in the plot, or additional variables may need to be assessed to check a new hypothesis derived through new insights. This refers to the *Zoom, Filter and Analyze Further* and *Details on Demand* step in Keim's VA Mantra. This also incorporates the introduction of new linked views. Brushing elements yields updates in all other views. This can either be used to inspect how the variables represented by the views are connected or to create detailed groups which can then be an-

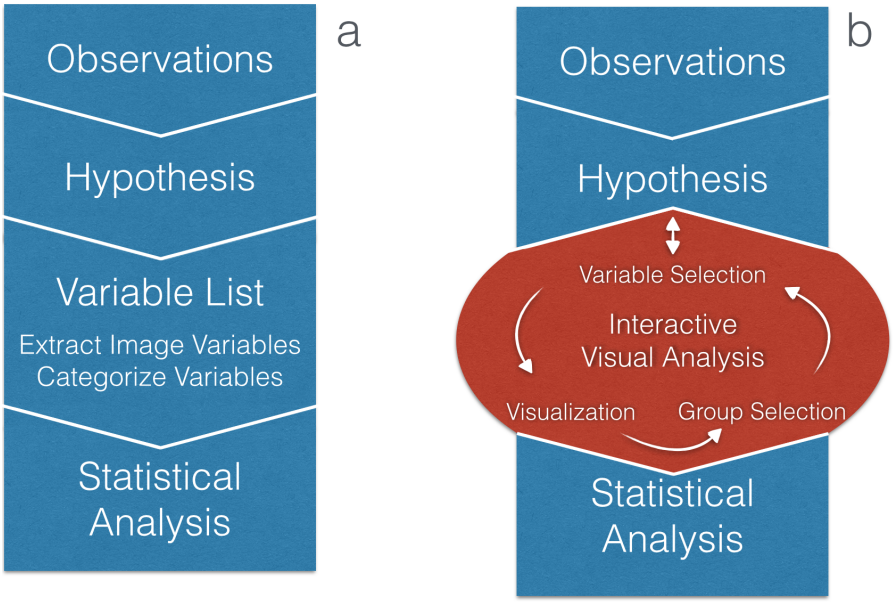


Figure 26: The standard epidemiology workflow consists of four steps (a). *IVA* tools complement parts of this workflow instead of replacing them (b). The combination of statistical and interactive analysis shows promising potential to unveil information in the data. We call the iterative red highlighted part *IVA Loop*, described in detail in Figure 27. Image from [293].

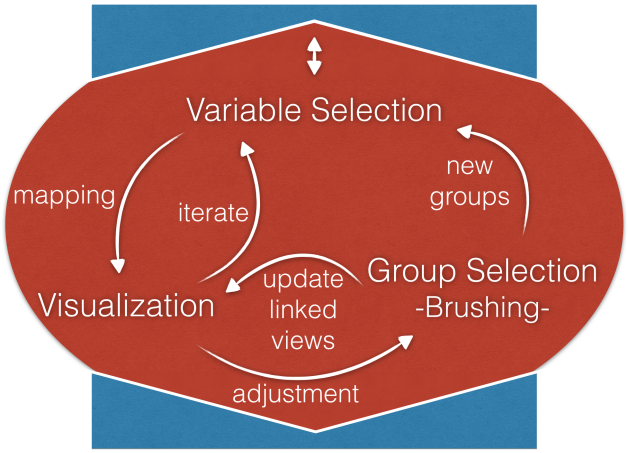


Figure 27: Detailed *IVA Loop* as extension from Figure 26. Usually starting with a selection of a variable of interest (user-driven or via data mining techniques), the data are mapped using a visualization technique appropriate for the selected data types. The data are visualized and can be brushed, yielding new groups to be investigated using further variables. Note that adjacent steps are directly connected via feedback loops, allowing for an iterative refinement and giving as much freedom to the user as possible. Image from [293].

alyzed further. For the latter, the user most likely has a hypothesis about the selected subgroup. To assess the hypothesis, he or she may add new variables to the analysis and observe the behavior of the subgroup. Also, the analysis for different groups can account for confounders, for example when investigating gender or age groups. This step can also be carried out automatically by a data mining algorithm, which yields variables depicting a distribution difference of the group compared to the whole subject set. Note that the workflow does *not* rule out the necessity of conducting statistical analysis to verify the findings. Without these and cross-references with other studies, the finding will not be accepted in the epidemiological community. This is, however, not the focus of this thesis.

As described in Section 3.2.2, *IVA* consists of different analysis levels. Their implementation, however, strongly depends on the underlying data and the analysis type and will be discussed in detail in the following chapters as part of the method designs. The methods proposed in this thesis achieve the fourth *IVA* level of **proprietary analysis**, where applications are custom-built w.r.t. the application domain and underlying questions in mind.

All methods presented in this thesis are based on enhancing the variable listing step, as seen in Figure 26. The methods can be distinguished using two criteria.

1. **Is the analysis hypothesis-free (explorative) or hypothesis-based (confirmative)?** A confirmative approach requires means of quickly selecting the features of interest for the investigated hypothesis. Hence, the variable selection step in Fig. 26 is carried out manually, leading to visualizations which can then be further assessed. This is the closest approach to the classical epidemiological workflow. An explorative analysis requires either an overview visualization over all available features in a data set or a data mining method extracting the important variables. The analysis requires the user to specify a phenotype (target disease or condition) he or she is interested in. An overview visualization can display hotspots correlating with the target. The analysis starts at the *Visualization* step in Figure 27. Data mining algorithms can identify specific risk groups as well as parameters related to the target. Analogous, the analysis starts at the *Variable Selection* step in Figure 27. The first one uses the visual system of the human as pattern recognition method to identify these features, the latter automates this task. Which method is chosen depends on the preference of the user as well as the suitability of the methods to observe patterns of different complexities, such as linear, quadratic or higher dimensional relationships. New questions will arise throughout the analysis process, regardless of how it started. As a result, the borders between explorative and confirmative analysis become more and more blurry as the investigation progresses. The proposed *IVA* workflow is therefore cyclic. Another consequence is the projection of the workflow back into the hypothesis step, as new insights are derived through the analysis, which lead to new questions. A good *IVA* system also reacts to the current analysis phase and proposes useful information to the conducted analysis. If the user, for example looks at a plot of features, the system might suggest features that correlate with the current investigated features, which can raise new questions [293].

2. **Which data types are involved in the analysis?** The involved data types restrict the number of suitable visualization and data mining methods. Most data mining methods, for example, expect categorical variables and do not work well with numerical data (recall Sec-

Table 3: The methods proposed in this thesis are categorized whether they contain spatial medical image data or not. The data incorporated by the methods listed in the second column include parameters derived from medical image data, such as ratios between tissue types or diameters of segmented tissue. The methods analyze these data, but do not treat the image-derived features in any special way.

	Explorative (Hypothesis-Free) Analysis	**Confirmative (Hypothesis-Based) Analysis**
Contains Medical Image Data	Chapter 5: Image-Centric Data Analysis • Segment the image data [290, 291, 296, 297] (Sec. 4.2) • Clustering of segmentation masks [291] (Sec. 4.4) • Shape variance visualization of subject groups [290, 293] (Sec. 4.3, 4.5)	Chapter 5 Image-Centric Data Analysis • 2D information visualizations augmented with 3D image data [293] (Sec. 4.5) • Extract correlations for subgroups [293] (Sec. 4.5) • Automatic shape-based clustering for selected subgroups [291, 293] (Sec. 4.4, 4.5)
Contains No Medical Image Data	Chapter 6: Data-driven Analysis of Sociodemographic, Medical and Lifestyle Factors • Extract Decision Tree Quality Plot [294] (Sec. 5.1) • Retrieve subject groups from clustering algorithms [292] (Sec. 5.2) • Overview visualization using 3D regression Heat map [295] (Sec. 5.4)	The classical epidemiological workflow applies. Support the workflow by modeling hypotheses: • Compile a set of linked visualizations depicting the variables of interest [293] (Sec. 4.5) • Describe the hypothesis using regression notation [295] (Sec. 5.4)

tion 3.1.5). The major distinction made in this thesis is whether the data involves medical image data. As shown in Chapter 4, the concurrent analysis of spatial image data with non-image data requires different approaches with a distinct set of visualization as well as data mining algorithms in contrast to methods without spatial data, as described in Chapter 5.

Almost all scientific publications associated with this thesis propose methods for both explorative *and* confirmative analysis approaches. According to the criteria described above, the methods proposed in this thesis can be distinguished as follows. Table 3 categorizes the methods proposed in this thesis according to the analysis approach and the involvement of spatial medical image data. It is worth noting that *all* publications associated with this thesis comprise medical image features. Chapter 4: Image-Centric Data Analysis focuses on publications, where medical image data is processed directly without converting it into numerical or categorical features [291, 293, 296, 297]. The focus of these methods is the concurrent visualization and analysis of spatial image data with non-image data. They tackle one major task, which epidemiologists could not solve until now–the calculation of mean shapes and models of specific tissues w.r.t. other variables. For example, what defines the mean shape of the spine of males and females and which other parameter influence it? What are the differences of this shape to the global average shape or to other shape classes?

Chapter 5 includes features derived from medical image data. This is a result of the focus of the clinical partners providing the data sets. The methods presented in Chapter 5 [292, 290, 295], however, do not treat those features in any special way. They might as well be data derived from other examinations. The image-derived features often constitute the target phenotype of the data. If, for example, the phenotype is the mean curvature of the spine, the analysis aims to find features correlating with the curvature. Epidemiologists already comprise a rich set of statistical methods to analyze data when they already have a hypothesis about the data and the data is

only of categorical or numerical type. In the proposed data-driven analysis approach, this analysis can be supported by offering the domain experts with an IVA tool of multiple linked views, which may lead to new insights and new questions. One example of such a hypothesis would be: Smoking habits are associated with back pain. The user then moves on to investigate the feature indicating whether the subject smoked as well as the phenotype feature back pain. Hence, the IVA system displays a suitable information visualization for this relationship, such as a mosaic plot (since both variables are ordinal). Maybe there are multiple features indicating the smoking behavior of a subject, such as *age when started smoking*, *number of years smoking* and *number of cigarets per day*. A plot matrix of the selected features allows for assessing their influence with the investigated phenotype. In these situations, visualizations are superior to classic statistical analysis, because the information is presented in a cognitively feasible way. The statistical analysis, however, still has to be carried out using dedicated statistical processors, such as R, SPSS or STATA, to quantify the hypotheses.

The following chapter covers the methods incorporating spatial medical image data.

4 IMAGE-CENTRIC DATA ANALYSIS

The introduction of this chapter is based on the VAST'14 publication, Section 1 "Introduction" [293] as well as Section 4.1. "Medical Image Analysis" of Preim et al. [296]. Large-scale population studies often include medical image data. The concurrent analysis of image data and non-spatial epidemiological factors requires techniques that reach beyond standard statistical methods. For instance, segmentation of the image data is required for an analysis of anatomical structure and of possible correlations between this structure and epidemiological factors. Semi-automatic segmentation techniques are promising but also challenging, since the employed modalities, such as magnetic resonance imaging (MRI) and ultrasound, are subject to inhomogeneity and noise.

Compiling a list of features for tests of statistical resilience based on experience-driven hypotheses leaves out other features in the data which potentially interact with a disease. This also applies to the chosen landmarks used to quantify medical image data information. Also, only a small subset of features can be concurrently analyzed. The standard workflow lacks methods for automatically identifying correlations possibly buried deep in the data or overseen by the expert. Also, only a small subset of factors can be concurrently analyzed.

One major purpose of incorporating medical image data in population studies is to *quantify* the underlying anatomy, e.g., using volumes, diameters, or spatial relations. Quantifications are of special interest for diagnostics, as mean values can be established for each patient derived from the mean of *similar* subjects. Similarity can be defined using age groups or gender, but also based on the influencing factor of the investigated condition. For example, the liver fat value of an alcoholic patient could be compared to other subjects with similar conditions to assess the disease severity. Additionally, variations can be associated with pathologies. For example, it may be detected which spine shapes are associated with strong back pain. MRI data is of special interest for epidemiologists, since it does not emit harming radiation to the patient and still allows whole body scans [296].

4.1 THE SPINE DATA SET

The methods proposed in this chapter are applied to a spine data set compiled to analyze *lower* back pain. Therefore, this section aims to give an overview of the disease as well as the data set.

LUMBAR BACK PAIN Back pain is one of the most common diseases in the Western civilization [267]. It is focused on the lumbar spine, as seen in Figure 28. While the understanding of genetic mutations regarding back disorders made progress, the correlations with different environmental factors as well as physical stress are not sufficiently understood. There are no correlations between degenerative changes of the intervertebral disc and adjacent vertebrae [69]. The main aging effect of the spine is the reduction of bone minerals and the development of degenerative diseases (osteoporosis) [194]. This yields a loss of height for the vertebral body, resulting in a dented surface, which may lead to herniated discs. The elasticity of the intervertebral discs is reduced with age. Lordosis defines the inward angle of the spine, which can be seen as the typical 'S' shape when seeing the subject from the

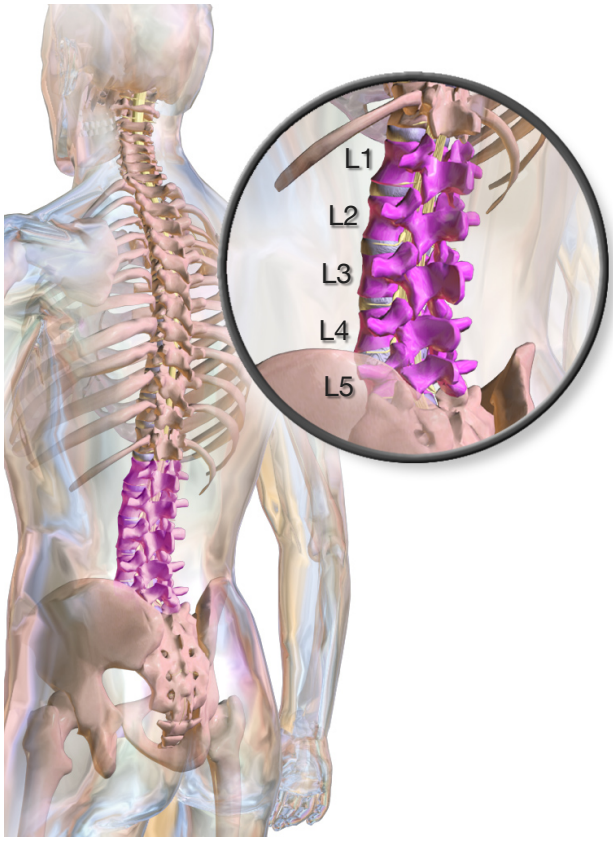


Figure 28: The highlighted lumbar spine consists of 5 bones. They are denoted as L1 starting with the top vertebra and ending with the L5 vertebra. Between the vertebrae lie the intervertebral discs, which buffer the movement-induced positional changes of the vertebrae. Lower back pain, the most common back pain type, is related to the structural changes of the lumbar spine. No clear single cause of lumbar back pain is known [35]. It is believed to be a result of skeletal and muscle issues, such as sprains or strains. Other risk factors comprise obesity, smoking, weight gain through pregnancy, stress, low physical activity, poor posture or sleeping position [238]. Possible physical causes include osteoarthritis, vertebra disc degeneration, broken vertebrae or spinal disc herniation [13]. Due to changes in posture and center of gravity, pregnancy shows a strong correlation with lumbar back pain, as nearly 50% of pregnant women report lumbar back pain [163]. The image is provided by the blausen.com staff under the Creative Commons Attribution-ShareAlike License [245].

side (sagittal plane). Scoliosis is a c-shaped deviation of the spine when the subject is viewed facing towards the observer (coronal plane). The spinal alignment and its shape are associated with lumbar back pain. The shape also seems to be influenced by age [230].

Manek et al. [166] reviewed the progress made in understanding causes of back pain and present influencing factors like age, gender, weight and different lifestyle aspects, such as smoking behavior and work conditions. Tucer et al. [258] conclude that depression is one of the independent risk factors for experiencing low back pain, although their analysis is based on surveys of the subjects and does not rest upon clinical analysis. Van Tulder et al. [267] conclude that the value of such identified risk factors as prognostic value remains low. No factor arose as strong indication for back pain through many different studies. Epidemiological analysis of lumbar back pain, such as the work of Harreby et al. [95], is largely focused on non-image information. Harreby et al. identified a risk group of female subjects

who smoke and have heavy jobs with a 46% probability of back pain. In comparable studies, only a few shape-related variables are included [147].

Lang-Tapia et al. [147] used a non-invasive method for analyzing spine curvature using a so-called "SpinalMouse". They correlated spine curvature with age, gender and body weight. They did not observe correlations between lumbar spine deformation and body weight.

Determining risk factors in this area can lead to [74]:

- a better understanding of effects of preventive measures, such as occupational health and safety regulations,
- prognostic features for diagnosis and treatment of lumbar back pain, and
- determination of particularly affected risk groups.

Characterizing the healthy aging process of the spine is a long-term goal for determining age-normalized probabilities for spine-related diseases by incorporating individual risk factors. As described above, however, no single factor for strong back pain arose throughout many studies.

THE DATA SET The data set comprises 127 features describing diagnosed diseases, lifestyle factors, women-specific factors, pain indicators, laboratory values and somatometric features for 6,753 subjects (4,420 from `SHIP-Trend-0` and 2,333 from `SHIP-2`). Since data acquisition protocols between these two cohorts are identical, the features between the two cohorts are comparable. The data contains 30 metric, 7 nominal, 29 ordinal and 62 dichotomous features. Somatometric features include measures of the human body, such as body height, weight and body fat percentage as well as gender. These measures are reliable and complete. Other features, such as pain indicators or lifestyle indicators (e.g., physical activity) are more subjective and less reliable. There are also features missing for each subject, such as features building upon each other (e.g., "Do you have high blood pressure? Which medication is prescribed against it?"). Therefore, some manifestations are sparsely populated, which makes statistical evaluation challenging.

The image data was acquired for each subject on a 1.5 Tesla scanner (Magnetom Avanto; Siemens Medical Solutions, Erlangen, Germany) by four trained technicians in a standardized way. The spine protocol consisted of a sagittal T1-weighted turbo-spin-echo sequence ($1.1 \times 1.1 \times 4.0$ mm voxels) [101]. The lumbar spine was detected in the image data using a hierarchical finite element method by Rak et al. [213]. The tetrahedron-based Finite Element Model (FEM) is initialized with three clicks on a vertebra – the center to initialize the position, as well as the top and bottom to determine the rough height of the model. Initial rough segmentations are then refined with a model-driven segmentation to finalize the data-driven correction step. The model uses a weighted sum of T1- and T2-weighted MR images to detect the lumbar spine shape. Once registered, it captures information about the shape of the lumbar spine canal as well as the position of the L1-L5 vertebrae. Due to incorrect initialization, strongly deformed spines, contrast differences and artifacts, the model was not able to detect lumbar spines for all subjects. Therefore, 2,540 tetrahedron models of the lumbar spine were obtained.

The epidemiologists are interested in the influence of the shape of the spine and its influence on back pain. Which deformation levels are healthy and which may indicate pathologies? Additionally, they are interested in other factors, which may influence back pain together with the shape of the spine. These factors include smoking behavior, nutrition, heavy physical work or medications. Characterizing the boundaries of healthy shape

changes of the spine shape is of great interest for epidemiologists. It allows them to infer pathological shape ranges associated with back pain and potential diseases. The vision of epidemiologists is a set of clinically measurable data, which is then compared to reference values to indicate pathologies.

4.2 IMAGE SEGMENTATION

The methods proposed in this thesis are not designed to help image segmentation for population study data. The foundations are, however, discussed here briefly, since the segmentation is part of the workflow and the foundation for the methods presented in this chapter.

The quality of medical image data assessed in the context of a population study data is often inferior to the clinical standards due to cost and time factors. Therefore, many standard tools for image quantification will not work for the data or yield inaccurate results. Therefore, epidemiologists often segment medical image data derived from population studies by hand, which is a time-consuming and tedious work. The fact that this effort and associated costs are accepted underlines the scientific importance of the results. Manual segmentations are acceptable in clinical settings, such as radiation treatment planning. They are, however, not suitable for large-scale population studies. The inter and intra-expert variability of segmenting the data is too high. Also, for cohort studies, which comprise multiple waves, follow-up data has to be assessed.

In order to make the results comparable, the segmentation has to be carried out in exactly the same conditions with the same experts, or it has to be done again from scratch. Segmentations are still carried out by hand because of the poor availability of standardized segmentation methods suitable for a wide range of image modalities and settings. Hence, most image segmentation algorithms are custom-tailored solutions based on the underlying image acquisition sequence and the structure of interest for the expert. This often includes the fine tuning of numerous parameters of different image filters and pattern recognition algorithms to the data, which is usually carried out by a computer scientist. Even after that, segmentation algorithms still require user input, such as setting landmarks to initialize the underlying model or to correct segmentation errors.

One example are the FEMs of Rak et al. [213], which are used to detect the lumbar spine shape in MRI scans as described above. Fully automatic methods, such as the liver segmentation of Gloger et al. [80] often follow a set of concurrent detection and localization steps. The approach can be reused and recombined for other applications, such as MRI data of the kidney [81]. For more details on image segmentation algorithms in the epidemiological application domain, see the survey paper of Tönnies et al. [297]

DATA STORAGE The non-image data is stored and accessed as ASCII-encoded text files in open formats, such as `JSON` or `CSV`. Other scientists can use the methods proposed in this thesis with their own data without having to convert the data into a specific database scheme. Major advantages of databases comprise sophisticated methods for structuring the data and provide fast subject filtering using queries. The statistical language `R`, which is an important tool in this thesis, already has libraries optimized for fast access, such as `dplyr` [276]. Fast and efficient filtering libraries are also available for Javascript, such as Crossfilter [117]. Therefore, fast filtering can be achieved in both languages without utilizing a database back end. The major performance bottlenecks are the complex calculations in the various analytics methods incorporated in this thesis. Hence, no database systems

are applied. A database should be employed when data filtering steps become a performance issue. Incorporating a suitable database scheme also allows the combination of different data sources, such as observed and simulated data.

4.3 EXAMPLE OF IMAGE SEGMENTATION BY DISSIMILARITY ANALYSIS USING SHAPE DEFORMATION MODELS

This section is based on

> **Paul Klemm**, Steffen Oeltze, Katrin Hegenscheid, Henry Völzke, Klaus D. Tönnies, and Bernhard Preim. Visualization and exploration of shape variance for the analysis of cohort study data. In *Proc. of the Vision, Modeling, and Visualization Workshop*, pages 221-222, 2012.

The image analysis workflow was developed jointly with Klaus D. Tönnies, Bernhard Preim and Steffen Oeltze. Katrin Hegenscheid and Henry Völzke provided the technical details on the epidemiological workflow and associated hypotheses and problems. Steffen Oeltze helped with questions regarding the MeVisLab [218] related implementation issues.

At the beginning of the work covered in this thesis, no image segmentation results were available for the given data sets. Therefore, approaches for deriving the image data as well as creating shape variance visualizations were investigated. The results are described in this section, which presents two methods for creating data structures suitable for shape variance analysis and provides suggestions for their visualization. A pipeline for analyzing shape

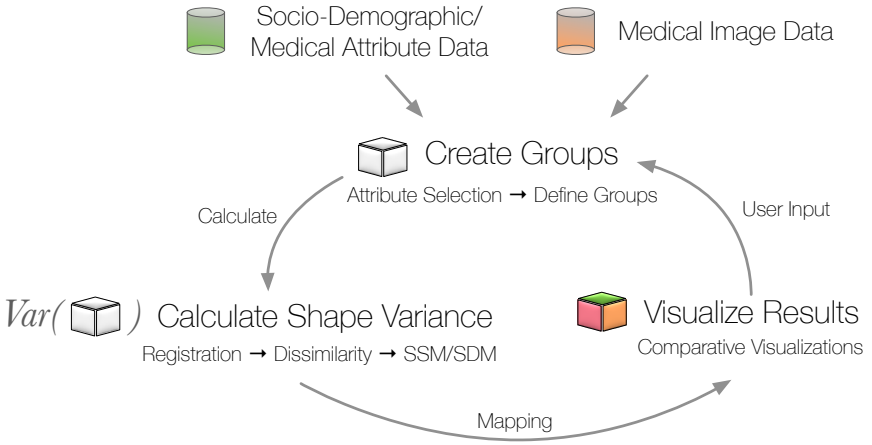


Figure 29: Workflow for the shape variance analysis of population study data. As a first step of the shape variance analysis, the user can define population groups using the different data types. Alternatively, this can be carried out automatically using data mining algorithms. With the selected groups the shape visualization model is then calculated as basis for the following visualization comparing the structures to reference groups. The visualization-derived insights directly influence the user-supported hypothesis generation.

variance population study data is shown in Figure 29. It is derived from the IVA analysis workflow, but displays the necessary steps for analyzing image data using shape deformation models based on dissimilarity. This does not only involve segmenting the tissue in each data set, but also requires correspondences between each segmentation instance. The latter is required for determining differences between subjects and groups. The shape variance

analysis step comprises two different approaches toward creating structures that allow for the visualization of inter-object differences. They arise from different requirements given by the underlying image data and the shape of the structure of interest.

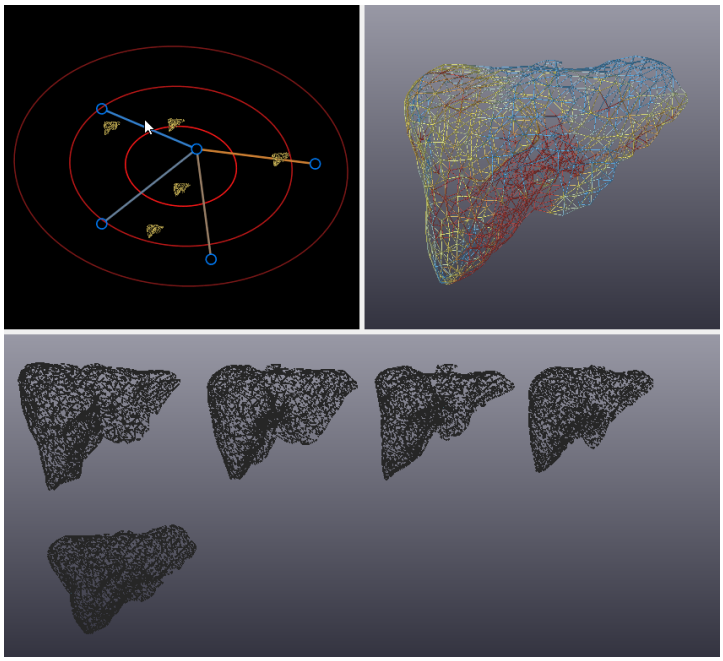


Figure 30: Visualization of five liver meshes generated from `GAMEs` algorithm using `ShapeSpaceExplorer` from Busking et al. [29]. The currently selected shape in the space (cursor-icon) is mapped via mesh morphing on the reference mesh. The color scale indicates the amount of deformation to the underlying reference.

ESTABLISH POINT CORRESPONDENCES FOR EXISTING SEGMENTATION MASKS If segmentation masks already exist for the structure of interest, the problem remains to establish inter-subject point correspondences. The growing and adaptive meshes (`GAMEs`) algorithm of Ferrarini et al. [70] allows to create shape distribution models of such masks. Prior to this step, however, the segmented structures have to be registered onto each other. This is carried out using the `elastix` toolbox [133]. The GAMEs algorithm yields a surface mesh for each segmentation instance, where the mesh points correspond between all instances. This structure can be used to calculate a *shape distribution model*, which can be visualized using the `ShapeSpaceExplorer` tool from Busking et al. [29] (see Section 3.1.3). The tool requires one of the created meshes as the basis for the mesh morphing algorithm which interpolates between the different volumes (see Figure 30). This space allows to navigate the different shapes, while the object space is a morphed representation of the currently selected object in shape space, showing the amount of local deformation to the reference volume.

DERIVE COMPARABLE VOLUME DATA USING NON-RIGID REGISTRATION *Statistical Deformation Models* (SDMs) capture deformation a model of the whole volume instead of only the surface (recall Section 3.1.3). SDMs are derived from MRI scans of the spine. Cuboid blocks containing the single vertebrae were manually cut out of a reference data set. Then, the reference models were aligned on each data set using affine registration of the elastix toolbox (B-spline grid size of 16 voxels and 500 gradient descent iterations).

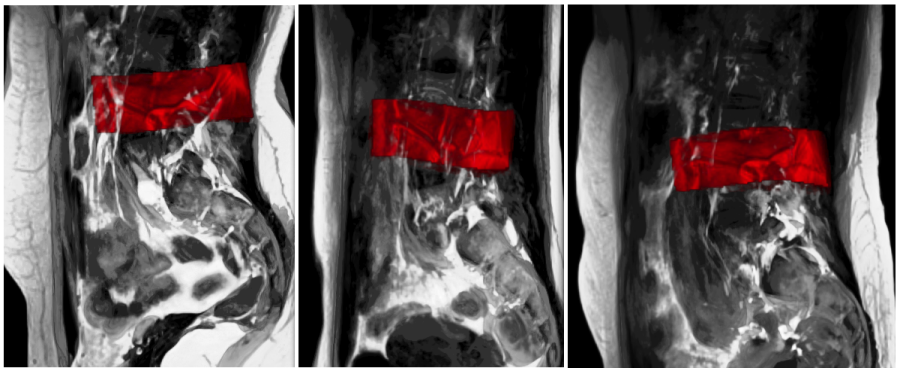


Figure 31: Sagittal MR images of the lumbar spine for three different subjects. The red pictured reference cuboid block of the L4 vertebra was first aligned via affine registration to the counterpart in the data set and then deformed using B-spline registration.

Since always the same reference model is deformed, the voxels of the resulting models are comparable (Fig. 31).

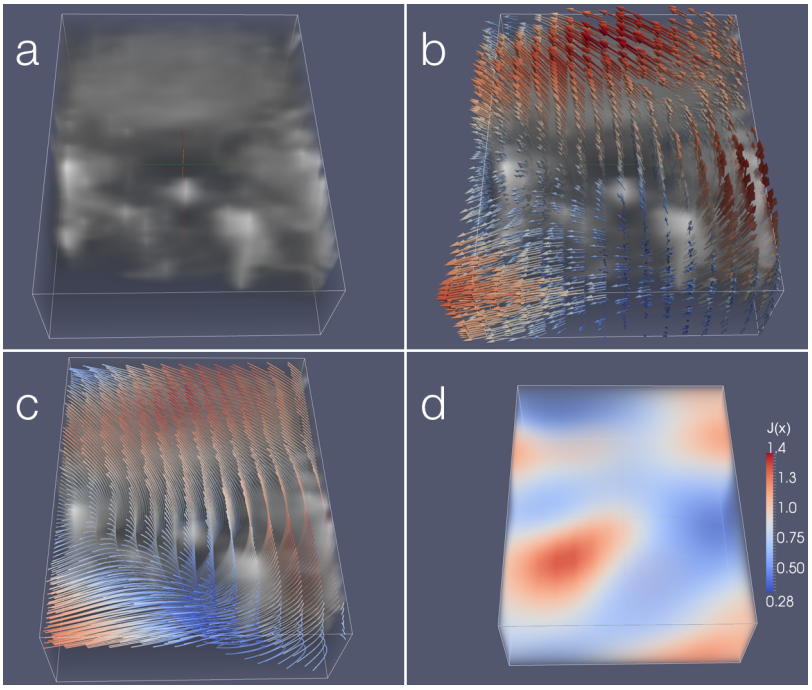


Figure 32: Four different visualization techniques for a deformation field of a L4 vertebra. (a) Visualization of a B-spline registered L4 vertebra. (b) Deformation field rendered with glyphs and (c) streamlines. (d) Visualization of the Jacobian determinant. Figure adapted from [290].

The resulting data model can be visualized using different techniques, such as glyphs and streamlines, as presented in Figure 32 (b) and (c). The Jacobian determinant of the deformation field which describes the local expansion and compression can be seen in Figure 32 (d). Another possibility is, similar to the `ShapeSpaceExplorer` approach, the deformation of a reference object using morphing algorithms as the user specifies its parameters. As it becomes clear which deformation is associated with a disease, quantifying metrics for these deformations can be derived. These metrics can then be assessed using the standard epidemiological pipeline.

The following section describes a system which employs SDMs to create sub-

ject groups based the shape of their lumbar spine canal. Section 4.5 shows how SDM data is visualized concurrently with non-image data.

4.4 EXAMPLE OF IMAGE SEGMENTATION AND PROCESSING ON LUMBAR SPINE VARIABILITY

This section is based on

> **Paul Klemm**, Kai Lawonn, Marko Rak, Bernhard Preim, Klaus D. Tönnies, Katrin Hegenscheid, Henry Völzke, and Steffen Oeltze. Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data. In *Proc. of the Vision, Modeling, and Visualization Workshop*, pages 121-128, 2013.

Katrin Hegenscheid and Henry Völzke provided the underlying data as well as the domain knowledge regarding lumbar back pain and potential associated risk factors. They also provided the medical knowledge required for the evaluation of the results. Marko Rak and Klaus D. Tönnies provided the tetrahedron-based segmentation models of the lumbar spine, which are the foundation for further analyses. Kai Lawonn helped with abstracting the tetrahedron models to line segment representation and the underlying MATLAB[1] implementation. Steffen Oeltze provided the clustering algorithm used for the shape-based grouping of subjects. He also provided major contributions to the VTK-based implementation of the group visualizations. The technical details were developed and discussed in detail together with Bernhard Preim, Kai Lawonn and Steffen Oeltze.

The methods in this section show how information from medical image data can be extracted and incorporated with non-image parameters. The Interactive Visual Analysis presented in Section 4.5 incorporates these information and enhances them.

In this section, an approach for the reproducible analysis of the lumbar spine canal variability in a population is proposed. It is based on the centerline of each individual canal, which is derived from a semi-automatic, model-based detection of the lumbar spine. The centerlines are clustered to form groups with low intra-group and high inter-group shape variability. The clusters are visualized by means of representatives to reduce visual clutter and simplify a comparison between subgroups of the population. Special care is taken to convey the shape of the spinal canal also orthogonal to the view plane. The approach is demonstrated for 490 individuals drawn from the SHIP data presented in Section 4.1. The reason why it was only performed on a subset of the data is that not all data was available at that time. Also, the automatic detection failed for several subjects due to wrong initializations and failed preprocessing steps. Preliminary results of investigating the clusters with respect to their associated socio-demographic and biological factors are presented. The contributions are:

- generation of groups of individuals sharing a similar shape of the lumbar spine canal,
- visualization of these groups by means of representatives,
- illustration of 3D shape in a 2D view.

While the processing of the 490 data sets represents first results, expected behavior like decreasing spine curvature with increasing subject body height was observed. Unexpected clusters of unusual shape, which are now subject to further epidemiological analysis, were found.

1 Owned by The MathWorks, mathworks.com

4.4.1 *Detection of the Lumbar Spine*

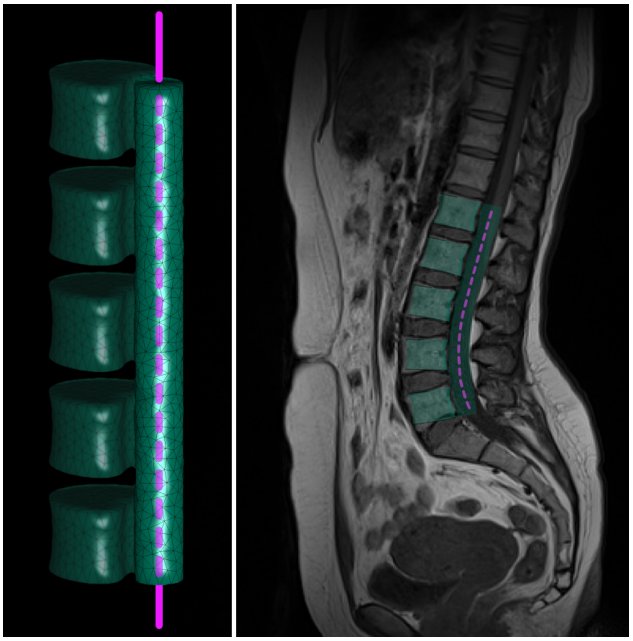


Figure 33: The layered finite element model consists of more than 2,000 tetrahedrons (left). The spine canal center line is indicated by the dashed line. The model uses the image-induced potential field to align itself to find a local minimum after the initialization (right). Image from [291].

A hierarchical finite element method according to Rak et al. [213] was applied to create the detection mask. FEMs of vertebrae and spinal canal are connected by a bar-shaped FEM (Fig. 33). The model comprises a fixed number of points which are pairwise relatable between instances of the model. Hence, correspondences between lumbar spine representations of different data sets can easily be established. The model is placed in the scene using an empirically chosen initialization point. The force acting on the model stems from aggregation of loads, which are derived from a potential field resulting from a weighted sum of the T1- and T2-weighted MRI data, see Rak et al. [213]. After detecting all spines, the models are registered using the Kabsch Algorithm [125], which is designed to minimize the root mean squared deviation between paired sets of points. The weights are gathered empirically, where the vertebrae appear as dark spots (and local minimum). For the detection of the spinal canal the images need to be smoothed. The model-based detection captures information about the spine canal curvature as well as the alignment of the vertebrae. It is not meant to capture information about vertebrae deformation and differences in spine canal extent.

4.4.2 *Analysis of Lumbar Spine Canal Variability*

The variability of the lumbar spine canal is investigated based on the deformed and registered models of the detection step. Since the primary interest is on the curvature of the spine, the analysis focus lies on the spinal canal. Centerlines capture curvature and are simpler to handle than the tetrahedral mesh. Agglomerative hierarchical clustering is carried out to form groups that exhibit low intra-group and high inter-group shape variability. The clusters are visualized by means of representatives to reduce visual clutter and simplify a comparison between groups of the population.

Centerline Extraction

The centerline extraction was carried out together with Kai Lawonn. In this subsection, the computation of the centerline $c_{\mathcal{S}}$ of the lumbar spine model $\mathcal{S}$ is described. The model is given as a cylindrically shaped tetrahedral mesh. The axis of rotation is aligned to the z-axis. Therefore, the parametric curve $c(t) = p_0 + t \cdot v_z$ is used, where the z-component lies in $[h_{min}, h_{max}]$. Here, h_{min} and h_{max} are the minimal and maximal height of the mesh, respectively. The parametric curve $c(t)$ can be written as:

$$c(t) = \underbrace{\begin{pmatrix} 0 \\ 0 \\ h_{min} \end{pmatrix}}_{p_0} + t \cdot \underbrace{\begin{pmatrix} 0 \\ 0 \\ h_{max} - h_{min} \end{pmatrix}}_{v_z}, \qquad t \in [0, 1]. \tag{4}$$

The intersection points of the parametric curve with the faces of the tetrahedra $\tau \in \mathcal{S}$ are determined from the undeformed lumbar spine model $\mathcal{S}_0$. Thus, the vertices are combined to obtain the triangles and faces and assess the intersection points with the curve. For this, the vertices v_0, v_1, v_2, v_3 of every tetrahedra $\tau = \{v_0, v_1, v_2, v_3\}$ are incorporated and solve the following matrix equation:

$$\begin{pmatrix} v_k & v_l & v_m & v_z \\ 1 & 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ -t \end{pmatrix} = \begin{pmatrix} p_0 \\ 1 \end{pmatrix}, \tag{5}$$

with different permutated $k, l, m \in \{0, 1, 2, 3\}$ for the four faces of the tetrahedra. The equation combines the parametric curve with the triangle face according to barycentric coordinates to obtain the intersection point. If a positive solution $\alpha, \beta, \gamma > 0$ is obtained, the considered curve point lies in the interior of a triangle of τ. Thus, the corresponding tetrahedron with its triangle and their barycentric coordinates is assigned to the curve point $p_i = p_0 + t \cdot v_z$. If one curve point lies on the boundary of a triangle, i.e., one of the coordinates is equal to zero, only one tetrahedron to the curve point is assigned. Using these values, the centerline of every deformed lumbar spine model is obtained by applying the stored barycentric coordinates to the corresponding tetrahedron. Having one intersection point p_i of the undeformed lumbar spine model with the assigned tetrahedra τ, the corresponding triangle face v_k, v_l, v_m, and the assigned barycentric coordinates α, β, γ, the new point p_i' is extracted of the deformed lumbar spine model by applying:

$$p_i' = \alpha v_k + \beta v_l + \gamma v_m. \tag{6}$$

Hence, the new centerline is derived.

Centerline Clustering

The centerline clustering was carried out by Steffen Oeltze-Jafra.

To cluster the centerlines, an agglomerative hierarchical clustering (AHC) approach is employed. It has been demonstrated that AHC delivers meaningful results in the clustering of other plane and space curves, such as fiber tracts from Diffusion Tensor Imaging (DTI) data [176], streamlines from flow data [283], and brain activation curves (time series) from functional Magnetic Resonance Imaging (fMRI) data [153]. Furthermore, it is flexible with

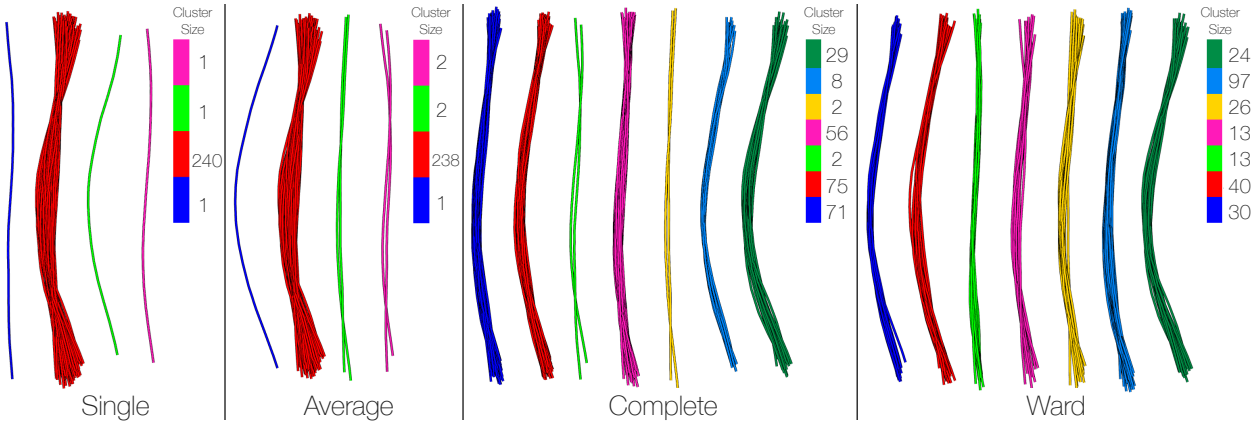


Figure 34: Spinal canal centerlines of 242 female subjects clustered with agglomerative hierarchical clustering using four different proximity measures and a technique for automatically computing the cluster count. Single link and average link suffer from the chaining effect (single large cluster), complete link produces compact, tightly bound clusters and Ward's method is biased towards generating clusters of similar size. The difference in centerline shape also occurs orthogonal to the view plane. Image adapted from [291].

respect to cluster shape and size (non-convex clusters are possible). AHC relies on the difference/similarity between data entities. Thus, a definition of centerline similarity is the prerequisite for AHC of centerlines.

Similarity is often evaluated by a distance measure. General requirements for such a measure are positive definiteness and symmetry. An example that has been successfully employed for clustering fiber tracts and streamlines [176, 283] is the *mean of closest point distances* (MCPD) proposed in [48]. For two centerlines c_i and c_j with points p, the MCPD is computed as:

$$d_M(c_i, c_j) = \mathrm{mean}(d_m(c_i, c_j), d_m(c_j, c_i)) \tag{7}$$
$$\text{with } d_m(c_i, c_j) = \mathrm{mean}_{p_l \in c_i} \min_{p_k \in c_j} \|p_k - p_l\|$$

CLUSTER PROXIMITY In advance, AHC requires the computation of all pairwise centerline distances and their storage in a quadratic and symmetric distance matrix $\mathbf{M}$. The algorithm operates in a bottom-up manner. Initially, each centerline is considered as a separate cluster. The algorithm then iteratively merges the two closest clusters until a single cluster remains. The merge step relies on $\mathbf{M}$ and a measure of cluster proximity. Various cluster proximity measures have been published, among which *single link*, *complete link*, *average link* and *Ward's method* [193] are the most popular ones. In single link, the proximity of two clusters is defined as the minimum distance between any two centerlines in the different clusters. Complete and average link employ the maximum and the average of these distances, respectively. Ward's method aims at minimizing the total within-cluster variance at each iteration. It defines the proximity of two clusters as the sum of squared distances between any two centerlines in the different clusters (SSE: sum of squared errors). The focus lies on automatically computing a reasonable number of clusters k before elaborating on the most suitable proximity measure for the epidemiological application. This computation helps in providing a good initial visual summary of the variants in spinal canal shape and facilitates a reproducible analysis.

NUMBER OF CLUSTERS Salvador and Chan propose a method for automatically computing the number of clusters in hierarchical clustering algorithms [227]. Their *L-method* is based on determining the *knee/elbow*, i.e., the point of maximum curvature, in a graph that opposes the number of clusters and a cluster evaluation metric. The knee is detected by finding the two regression lines that best fit the evaluation graph, and then, the number of clusters that is closest to their point of intersection is returned. Locating the knee depends on the shape of the graph, which again depends on the number of tested cluster numbers k. Salvador and Chan recommend using a full evaluation graph, which ranges from two clusters to the number of data entities. Starting with the full graph, the L-method is carried out iteratively on a decreasing focus region until the current knee location is equal to or larger than the previous location. As evaluation metric, the proximity measure used by the different link versions of AHC is applied. Furthermore, the evaluation is not based on the entire dataset but only on the two clusters that are involved in the current merge step.

EVALUATION OF CLUSTER PROXIMITY MEASURES In an informal evaluation based on 16 datasets, the AHC was tested with the four proximity measures and the L-method. The 16 datasets represent the complete set of centerlines ($n = 490$) and epidemiologically relevant subsets derived according to gender, age, e.g., 20-40, 41-60 and 61-80, body weight and body height. For each dataset, the four proximity measures are applied and all clustering results are visualized side-by-side. A visual inspection of the results confirmed textbook knowledge with regard to the strengths and weaknesses of the proximity measures [193] (Fig. 34 shows an exemplary scenario).

In single link clustering, the *chaining effect* could be observed for every dataset. Here, a single large cluster arises containing almost the entire set of centerlines. This cluster contains very dissimilar centerlines but they are connected by a chain of similar ones via some transitive relationship. For the majority of datasets, average link failed to avoid this effect. Instead, strong outliers were represented as individual clusters while the remaining centerlines, being dissimilar and still comprising outliers, were grouped in a single large cluster. Complete link clustering produced small, compact, and tightly bound clusters. Ward's method was biased towards generating clusters with similar size. These clusters showed less diversity than the ones generated by means of complete link. In summary, due to the chaining effect of single link and average link, and the arbitrary assumption of similar cluster sizes in Ward's method, the complete link is favored as a proximity measure.

The bottleneck of AHC in terms of time complexity is the computation of $\mathbf{M}$, in particular when a multitude of closest point distances must be calculated (Eq. 7). However, the total number of centerlines ($n = 490$) and the number of vertices per centerline ($v = 93$) are relatively small. Furthermore, the computation has been parallelized and the matrix must be computed only once and may be stored. The computation of $\mathbf{M}$ based on the complete set of centerlines, i.e. the entire population, can be considered as the worst case. On a 3.07 GHz Intel 8-core PC with 8 GB RAM and a 64 bit Windows operating system, the computation took 7.9 s. The L-method for determining the number of clusters took 24.2 s and represents the bottleneck in processing the data. This is due to the multitude of computations required for finding the two best fit regression lines but may be mitigated by cutting off unlikely high numbers of clusters from the full evaluation graph [227].

The clustering implementation is based on the AHC algorithm and the proximity measures being part of MATLAB's Statistics Toolbox (MathWorks, Natick, MA, U.S.). The source code of the L-method is provided by A. Zagouras as part of MATLAB Central's file exchange [284].

Visualization of Clustered Centerlines

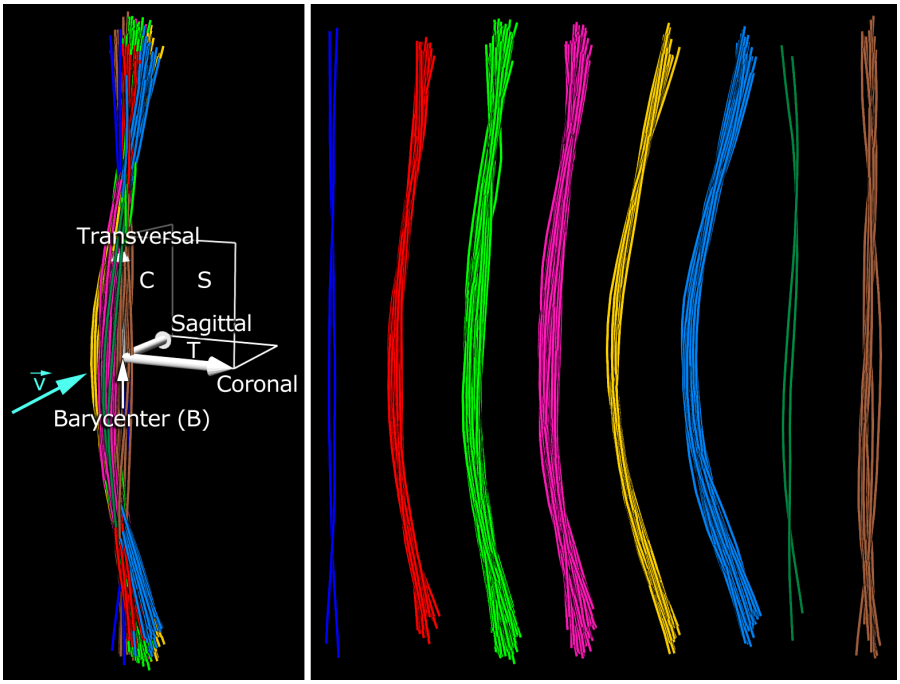


Figure 35: Visualization of the hierarchical agglomerative clustering results. Initially, all centerline clusters are closely intertwined (left). To simplify their interpretation, they are translated along the coronal axis and lined up at equidistant locations (right). The annotations illustrate typical medical view planes/axes: sagittal (S), coronal (C), and transversal (T). The default viewing direction $\vec{v}$ is parallel to the sagittal axis (as can be seen in the right view). Image from [291].

A standard medical view for inspecting the spine in MR images is the sagittal view with the vertebrae located to the left of the spinal canal (Fig. 33, right). Hence, it is chosen as the default view for the presentation of the clustering results. Initially, all centerlines and hence also the clusters are closely intertwined in space due to the co-registration of all spine detection results (Fig. 35, left). In order to get a better overview of the individual clusters, they are translated along the coronal axis and lined up at equidistant locations (Fig. 35, right). The centerlines are visualized with GPU support as illuminated streamlines with halos [68]. The halos improve the visual separation of individual lines. Before the centerlines are translated, the barycenter B of the entire bundle of lines is computed (Fig. 35, left). It will be used for positioning visual hints in the scene.

CLUSTER REPRESENTATIVES In order to simplify the interpretation of a cluster, to further reduce visual clutter, and to improve a visual comparison of clustering results between groups, e.g., younger and elder subjects, a representative centerline for each cluster is computed. This is inspired by the computation of a representative fiber tract for a bundle of fibers derived from DTI trajectography data [26]. Here, the fiber with the smallest sum of distances to all other fibers, i.e., the centroid fiber, of the bundle is chosen. Since all pairwise centerline distances are stored in $\mathbf{M}$, the selection of a centroid centerline is straightforward. Each such centroid is then visualized by a ribbon whose width is scaled according to the size of the corresponding cluster (Fig. 36). Please note that the location of the vertebrae corresponding to this centroid centerline is intentionally not indicated since the ribbons are

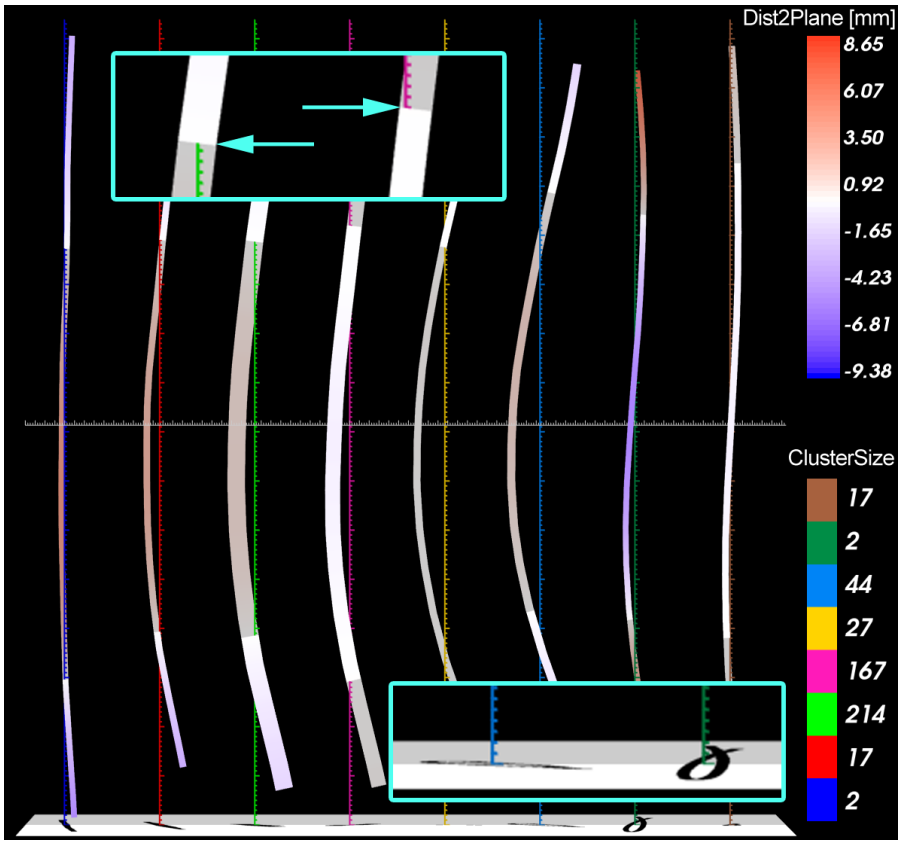


Figure 36: Spinal canal centerlines of all subjects ($n = 490$) clustered with agglomerative hierarchical clustering employing complete link. For each cluster, a representative centerline is visualized as a ribbon. Ribbon width encodes cluster size. Ribbon color encodes the distance to a view-aligned, highly transparent, sagittal plane passing through the barycenter B of the original centerline bundle (Fig. 35, left). The sequence of a ribbon's intersection with the plane supports an assessment of its curvature (upper inset). Shadow projections reveal how far a representative extends to either side of the plane (lower inset). Image from [291].

representative for the course of the spinal canal but not necessarily for the vertebrae location.

VISUAL HINTS The curvature of the spinal canal along the coronal axis is perceived well in the sagittal view. However, the curvature along the sagittal axis, i.e., the viewing direction, is only deducible by rotating the scene. Hence, the sagittal view is augmented by three visual hints improving the curvature perception. (1) A highly transparent sagittal plane passing through B is added to the scene. The position of the ribbon parts with respect to the plane (in front/behind) and the visible intersections of ribbons and plane support the differentiation between spinal canals being mostly bended towards the viewer from those being bended away (Fig. 36, upper inset). (2) The ribbons are colored according to their distance to the sagittal plane. A diverging color scale is used to distinguish between parts in front of the plane (blue), close to the plane (white), and behind the plane (red). (3) A transversal plane is positioned below the ribbons and a light source is positioned above them. Shadow projections are computed and drawn on the plane. They provide an estimate of how far the representatives extend to either side of the plane (Fig. 36, lower inset). In some cases, the projections revealed subtle differences in shape, which could hardly be inferred from the other two hints.

MEASUREMENT AND INTERACTION In order to facilitate a more quantitative analysis of the centerlines and to support a comparison of individual representatives, a vertical and a horizontal axis including tick marks are added to each cluster representative (Fig. 36). All axes are located within the sagittal plane (1). An initial pair of axes running through B has been computed based on the entire set of centerlines and then copied and translated together with each cluster along the coronal axis (Fig. 35). The vertical axes are assigned a unique cluster color to interrelate the representatives and the cluster size legend.

The interaction with the visualization exceeds standard 3D scene navigation. Individual representatives may be picked by the user and all centerlines of the corresponding cluster are visualized. The measurement of the spine based on neuralgic points is of crucial importance and has a long tradition in orthopedics. Hence, two measurement widgets have been added for measuring distances and angles (Fig. 37). Both widgets are bound to the geometry of the ribbons in order to simplify measurements in 3D space. The visualization has been implemented in C++ and the Visualization Toolkit (Kitware, Inc., Clifton Park, NY, U.S.).

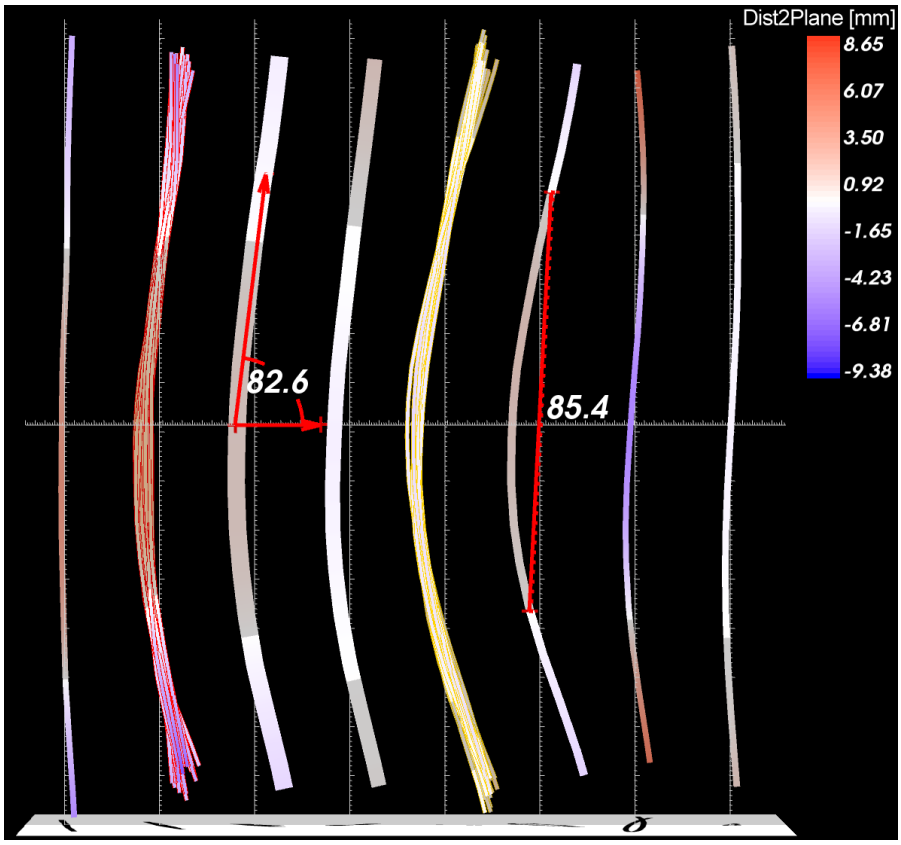


Figure 37: The prototype comprises of various interaction facilities. The user may pick a cluster representative, i.e., a ribbon, causing the corresponding cluster to be visualized (centerlines with red and yellow halos). Widgets for measuring distances and angles facilitate a quantitative analysis of the spinal shape. Image from [291].

4.4.3 *Results & Discussion*

In this section, preliminary results combining the shape visualization with associated population study data are presented. As seen in Fig. 36, the clustering step is a good way to detect outliers in the data as clusters with very few subjects that have an unusual shape. This can be utilized for finding

pathological spine shapes–even for subjects that do not have a diagnosed back disorder. The technique scales well regarding the number of input center lines. It is possible to generate an overview for hundreds of subjects as well as for smaller subsets, e.g., subjects that share certain similar attributes. A subset visualization can be applied to detect if the different shape clusters imply a significant difference in associated variables of interest. Does, for example, a physically demanding job correlate with an extraordinary curved spine?

The clinical partners expected the lumbar spine to be more straight along the coronal axis for tall people, while being more sinuous ("*lordosis*") with decreasing body height. To check the results for medical plausibility, subsets of the data based on *body height* are created. For each cluster the distance to the arithmetic mean of *age*, *body height*, and *weight* is calculated. The means of the absolute lordosis curvatures K are computed using the Frenet formulas [76].

While the mean curvature K for people sized 150 – 160 cm is $38.99 \cdot 10^{-4}$ ($\sigma = 9.99 \cdot 10^{-4}$), it gets smaller the larger the subjects are, being at $34.59 \cdot 10^{-4}$ ($\sigma = 9.98 \cdot 10^{-4}$) for 160 – 170 cm and at $31.95 \cdot 10^{-4}$ ($\sigma = 8.88 \cdot 10^{-4}$) for 180 – 190 cm tall people. The expected differences in the distinct groups could not only be confirmed; also clues can be given for groups which share similar curvature. When looking at subject groups of body height 150 – 160 cm, 160 – 170 cm and 170 – 180 cm a cluster of subjects who are about 10 years older than the rest of the group could always be observed. They all presented a lordosis shape as well as an "S" shape in sagittal direction ("*scoleosis*", see Sec. 4.1). Since a cluster showing the same characteristics was found in distinct subject groups, it is subject of further investigation.

This finding is an example of how a clustering result can create groups related by shape in order to find other correlations in the associated socio-economic and medical attribute parameters. It can also serve as starting point for a visual analytics tool to detect risk factors. This finding is an example of how the clustering step can also support a hypothesis generation step by creating subject groups with similar shape characteristics which can then be projected back to the associated subject data to find new correlations. It can also serve as starting point for a visual analytics tool to detect risk factors.

The visualization aims for a visual comparability of the clusters. Additionally statistically reliable shape-describing features would enhance the method by making statistical calculation applicable to deformation information. This can be achieved by storing the curvature and position of several fixed points in the FEM. While the visualization allows for the characterization of the lumbar spine curvature, it is currently not possible to predict information about spinal canal narrowings, which can also be an indicator for pathologies like spinal stenosis. This is also the case for a vertebrae deformation, which is an indicator for osteoporosis.

4.4.4 *Summary and Conclusion*

Applying the analysis of medical image data associated with non-image data in a population study context is both promising and challenging. The multitude of subjects requires robust yet precise and at least semi-automatic detection and segmentation algorithms which capture the shape of a structure of interest over a large space of subjects. Subjects with morphologically manifested pathologies render the automatic and semi-automatic detection and segmentation of the image data difficult. Assessing the resulting information space demands visualizations that map relevant information among large groups of subjects.

The goal of the following section is to include more shape descriptors and apply the technique to all population study subjects. This allows for a statistically reliable comparison of clusters. With the methods presented in this section, only the overall curvature and torsion is calculated. Those can be misleading metrics, since coronal as well as sagittal deformation can induce a large curvature. Healthy and pathological subjects can be analyzed based on their shape differences. Those and other shape-describing metrics can be transferred to the population study data dictionary. Information about vertebrae alignment is also of great interest.

The presented approach implements a pipeline for analyzing the lumbar spine canal in order to correlate its shape to other variables associated with the population study. This was done using an association to *body height*, *gender*, *age* and *weight*. While this was a first step to confirm the expected shape in different subject groups, it has to be enhanced to be applicable to all data variables measured in the population.

In the following section, a web-based visual analytics framework is presented that allows for information visualization on non-image data in combination with complex data set queries including the shape of structures. This also allows for complex queries based on the shape of tissue types, which are hard to recreate using classic statistical processors. It provides the epidemiologists with a fast and effective way to analyze their data sets exploiting the potential that lies beneath the numbers. The abstraction of complex data models of the lumbar spine as well as the agglomerative hierarchical clustering technique serve as image-based input for the methods developed in the following section.

4.5 INTEGRATING IMAGE DATA WITH NON-IMAGE VISUALIZATIONS

This section is based on

> **Paul Klemm**, Steffen Oeltze-Jafra, Kai Lawonn, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Trans. on Visualization and Computer Graphics*, 20(12):1673-1682, 2014.

Steffen Oeltze-Jafra, Kai Lawonn and Bernhard Preim helped with adapting the IVA methodology to the epidemiological application domain. Steffen Oeltze-Jafra provided valuable input on how to define *range* and *domain* features in this domain. Katrin Hegenscheid and Henry Völzke provided the data, research questions, hypotheses and background knowledge to the application domain. They are also the evaluating experts and provided important input for improving the iteratively designed prototype.

An *IVA* approach [255] for the combined analysis of image and non-image data is proposed. It is an implementation of the workflow described in the introduction of Part ii. Visual queries and direct feedback of Visual Analytics systems allow for a fast exploration of the data space incorporating many different variables. Intended as an extension to the well-established epidemiological tools it provides a way to rapidly validate hypotheses and to trigger *hypothesis generation* using data mining methods, such as clustering. *Hypothesis generation* gains importance since the number of epidemiological variables increases and the focus shifts towards more complex relations involving more than two variables. The contributions are:

- an IVA workflow for population study data to allow both hypothesis-driven analysis and hypothesis generation,

- visualization techniques that incorporate both information visualization and 3D rendering of organ shapes as well as combining them with epidemiological graphics and key figures,
- highlighting subject groups and variable associations using shape-based clustering and statistical contingency measures.
- an implementation of the presented methods in a web framework based on `WebGL`, `D3.js` and `NodeJS`.

The approach is applied to a data set compiled to analyze lower back pain and aiming to determine variables that indicate pathological changes. This data set comprises 127 variables and 2 sequences of MRI data from 6,753 subjects. The presented method is implemented using modern web technologies, such as `WebGL`, `D3.js` and `NodeJS` to make them easily accessible for the domain experts to enable a fast feedback loop.

Some methods used in this section are not restricted to analyze image-based data, such as the contingency matrix or the pivot table. They could also be discussed in Chapter DATA-DRIVEN VISUAL ANALYSIS OF SOCIODEMOGRAPHIC, MEDICAL AND LIFESTYLE FACTORS, since they mostly represent non-image data or image-derived metrics. The methods are incorporated in this chapter, because it is part of the analysis workflow, which is specifically designed to analyze epidemiological hypotheses focused on, but not restricted to image data. The important related work for this method can be found in Section 3.3.4.

4.5.1 *Image-Centric Population Study Data in an Interactive Visual Analysis Context*

In this section, the implementation of the *IVA* principle on population study data is presented. As described in Section 3.2.2, IVA comprises different analysis patterns, depending on the type of the conducted analysis as well as the underlying data type.

DOMAIN AND RANGE VARIABLES As stated in Section 3.2.2, data are characterized by a combination of independent variables, such as space and/or time, and dependent variables, like temperature or pressure. These are viewed either using physical views, which usually employ volume rendering to display spatio-temporal behavior, as well as attribute views, such as scatter plots, which show relationships between data attributes associated with the spatio-temporal observation space. Transferred to epidemiological data, the residential area of population members could be interpreted as *space*, the different assessment cycles of a longitudinal study as *time*, and the image and non-image data as *dependent variables*. This method neglects geographical and temporal aspects. Instead, an abstract model is employed, which considers the subjects as living in a joint image space where each of them is represented by a segmented organ or structure. For instance, the lumbar spine is segmented over all subjects and all lumbar spines are co-registered spanning a joint space. Then, two types of dependent variables exist:

1. the socio-demographic data and medical examination results and
2. variables derived from the segmented structures, e.g., spinal curvature or misalignment of the vertebrae.

An alternative of the image space would be the shape space generated by extracting the major modes of variation from all segmentation results as

presented by Busking et al. [29], which is incorporated in Section 4.3. Based on this model, the three analysis patterns of *IVA* can be employed.

Local Investigation refers to the inspection of dependent variables with respect to subsets of the image or shape space. For instance, the epidemiologist selects several lumbar spines with a common characteristic in the image or shape space and inspects the associated dependent variables in an attribute view [104]. The selection step requires dedicated interaction techniques for defining a subset. Alternatively, derived shape-related variables opposed in an attribute view or automatic techniques for shape clustering may be employed, as presented in Section 4.4. Clustering algorithms can be used to investigate associations between shape groups and other non-image variables. Analysis of outliers can indicate segmentation errors or a group of subjects sharing a pathology.

Feature Localization refers to the search for structures in the image or shape space with a defined characteristic. The epidemiologist may be interested in all female subjects with lower back pain and wishes to see the corresponding spines in a physical 3D view.

Multivariate Analysis refers to an investigation of multi-variate properties of the dependent data by specifying a variable in one attribute view while analyzing the value distribution with respect to other variables in other attribute views. Epidemiologists may define a variable in a scatter plot of the body mass index (BMI) and age to inspect the result in a histogram of body height. These associations may also be summarized using pivot tables, which are widely used in epidemiology.

4.5.2 *Data Preprocessing*

Transformation operations on the data to prepare it for an *IVA* system are denoted as data preprocessing.

NON-IMAGE DATA Data obtained using questionnaires or medical tests are often stored using statistical packages such as `SPSS`, which have a proprietary data format with limited export capabilities. Exporting the data in the respective tool to a `CSV` file and then converting it to file types that are easily manageable, such as `JSON`, makes it readable for modern programming languages. This can be achieved by using data wrangling tools such as `OpenRefine`[2], which also validate the data (find missing data, clean up bad formatting, transform scales). A data dictionary stores information about each manifestation of a variable, such as a detailed description, its meaning as well as the unit of measurement. Exporting the data dictionary, which stores information about each manifestation of a feature, is also an important step to get a detailed description of data variables and the meaning and unit of measurement of their values. Missing data are denoted using error codes indicating their cause ranging from ethical to medical and personal issues. Therefore, these are also included as error codes which have to be marked as such in the data dictionary. It is important to not simply interpolate missing features, because it imposes a high change of introducing a new bias into the data. `SPSS` allows to store the data dictionary as part of its proprietary format. In order to export it to an open format, we choose to include the description, range and unit of measurement in a structured `JSON` file, where the variable ID acts as reference, which allows for linking between the `CSV` and `JSON` file.

IMAGE DATA Information about anatomical structures, such as diameter or volumes, is extracted from the image data. This is either done manually

2 Developed by Google, Open Source; openrefine.org

by expert setting, landmarks or a (semi-)automatic detection, registration and segmentation. These algorithms have to deal with a large inter-subject variability of the anatomical structure [296]. Grey value comparison is used to measure the quantity of fat, water and–application-specific–the iron content (liver) or the distribution of grey and white brain tissue. Morphometric variables are derived to allow for statistical comparison of the tissue, which incorporates mostly positions, diameters, volumes as well as distances and alignment to other structures [123]. For more details, see Section 4.2.

4.5.3 *Analysis Workflow*

The proposed *IVA* workflow consists of three major steps, as illustrated in Figure 27: Variable selection, visualization and brushing. A hypothesis-driven analysis usually starts with the selection of variables or shape groups derived from a shape-based clustering. Hypothesis generation with focus on image data starts with a shape-based clustering or an *overview visualization* of all variables. The variable is mapped using an automatically chosen visualization appropriate for its data type (described in detail in the following section). The visualization techniques have to combine both image- and non-image data to set domain and range data in relation to each other. In our system, the visualization can either be brushed or new variables can be added to the analysis. Brushing methods are subdivided using the different IVA levels presented in Section 3.2.2. In this method, brushed regions are treated like categorical variables, as they divide the subject space in the same way. Selecting variables also triggers a *multivariate analysis* using contingency values (described in the following section) to highlight associated variables. A sample workflow using interaction and visualization techniques described in the next section can be seen in Figure 38.

4.5.4 *System Design and Implementation*

The suitability of visualization techniques for epidemiological data depends on their ability to compare multiple data variables while highlighting associations. The methods have to reflect the routines that epidemiologists take into their research. Visual evaluations of data are therefore as important as methods allowing for numerical data analysis. In the following sections the different parts of the system are presented.

Design and Visualization Techniques

The epidemiological experts Katrin Hegenscheid and Henry Völzke are located in Greifswald, while the methods were developed in Magdeburg. Therefore, it became clear early that the communication and exchange of software has to be focused on web technologies to facilitate rapid feedback cycles. By running the prototypes on server machines, software exchange became as easy as sharing a weblink, giving the opportunity to include the clinical experts in the development process with little effort. Incorporating the *IVA* workflow for image-centric population study data requires *overview visualizations* as well as *multivariate visualizations* that bring image-derived information in context to non-image variables.

The focus on web technologies is not without trade-offs. Classical UI elements, such as the menu bar or custom right-click menus, are technically possible, but not common in this domain. In favor of a clean layout, the system was designed without such components. Since the previously described *IVA* workflow allows for many different ways to analyze the data, the interface was designed in a minimalistic manner, treating the resulting space as

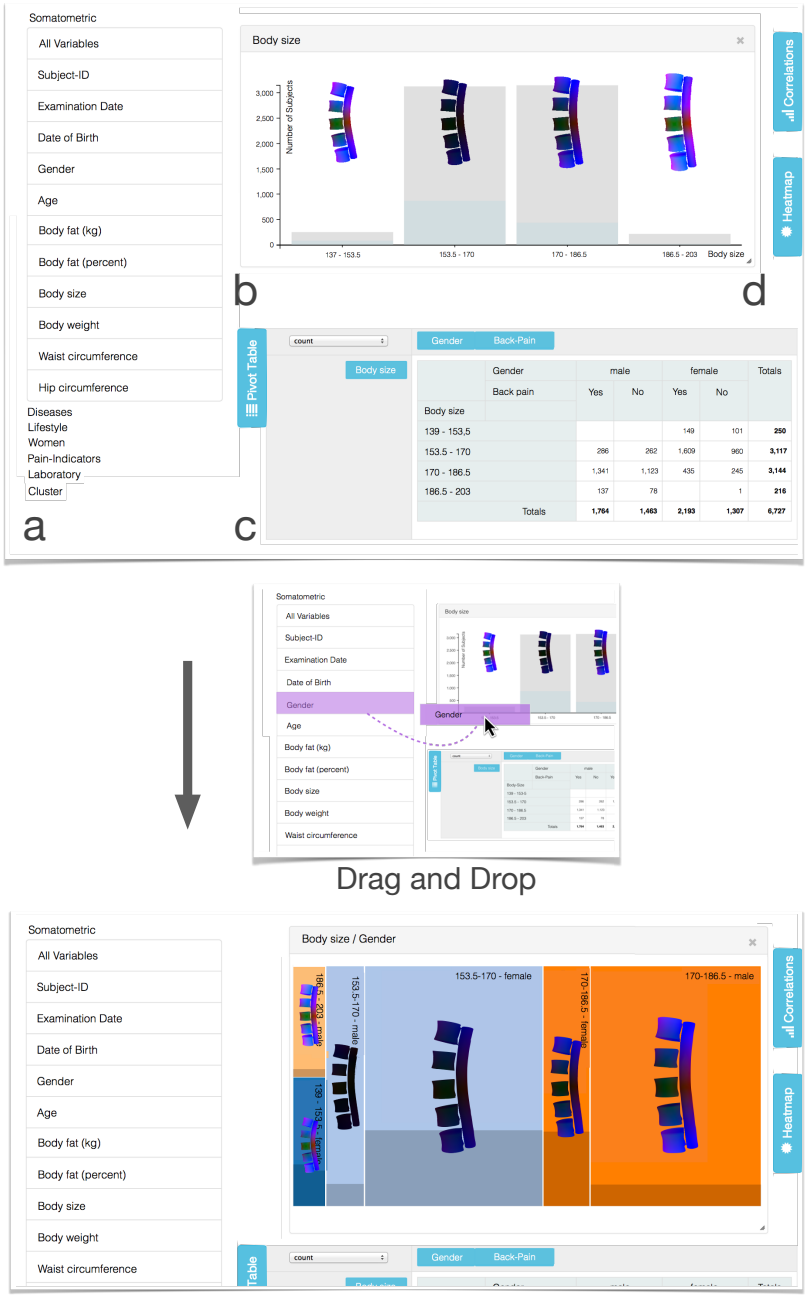


Figure 38: Workflow of introducing new features into the canvas area. (Top) Screenshot from the front end, which is divided as follows: (a) The sidebar containing all variables as well as the groups defined in the analysis process; (b) the canvas area where variables can be added via drag and drop and the visualization is chosen automatically according to the data type; (c) the interactive pivot table showing the exact numbers for each displayed variable combination; (d) buttons to open panes containing the contingency matrix, contingency pane and pivot table. The data displayed is used to analyze the lumbar spine. Dropping the *gender* parameter on the already plotted *body size* container creates a mosaic plot combining both variables (bottom). In a prior step, the user selected all subjects with diagnosed thyroid disorder. Their shares are denoted in the bar chart as blue and the mosaic plot as dark shade. Subjects between 153.5-170 cm body size are more affected by thyroid disorder (box plot) and are mostly female (mosaic plot). Distance to the mean mesh of subjects with thyroid disorder is encoded as red for x-axis, blue y-axis and green z-axis. Image adapted from [293].

canvas for the data. The workspace was divided into four major parts, as illustrated in Figure 38 and 39.

- The *sidebar*, which contains all epidemiological variables. The cluster results group variables like categorical variables and are part of the sidebar as well (Fig. 38 a).
- The *canvas* holding all visualizations. Elements can be added, arranged, resized and removed freely (Fig. 38 b).
- The interactive *pivot table* gives detailed numerical information of the variables in the canvas view. This view on the data is familiar to epidemiologists (Fig. 38 c).
- The *contingency view* depicts relations for variables in the canvas in a contingency matrix (Fig. 39) and a *contingency list*.

The design of the system follows Tufte's design principles for good visualizations (see Sec. 3.2.1). Therefore, the number of user interface elements is minimized to allow the user to focus solely on the visualizations. This is achieved by facilitating most of the interactions using drag and drop mechanics. The system does not use context menus to avoid hiding functionality from the user. Also, no menu bar is incorporated. The visualizations themselve have as little descriptive labels as possible to maximize the data-ink ratio.

SYSTEM LAYOUT Several layouts were tested with this prototype. The initial idea was to make all components freely arrangeable and resizable on a large *canvas* area. This idea was soon dropped, since domain experts reported a cluttered workspace, which required a lot of scrolling. The introduction of separate panes for the contingency matrix, pivot table and sidebar, displayed with a mouse click on the corresponding button and sliding on top of the *canvas* was considered more feasible (Figure 38 shows the system with reeled-out pivot table pane). All user-generated visualizations are part of the *canvas* and can be arranged freely.

SIDEBAR Only the *sidebar* is visible at system start. It categorizes all variables into different types, for example somatometric (measurements of the human body dimensions), disease- or lifestyle-related, pain indicators and laboratory data (Fig. 38 a). It also contains subject groups defined by automated shape clustering. Groups are treated like categorical variables. Variables can be dragged from the sidebar into the canvas area for a *feature localization*, which works as follows. This triggers an adaptive feature visualization suitable for the current data type.

ADAPTIVE VARIABLE VISUALIZATION The visualization type, inspired by `GPLOMS` (see Sec. 3.2.2), is dynamically chosen based on the variable types and number to allow for *multivariate analysis*. Categorical data are either mapped to bar charts (single variables) or mosaic plots (multiple variables). Figure 38 describes this dynamic adjustment. Continuous data can be visualized using scatter plots (two variables) or parallel coordinates (multiple variables), but in epidemiology, this data type is usually categorized into ordinal groups of *equal size*. Since the number of categories often depends on the hypothesis, the discretization steps can be adapted dynamically. Too many groups potentially generate sparse bins not suited for statistical evaluation. Not enough groups overgeneralize information. Adaptive discretization is an option, but imposes possible overfitting to the data. Conclusions based on statistical relationships derived from groups already biased by variable

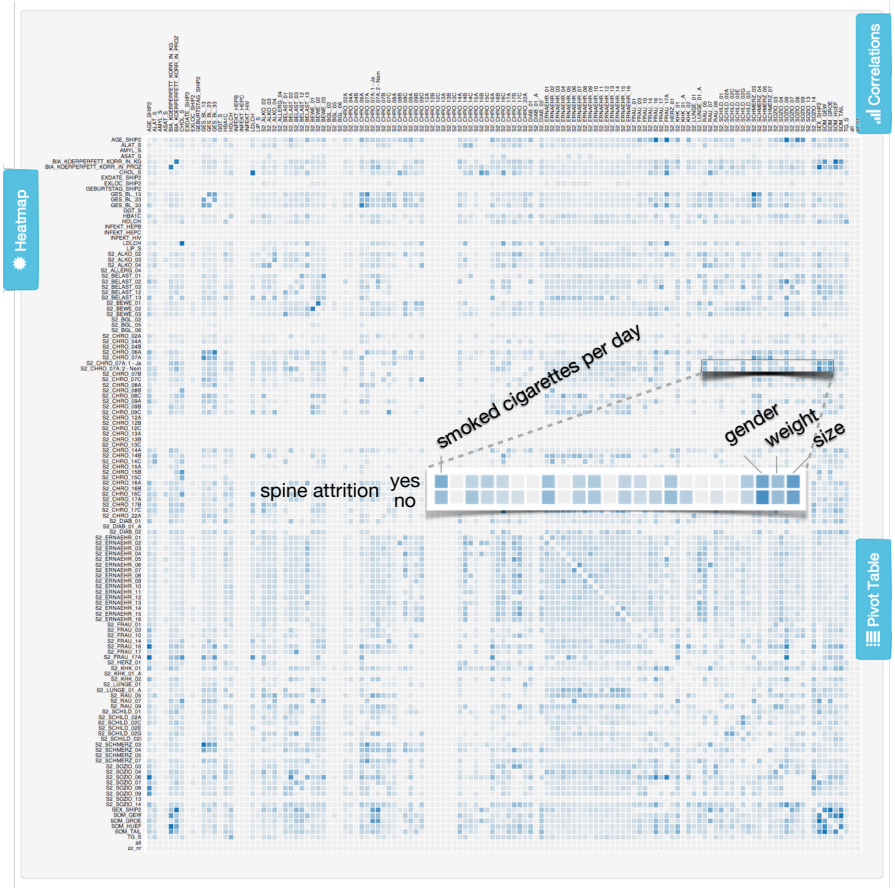


Figure 39: Contingency matrix of 129 variables (127 data set variables, 2 cluster results) showing 16,641 combinations. Similarity is calculated using the *Cramér's V* contingency value. Color brightness encodes association strength. Moving the mouse over an entry enlarges the variable names for better readability. The enlarged excerpt shows associations for shape clusters of subjects with and without diagnosed spine attrition, which show associations between gender, weight, body height and smoking behavior. The contingency pane is not shown here. Image from [293].

distribution are heavily influenced by the used discretization. Therefore, the convention to use bins of equal size is applied.

Following Tufte's concept of *small multiples* [259], information derived from the medical image data are incorporated into the plot by including color-coded mean shapes for each manifestation (Figure 38 b). The 3D plots can be navigated using standard mouse input, the camera is synchronized between all views to enable direct comparison. The distance from a group mean shape is mapped to the global mean using color. This allows to assess local shape changes (Fig. 38) and is an important information to the epidemiologist. Until now, epidemiologists were not able to inspect shape differences based on non-image variables. Dropping a variable on an existing plot adapts the visualization dynamically to allow for comparison (Fig. 38 right).

To support *feature localization*, subject groups can be brushed via a double-click on its representative in the visualizations. Holding down the shift key allows users to select multiple manifestations. Brushed groups act as reference for the shape visualization, calculating distances based on the mean shape of the brushed selection. This allows to highlight distances between subjects. The share of subjects of this subgroup is linked to all other views (Fig. 38 left). If the user selects all female subjects in a visualization of gender

distribution, all other displayed meshes are color-coded with their distance to the female mean and the share of female subjects is highlighted in the information visualization. Brush selections are propagated to all visualizations allowing for fast feature querying.

PIVOT TABLES Pivot tables are frequently used to present the data in epidemiological publications. Epidemiologists are used to perform *multivariate analysis* of groups based on table representations. Thus, an interactive pivot table was introduced. These tables clearly convey the subject count in each group (see Figure 38 c). However, they quickly get confusing when they are divided into many subgroups. This problem was tackled by making the order and number of displayed variables adaptable. This also applies to the assignment of row or column variables. Another way to avoid clutter is the user-driven selection of displayed variables. To allow for better comparison with respect to variables, the values of each cell can also be displayed as percentage of the variable represented of either the row or column.

AUTOMATED VARIABLE SUGGESTION USING A CONTINGENCY MATRIX Highlighting potentially interesting associations in the data set is one major benefit of the *IVA*-powered approach and is part of the *multivariate analysis* pattern for analyzing variables outside the shape space. Turkay et al. [263] used the approach to calculate statistical key figures based on the distribution functions of each variable derived from the image data. Since the majority of the data are categorical variables, different solutions have to be employed. The *Cramér's V* contingency coefficient can be used to calculate correlations between categorical variables [52]. It is based on *Pearson's* X^2 distribution test [200], which uses contingency tables holding the counts of subjects for all possible manifestations of two variables. *Cramér's V* is defined as:

$$V = \sqrt{\frac{X^2}{N(k-1)}}, \tag{8}$$

where X^2 equals *Pearson's chi squared*, N is the total number of observations and k is either the row or column count, depending on which one is lower. V yields values between 0, meaning that two variables are completely independent, and 1 indicating that they are the same. *Cramér's V* does not allow to infer the dependency direction.

It shares the same restrictions as *Pearson's* X^2. The expected counts in the contingency table have to be larger than five for 80% of the entries and no expected value must be smaller than one [44]. Some manifestations and variable combinations, which are only exposed by small subject groups, cannot be assessed with this technique. They cannot be included into the epidemiological analysis, since statistical validation needs a minimum count to be valid. The contingency matrix highlights correlations between all variables. This aims to highlight variables possibly associated with the focused hypothesis and to trigger new hypotheses. Contingency is visualized using an interactive contingency matrix with association power mapped to color brightness. The distinction whether an association is a confounder or an effect depends on the context defined by the hypothesis and is a decision to be made by the domain expert. The contingency matrix visualization is an *overview visualization*–something the epidemiological community lacks and is in great need of.

CONTINGENCY PANE Dropping a variable into the canvas area adds an entry for each manifestation of it to the *contingency matrix*. Testing sessions revealed that it was tedious to open the matrix every time a new variable

is added. As a consequence, the *contingency pane*, a table containing correlating variables for the last added visualizations in descending order of the *Cramér's V* value was added. *Contingency pane* entries can be dragged and dropped into the *canvas area* just like variables in the *sidebar*.

INITIALIZATION AND CLUSTERING Using variable suggestion allows to initialize the system with a set of potentially interesting visualizations. After testing and domain expert feedback, this idea was dropped. Reasons for this are twofold. Very often, high correlations are obvious, such as gender with menstrual status. Also, it was observed that the variables of interest are dependent on the specialization of the domain expert.

Subject clustering is triggered automatically as *local investigation* for a variable after it was added to the canvas by the user. The clustering method and parameters are transfered from Section 4.4. A status indicator at the bottom of the screen keeps the user informed about the pending clustering result, since the process can take up to ten seconds. Clustering results are listed in their own category in the *sidebar*. Since a clustering process can take a couple of seconds, a status indicator at the bottom of the screen keeps the user informed.

Implementation

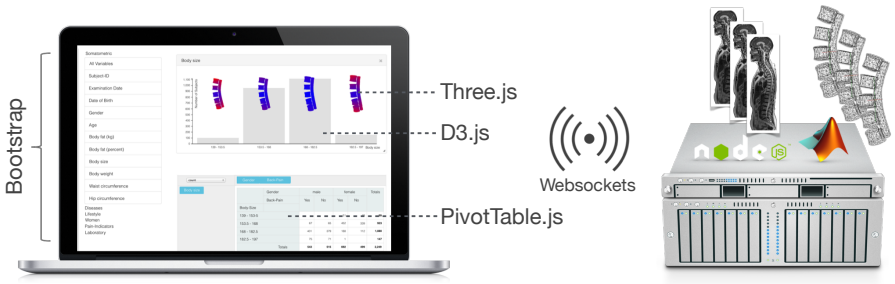


Figure 40: Overview of the technologies incorporated in the prototype. The front end solution (left) uses *HTML5/CSS3*, *WebGL* and *SVG* to display the data. The *NodeJS*-based back end (right) stores all image and non-image data and transfers it to connected clients. All computation-heavy operations, such as calculation of mean shapes or distances, as well as statistical processing are performed on the server side to keep hardware requirements of client systems low. Client-server communication is accomplished via the *Websocket* protocol. Image from [293].

In this section, the implementation of the presented methods using open web standards is discussed. To provide a fast communication loop between method development and expert input, modern web technologies are employed. In addition to the obvious advantages of web technologies, the following aspects are crucial for this work:

- No additional software needs to be installed, most people use a decent state-of-the-art web browser, even on mobile devices.
- The client-server structure allows for employing heavy computation on a server machine and transferring results to the client.
- Disk-space demanding image data remains on the server and elements can be transferred on demand.
- Since image data for thousands of subjects requires hundreds of gigabytes disk space, it can remain safely on the server and elements can

be transferred on demand. High confidentiality standards of the data are met by a password protecting the access.

- Recent developments in *WebGL* applications running in browsers with near-native performance result in many open source libraries, which are well documented and driven by active communities. *WebGL* is used for rendering shape information.

These advantages do not come without drawbacks. Sophisticated libraries and languages, such as the `Visualization Toolkit` (VTK) [3] or `R`[4] for statistics, are either not available at all or only accessible through complex client-server systems. Therefore, many standard methods had to be written from scratch. The back end is realized using `NodeJS`[5], which is based on the Google V8 Javascript runtime environment. Due to its event-driven non-blocking I/O model it is fast and responding even with heavy workload, such as mesh processing. Non-image data for all subjects including the data dictionary is stored in a `JSON` file on the server. Image data are available as raw `DICOM` files. Segmentation masks of anatomical structures are represented as meshes, suited for comparing subjects. The requested data are transmitted when a client connects. The server performs heavy statistical tasks, such as calculation of *Cramér's V* values for all variable combinations in order to keep the computation time on the client as low as possible.

The front end is created using `Bootstrap`[6] as foundation for the layout and basic UI elements using `HTML5`, `CSS3` and `Javascript`. Information visualizations such as scatter plots and bar charts are created using the popular `Data-Driven Documents (D3.js)` library [24], which works well for attaching data to visible elements like vector graphics and provides powerful transformation and mapping tools. The pivot table implementation uses `PivotTable.js`.[7] `Three.js`[8] allows GPU-accelerated data rendering using `WebGL`. The WebSockets protocol handles the client-server communication. Since the employed clustering algorithms are written in `MATLAB`, they are accessed using the `NodeJS` server. This is accomplished by converting it to a parameterized standalone console application, spawned by `NodeJS` on client request. The result is read from the console output and is returned to the client. All parameter-steered console applications can be incorporated in this context. Figure 40 summarizes the incorporated technologies.

4.5.5 *Application*

This section describes how the presented *IVA* workflow is used in the epidemiological application. A qualitative evaluation was conducted with two domain experts on the lumbar back pain data set, described in Section 4.1. Characterizing the healthy aging process of the spine is a long-term goal for determining age-normalized probabilities for spine-related diseases by incorporating individual risk factors. The assessment of the spine shape potentially yields new risk factors for these diseases. These shape-related risks can then be translated into metrics, which can be statistically assessed.

Data Preprocessing

NON-IMAGE DATA To ensure a fast and easy data access outside of statistical processors like `SPSS`, the data was exported to the `JSON` file format.

3 Developed by Kitware Inc; `vtk.org`
4 Open Source; `r-project.org`
5 Developed by Joyent Inc, `nodejs.org`
6 Developed by Twitter, `getbootstrap.com`
7 Developed by N. Kruchten, `nicolas.kruchten.com/pivottable`
8 Originally developed by R. Cabello, `threejs.org`

Since it lacks export methods for data dictionaries, `SPSS` is incorporated to export the data to the `SAS v9+` format, which saves the data labels, and exported the data values as non-labeled `CSV`. A short script combined both data sources to a `JSON` file. The data types had to be transferred manually. Each variable is stored as an object containing information about:

- the data as array of values; categorical data and error codes are stored using IDs,
- the data type (continuous, nominal, ordinal, dichotomous),
- a detailed description of the feature, and
- the data dictionary translating value or error IDs to values.

Each feature is stored as an object containing:

- the data as array of values–categorical data and error codes are stored using IDs,
- the data type (continuous, nominal, ordinal, dichotomous),
- a detailed description of the feature, and
- the data dictionary translating value or error IDs to values.

Continuous variables are discretized to allow for *Cramér's V* contingency coefficient assessment. Following epidemiological publications, the number of groups is set to five (the *quintiles*) to allow for contingency assessment.

IMAGE DATA The lumbar spine was detected in the image data using a hierarchical finite element method by Rak et al. [213], as presented in Section 4.4. This semi-automatic method requires the user to initialize the FEMs with a click on the L3 vertebra. Two user-defined landmarks on the top and bottom of the L3 vertebra describe an initial model height estimation. The model uses a weighted sum of T1- and T2-weighted MR images to detect the lumbar spine shape. Once registered, it captures information about the shape of the lumbar spine canal as well as the position of the L1-L5 vertebrae. Due to incorrect initialization, strongly deformed spines, contrast differences and artifacts, the model was not able to detect lumbar spines for all subjects. A total of 2,540 tetrahedron models were obtained of the lumbar spine. Using the methods from Section 4.4, the centerline of the lumbar spine canal was extracted, which captures information about lordosis and scoliosis (the medical terms for spine curvature).

Shape Visualization and Clustering

The tetrahedron-based detection model consists of corresponding grid points for each structure instance. This allows to calculate shape distance and similarity. This information is used to calculate mean shapes as described in Section 3.1.3. The shape distance between meshes is mapped to color (recall Fig. 38). For dichotomous variables, the color represents distances between mean shapes of the two groups, for variables with more than two manifestations it encodes the distance to the global mean shape of all subjects.

Shape-based clustering is carried out via agglomerative hierarchical clustering of the spine canal centerlines (recall Section 4.4). Since the "correct" number of clusters in a given group is unknown, an estimate is computed by means of the knee/elbow method [227].

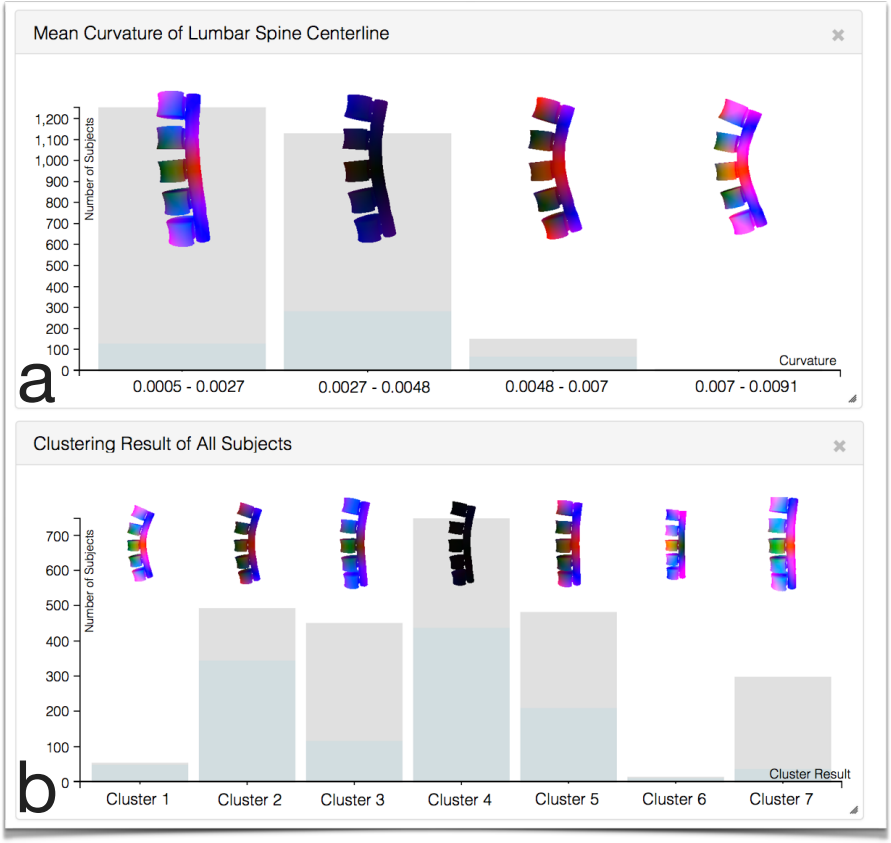


Figure 41: Screenshots from the hypothesis-free analysis. (a) Mean curvature of lumbar spine canal plotted against the mean shape of 58-74 years old female subjects (light-blue bars). Note the high amount of this subject group relative to the total count in the third group. The last group contains four outliers. (b) Clustering of all subjects yields seven groups, whereas cluster 4 assembles the mean. The light blue bars indicate the share of females in the group. Image adapted from [293].

Participants, Setup and Procedure

Inspired by Lam et al. [141], an investigation of *Visual Data Analysis and Reasoning* (*VDAR*) was conducted. This approach aims to characterize the system's ability to explore data, discover knowledge, generate hypotheses and help formulating decisions. Since it is hard to quantify these outcomes, Lam et al. suggest case studies for the *VDAR* by applying the think-aloud protocol to understand the domain expert's observations, inferences and conclusions when using the system.

The participants are two epidemiological domain experts. HV and KH are physicians with focus on epidemiological research. HV is a specialist in internal medicine (23 years of experience) and head of the SHIP, KH a radiologist (9 years of experience) and responsible for the SHIP MRI data acquisition.

SETUP Due to the large geographical distance between the participating institutes, the evaluation was done completely web-based. The experts accessed the prototype by entering the weblink into their browser. User input was observed using screen-sharing. Communication was enabled via webcam-supported Voice over IP. The total setup time including installing the screen-sharing application was about five minutes. Video recordings of the sessions allowed a detailed evaluation afterwards.

PROCEDURE At first, the computer scientist controlled mouse and keyboard of the participants' PC and demonstrated the basic functionalities of the system. This included the contingency matrix, the correlation view, how to introduce feature visualizations, how the color coding of differences works, and the pivot tables. As they understood the concepts, the computer scientist handed over the mouse and keyboard control and only observed from this point on. The epidemiologists were given two tasks: one hypothesis-free analysis of the data and one starting with an assumption. For each case, an analysis was conducted with each expert.

Case 1: Hypothesis-free Analysis

Analyzing the data set without prior hypothesis requires a starting point giving an overview of the data [239]. With the herein presented method, there are two ways to achieve this. Performing a *multivariate analysis* by viewing the contingency matrix (sized 127×127 tiles) or a shape grouping step using shape-based clustering. The first was chosen by both experts. Before, they were not able to look at all variables in the context of each other. To cite one expert, the contingency matrix "illumunates the data black box", making it possible to look at the data unbiased of assumptions.

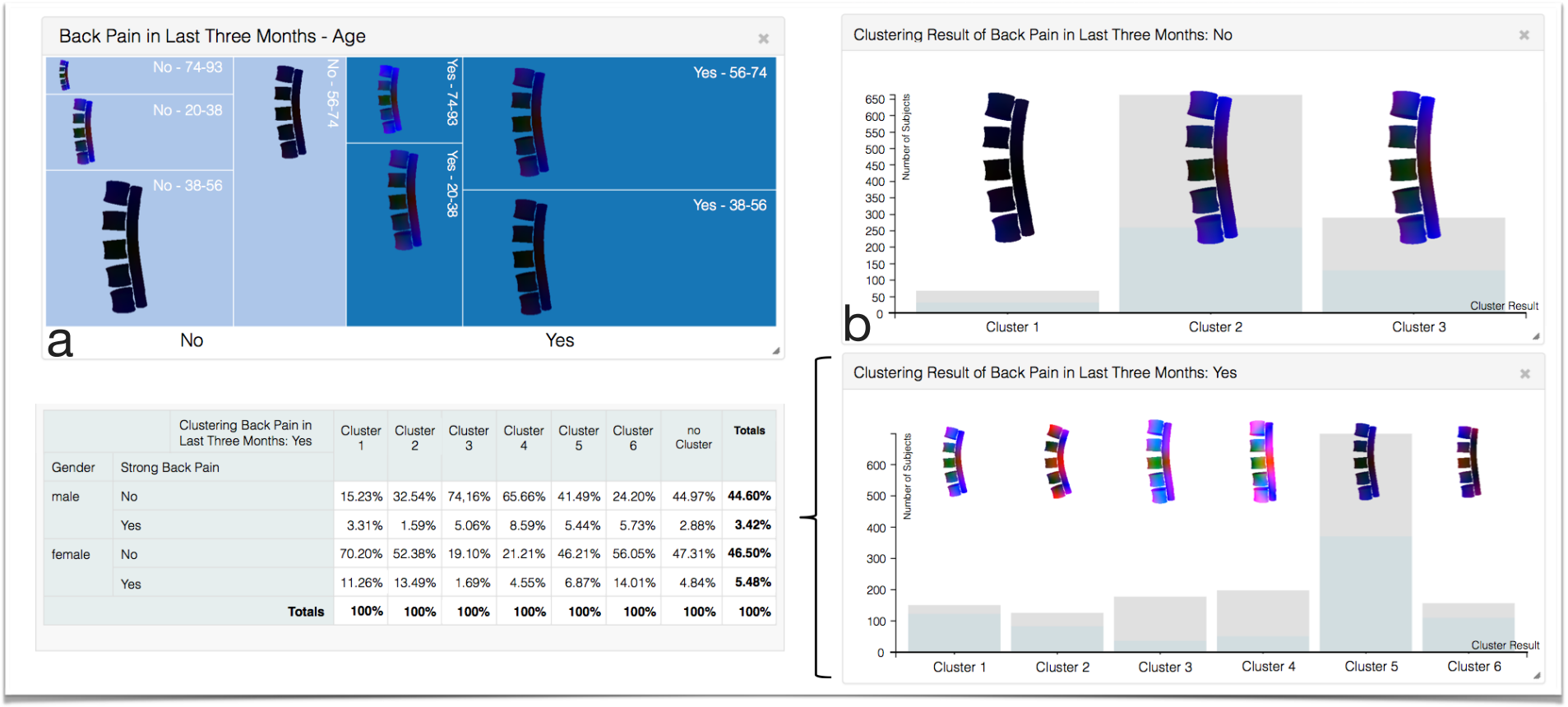
ANALYSIS 1 The radiologist (KH) was looking for correlations with shape-related variables in the data, finding that *spine curvature* correlates with *leg pain*, *age*, *body height* and *hormone replacement therapy status*. Due to the dense mapping of information in the contingency matrix, KH suggested to make this visualization full screen.

After this initial overview, KH performed a *multivariate analysis* by introducing variables, such as *age*, *waist circumference*, *weight* or *lumbar spine canal curvature* as bar chart views into the canvas area and selected subgroups to see how they are distributed and if they could observe unusual behavior in the mean shapes. This pointed out problems with the used categorization method splitting numerical variables into equally-sized ordinal bins. If a variable contains outliers, such as *waist circumference* (e.g., by subjects with morbid obesity), this approach leads to sparse categories, making it hard to calculate associations. The proposed expert solution for this is categorization using quartiles/quintiles and is described in detail later in this section.

A *multivariate analysis* using the *Cramér's V* contingency values for subjects with strong lumbar spine curvature showed that these subjects are primarily females between 58-74 years who also report pain radiating from their back into other body regions (Fig. 41 a).

ANALYSIS 2 HV also started with a *multivariate analysis* using the contingency matrix to analyze non-image variables, such as age-associated parameters like *income*, *blood fat values* or *number of born children*, but found no associations of interest. Therefore, he applied the *local investigation* pattern by a shape grouping step using shape-based clustering via dragging *All subjects* from the sidebar into the canvas area, triggering the shape clustering (Fig. 41 b).

Cluster 4 represents subjects with average shape. Other shapes differ with respect to size, such as cluster 2, 3, 7, whereas the last one and cluster 5 also represent a more straight spine, which is usual for subjects with larger body size. Cluster 1 and 6 contain outliers, characterized by their unusual shape and small number. Cluster 2 contains the second largest number of elements and was therefore of special interest. To get an overview of the suggested variables, the user opened the contingency pane (not shown here) to perform a *multivariate analysis* by looking at *Cramér's V* contingency values of



	Clustering Back Pain in Last Three Months: Yes	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	no Cluster	Totals
Gender	Strong Back Pain								
male	No	15.23%	32.54%	74,16%	65.66%	41.49%	24.20%	44.97%	**44.60%**
	Yes	3.31%	1.59%	5.06%	8.59%	5.44%	5.73%	2.88%	**3.42%**
female	No	70.20%	52.38%	19.10%	21.21%	46.21%	56.05%	47.31%	**46.50%**
	Yes	11.26%	13.49%	1.69%	4.55%	6.87%	14.01%	4.84%	**5.48%**
	Totals	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Figure 42: Screenshots from the hypothesis-based analysis. (a) A mosaic plot mapping age against the dichotomous questionnaire answer to *"Did you experience back pain in the past three months?"*. (b) Clustering result of *"Did you experience back pain in the past three months?"* Yes/no with female share in each group. Cluster 1 and 6 for answer *"Yes"* contain mostly women. The pivot table shows how many subjects with strong back pain are in each cluster for answer *"Yes"*. Subjects in cluster 1, 2 and 6 report strong back pain more often than subjects in other clusters. Image from [293].

all clusters, revealing a correlation with *gender* (0.29) and *body size* (0.24). Therefore, another *multivariate analysis* was carried out by dragging the *gender* variable to the canvas and selecting all female subjects (Fig. 41 b). Cluster 1 contained primarily female subjects. The contingency table showed contingency values ordered by magnitude shows correlations with *leg fatigue* (0.45), *physically heavy work* (0.43), *body weight* (0.32), *dyspnoea* (0.3) and *headache intensity* (0.3). No correlation with back pain was showed in the contingency table. Since it is a pain indicator, headache was of special interest and was further investigated by incorporating a pivot table setting *headache intensity* in relation to cluster affiliation. It was found that cluster 1 subjects report heavy headaches more frequently than other subjects (9.26% of subjects compared to 1.34% - 3.53% in the other clusters). Another *multivariate analysis* using a pivot table set gender and employment status in relation to clustering affiliation. It shows that cluster 2 contains mostly women and also has a larger unemployment rate, while the overall employment rate of women and men in the data set is almost exactly 50% (students and pensioners are not counted as employed). While all observed features seem to be plausible associations related to back pain, the values indicate that cluster 2 contains subjects with chronic back pain radiating to the legs. Metabolic parameters, such as *blood fat* and *blood sugar*, are also possibly associated features. The employment status is a feature relating to many different lifestyle factors such as income or nutrition as well as age, and might act as confounder.

The experts emphasized the importance of methods providing an overview of the data for hypothesis generation. With the presented *IVA* approaches they were quickly able to confirm medical knowledge and to elaborate new hypotheses. These hypotheses are focused on the correlations observed together with shape-based correlations. Namely, the correlation of a cluster containing females and their high share of reports of headaches provides an interesting subject group for further analyses. Also, the high share of back pain reports for the subject cluster with an unusually high unemployment rate may be analyzed further to see if employment status is a predictive factor for specific subject groups. One observation was that the domain experts are more likely interested in variables they are familiar with and have personal clinical experience with.

Case 2: Hypothesis-driven Analysis

If the user proposes a hypothesis about a relation between a non-image variable and shape, the workflow slightly differs from the hypothesis-free analysis. The starting point follows the *feature localization* pattern, where a variable of interest is selected by dragging it into the canvas area and viewing the subjects' distribution as well as their shape differences.

ANALYSIS 1 Hypothesis: "*Back pain is associated with age and lumbar spine shape*". To validate this hypothesis, a *feature localization* was performed by combining the dichotomous variable "*Did you experience back pain in the last three months?*" with age in a mosaic plot by dropping both variables on the canvas area (Fig. 42 a). HV was not able to observe the expected effect in the visualization. Reasons for this are twofold. Age influences the lumbar spine shape, while the differences between subjects with and without back pain are small. The major differences seen in the visualization are therefore related to the age variable, masking differences related to the back pain parameter. The second explanation is the commonality of back pain in our society. As seen in Figure 42 (a), subjects reporting back pain are the majority, which makes it difficult to extract parameters that reliably describe back

pain. A *multivariate analysis* using the contingency table showed a strong association between *back pain* with both *gender* (0.37) and *body height* (0.35). *Body height* was explained as a confounder for *gender*, since female subjects on average are smaller than male subjects. The analysis solely based on shape-accentuated body height differences in *gender*, which clouded the differences of *back pain*.

The epidemiologists pointed out that they would like to see a more intuitive and fast way to select subgroups based on different variables to make full use of the analysis capabilities, as discussed in the Section *Further Feedback and Lessons Learned*.

ANALYSIS 2 Hypothesis: "*Back pain is related to lumbar spine deformation*". The previously discussed analysis questions the suitability of the lumbar spine segmentation for analyzing back pain, leading to this analysis. Therefore, the dichotomous variable "*Did you experience back pain in the past three months?*" is dropped into the canvas area. Figure 42 (b) shows the results of the automatically triggered shape-based clustering for subjects with and without back pain. The clustering algorithm finds only three homogeneous clusters close to the global mean shape for subjects reporting no back pain. The cluster analysis for back pain yields six diverse clusters with pathological shape classes. Cluster 5 represents most of the subjects and is similar to the global mean shape. Cluster 1 and 2 represent a *hyperlordosis*, a strong curvature of the lumbar spine, while cluster 3 and 4 represent a more straight shape. A *multivariate analysis* using the pivot tables puts gender and strong back pain in context to cluster affiliation (Fig. 42 b). It shows that subjects in cluster 1, 2 and 6 reported strong back pain (cluster 1 14.57%, cluster 2 15.08%, cluster 6 19.57%, compared to 6.74% - 13.13% in the other clusters), while at the same time they also have a considerably higher share of females (cluster 1 81.46%, cluster 2 65.87%, cluster 6 70.06%, compared to 20.79% - 53.08% in the other clusters). To check for unusual correlations, the expert used the *Cramér's V* contingency table. It depicted strong associations with *body fat* (0.32), *body weight* (0.3) and *high blood pressure* (0.27) for cluster 1, *alcohol consumption* (0.32) and *attentiveness disorder* (0.28) for cluster 2, and *strong need of sleep* (0.26) for cluster 6. For the experts, these observations are a starting point for a number of new hypotheses about possible relationships, for example the association between overweight and cluster 1.

In summary, it can be stated that the hypothesis-driven analysis leads to hypothesis generation by design of the framework. It is not suited and intended to statistically validate hypotheses. It rather triggers the analysis of potentially associated variables with a pathology of interest.

FURTHER STATISTICAL ANALYSIS OF THE OBSERVED RELATIONSHIPS The following analysis is available open source `R Markdown` document.[9] The evaluation shows that there are no *obvious* relationships between spine shape and back pain. It, however, yields several features that are associated with certain subgroups derived through shape-based clustering. These features are now assessed in a follow-up statistical analysis for potential relationships with *back pain*. The data basis is the same subset of subjects, namely those who comprise a detection model for the lumbar spine (2,540 subjects). The target feature is always the binary (dichotomous) answer to the question "*Did you experience back pain in the past three months?*" and is referred to as *back pain* in this paragraph. The analysis was carried out for *all subjects*

9 http://paulklemm.github.io/StatisticalReview/VAST14_Statistical_Review.html
https://github.com/paulklemm/StatisticalReview

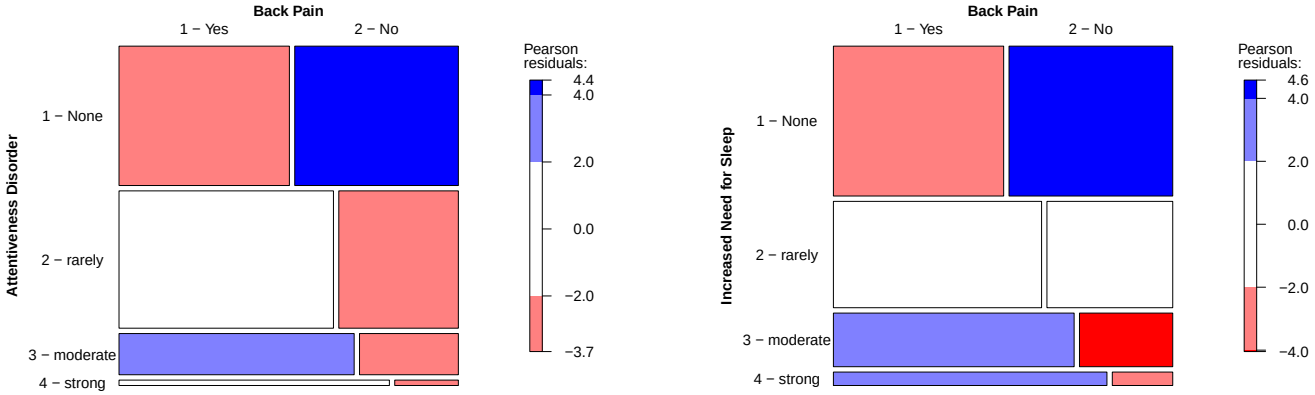


Figure 43: Mosaic plot of *back pain* and *attentive disorder* and *increased need of sleep*. Subjects with higher *attentive disorder* and *increased need of sleep* report *back pain* more frequently. The colors represent the residual level of each combination. Blue means that there are more observations in a cell than expected given a null model (referred to as independence of the features). Red means there are fewer observations than expected. Therefore, subjects with no *attentive disorder* and normal *sleep behavior* seem to have less reports of *back pain* as expected, while subjects with *attentive disorder* and increased required sleep show also increased *back pain* reports.

as well as *subjects older than 60 years* due bone erosion in the spine vertebrae for older subjects.

In order to assess the association between the continuous feature *age* and the categorical feature *back pain*, an analysis of variance (ANOVA) is conducted. It explains the variance of a target feature (in this case *back pain*) with one or more independent features. In the R implementation of the ANOVA, a linear model is used for predicting each subject group ("*Strata*") depending on the categorical target. An ANOVA calculates parameters called *F statistics*, comparing the variation between the strata. The F statistic can be depicted as $\text{F statistics} = \frac{\text{between-group variability}}{\text{within-group variability}}$. This statistic can be used to see if the between-group variability dominates over the within-group variability. The null hypothesis for this test states that *age* does not influence *back pain* and yields an equal distribution. The alternative hypothesis suggests that the distributions are not equal and suggests a relationship between *age* and *back pain*. The ANOVA yields an F-value of 0.009 with a p-value of 0.923. A p-value below 0.05 is the general standard to accept the alternative hypothesis. Since the p-value of *age* and *back pain* is above this threshold, the alternative hypothesis has to be rejected. Note, that this result is *not* a prove that the null hypothesis is true. Another association metric are the *odds ratios* (OR), which analyze the proportion of two dichotomous features, mostly a risk factor and a disease outcome (see Section 2.1). A odds ratio calculation requires a 2×2 contingency matrix. Therefore, *age* is dichotomized into two groups for the analysis. Subjects older than 60 years have an OR of 0.93 for *back pain*, subjects older than 70 years have an OR of 1.04. Both values are near to 1, meaning that there are only slight differences w.r.t. the *back pain*.

The association of *body fat* with *back pain* can be observed in the data with a high F-value of 24.33 and a very low p-value of $8.64e^{-7}$. For subjects older than 60 years, the effect can be still observed with an F-value of 5.699 and a p-value of 0.0172, but it is less powerful. The World Health Organization

(WHO) does not provide clear obesity levels w.r.t. *body fat* percentage [221]. To determine obese subjects based on *body fat* percentage, thresholds values from Romero-Corral et al. [221] are incorporated, who correlate obesity to cardiometabolic dysregulation and cardiovascular mortality. The chosen cut-off is 22.15% for men and 33.3% for women. These values are incorporated to calculate the OR for obese men and women w.r.t *back pain* based on *body fat* percentage. The OR for obese men to have *back pain* is 1.29, for obese women 1.13. This indicates that obese men suffer more likely from *back pain* than obese women.

The ANOVA of *back pain* and *body mass index*, which is calculated as $\mathrm{BMI} = \frac{\mathrm{body\ weight\ in\ kg}}{(\mathrm{body\ size\ in\ meter})^2}$ [210], shows no correlation in the data with an F-value of 3.354 and a p-value of 0.0672. For subjects older than 60 years, the alternative hypothesis is rejected with an F-value of 0.014 and a very high p-value of 0.906. Similarly, the data does not support a relationship between *body weight* and *back pain* with a low F-value of 0.021 and a p-value of 0.884. The group of subjects older than 60 years shows an F-value of 3.72 and a p-value of 0.0541, which is still above the 0.05 threshold. The WHO defines obesity based on a BMI above 30 independent of gender [281]. The OR of subjects with a BMI above 30 and *back pain* is 1.22 and shows a relationship between obese subjects and *back pain*.

Since *high blood pressure* is a categorical feature as well, the *Cramér's V* value was calculated for the target *back pain*. This value, calculated for all subjects, yields no sign of correlation with a low value of 0.065 and also for subjects older than 60 years with a value of 0.077. Similarly, the *alcohol intake in the last 12 months* shows a low correlation with *back pain* with a *Cramér's V* value of 0.058. Subjects older than 60 years, however, show a low correlation with a value of 0.108. *Attentiveness disorder* shows a low correlation with back pain with a *Cramér's V* value of 0.162 for the whole population and 0.216 for subjects older than 60. Similarly, *increased need for sleep* shows a correlation with back pain with a *Cramér's V* value of 0.186 and 0.204 for subjects older than 60 years. As seen in Fig. 43, the reports of back pain increase as *attentive disorders* as well as the *increased need for sleep*.

Even though the extracted features are influenced by lumbar spine shape, the relation between them and *back pain* is, with exceptions, low. This is, however, not unexpected, since the analysis was carried out on features that are extracted solely based on the shape, and does not include any pain indicators. A thorough analysis conducted in Section 5.1 will provide a deeper analysis of shape-related features as well as their influence on back pain using hierarchical clustering.

Further Feedback and Lessons Learned

Both domain experts rated the *IVA* approach positively. KH emphasized the way the image data are included into information visualizations, which comes much more natural to her due to her background in radiology. Great potential is also seen in communicating insights efficiently using the presented visualizations. Epidemiological publications often present data as lists containing results of statistical analyses. This representation reaches its limit for showing shape variance information for anatomical structures. Here, the epidemiologists see a great potential for communicating influencing factors of non-image data on shape information. The resulting plots can also be part of a final report to graphically underline results. The herein presented data-driven analysis approach may also be supported through videos of analysis sessions, which document the train of thought and rea-

soning process. This highlights the benefits of this approach and also allows to spot potential mistakes and questionable assumptions.

MULTIVARIATE ANALYSIS IS MOST IMPORTANT FOR HYPOTHESIS GENERATION Both experts emphasized the potential of the *multivariate analysis* capabilities of the contingency matrix for gaining insight into a large amount of variables simultaneously. It is also useful to verify established but still controversial risk factors, such as the metabolic syndrome for coronary heart disease and whether the data set provides more suitable risk factors. Creating contingency matrices for subgroups, such as different age bins, can help to characterize the aging process by deriving age-specific risk factors. *Multivariate analysis* can be improved by more ways of brushing the data as well as creating subgroups for comparison as a result of the hypothesis-driven analysis case. Too small variable ranges yielding sparse groups could hinder the calculation of statistical resilient measures, since they require a minimum amount of subjects exhibiting the selected variable ranges.

SEGMENTATION QUALITY IS CRUCIAL KH pointed out a unusual strong similarity of the L3 vertebra throughout the population. She would expect a higher shape variability of the vertebra. The medical explanation is that it represents an angular point of curvature of the lumbar spine. A second explanation is the use of the L3 vertebra as initialization point of the lumbar spine model. The experts also emphasized that associations related to shape strongly depend on the segmentation quality. The lumbar spine model used in this case study captures deformation of the spine canal well, but lacks precise definition in vertebrae height and shape. Since deformation of the spine canal is the last stage of pathological lumbar spine deformation and is preceded by vertebrae deformation, the system would strongly benefit from more precise segmentation results capturing these prior changes. For the visual comparison, KH proposed an abstraction of the representation into landmarks, such as centers of the vertebrae and cardinal points of the lumbar spine canal.

USAGE OF DIFFERENT CATEGORIZATIONS DEPENDING ON EXPECTED OUTCOME Categorizing numerical variables into equally sized ranges possibly creates sparse categories due to outliers, for example when analyzing body weight. These outliers are only of high interest for finding pathological subjects. The experts therefore suggested two modes of the tool. The outlier mode still creates categories of equal ranges, producing sparse categories for outliers. Balanced categories are created in the second mode, which uses quartiles or quintiles to set borders between categories.

WEB TECHNOLOGIES ARE WELL SUITED FOR RAPID FEEDBACK The web-based approach for both implementing the prototype and getting feedback via Voice over IP conference calls worked very well. Since the software does not need to be compiled, small changes can even be made on the fly during a testing session. All the data as well as associated medical images remains on the server machine and has not to be moved tediously using external hard disks. This approach is well suited for the *VDAR* approach to assess user thought processes using the think-aloud technique.

4.5.6 *Summary and Conclusion*

An *IVA* framework for the analysis of image-centric epidemiological data was proposed in this section. Hence, the framework allows the hypothesis-driven analysis *and* hypothesis generation. The visualization of multivariate

data using multiple connected views allows users to get fast visual feedback about subject groups. Brushing and linking allows to adapt the data to formulated hypotheses. The use of pivot tables is familiar to epidemiologists while embracing the power of interactive adjustment of the shown variables. The automatic *suggestion* of correlations using contingency methods, such as *Cramér's V*, triggers *hypothesis generation* by highlighting correlations potentially overlooked by the experts. Shape-based clustering assesses the variability of an anatomical structure in the context of non-image variables, such as disease indicators or lifestyle factors.

Epidemiologists are for the first time able to assess shape information of the lumbar spine and its influence on diseases. Findings from analyzing lumbar back pain using the *IVA* approach range from deriving shape-based groups of subjects to detailed descriptions of variables potentially associated with the disease, such as waist circumference, alcohol consumption and attentiveness disorder. The future work regarding this system comprises:

- shape brushing methods to intuitively query subjects using image data,
- the inclusion of more statistical methods and views that are familiar to the epidemiologists (odds ratios, box plots), or
- adapting the shape visualization to explore other organ data with different variance type (such as texture of liver or white/gray matter distribution in the brain).

To reduce the number of false positive findings, the data space can also be randomly cut in half. Then, the hypothesis can be cross-validated for statistical soundness. This requires a large number of subjects, especially if the investigated features are rare and only presented by a few subjects.

As the number of image-centric population studies, participating subjects, gathered variables and imaging modalities rises, and advances towards comparability between population studies are made, the gap between data complexity and analyzability increases. The methods proposed in this section focus on closing this gap, allowing the domain experts to dig deep into the data and potentially obtain unexpected findings.

5

DATA-DRIVEN VISUAL ANALYSIS OF SOCIODEMOGRAPHIC, MEDICAL AND LIFESTYLE FACTORS

This section focuses on IVA methods for analyzing data without a spatial (physical) relation. *Physical views* are absent in the methods presented in this chapter (recall Section 3.2.2). The workflow is specialized to focus on the IVA pattern of *multivariate analysis*, where independent variables and their connections are investigated.

The presented methods tackle different epidemiological problems. Section 5.1 and 5.2 incorporate machine learning algorithms to support hypothesis generation for population study data. Section 5.1 focuses on the analysis of metrics derived from the medical image data, which was also used in the prior chapter. It provides an answer to the question which non-image features can be predicted using solely the information of the image-derived metrics. While this is primarily a method to derive features that are associated with a set of target variables, it can also be used to investigate hypotheses. This can be achieved by dividing the data set according to features of interest (such as different age levels), as well as restricting the set of target features. In Section 5.2, the suitability of different clustering algorithms for epidemiological data is investigated.

Section 5.3 and 5.4 employ overview visualizations for feature correlations to derive insight into the data. Section 5.3 employs a purely visual method for conducting an explorative hypothesis-free analysis of a data set by providing interactive plot matrices. The 3D regression heat map presented in Section 5.4 employs an overview visualization using regression models of pairwise feature correlations with a specified target feature, which mostly indicates a disease or condition. The target can also be any feature in the data, but in epidemiology the target is usually a disease indicator. By providing means of adapting the underlying regression formulas, hypotheses can be integrated into the system, supporting a confirmative analysis.

5.1 DECISION TREE QUALITY PLOT

This section is based on

> **Paul Klemm**, Sylvia Glaßer, Kai Lawonn, Marko Rak, Henry Völzke, Katrin Hegenscheid, and Bernhard Preim. Interactive Visual Analysis of Lumbar Back Pain - What the Lumbar Spine Tells About Your Life. In *Proc. of Information Visualization Theory and Applications*, pages 85-92, 2015.

Marko Rak provided the detection masks of the lumbar spine. Katrin Hegenscheid and Henry Völzke provided the data as well as the epidemiological background knowledge together with associated problems and hypotheses. Sylvia Glaßer, Kai Lawonn and Bernhard Preim provided fruitful discussions for developing the idea and concept of the publication as well as the incorporated error term.

The methods presented in this section were developed to conclude the analysis of the spine data set, which was already incorporated to validate the methods developed in the prior chapter. It aims to determine which features can be directly correlated with the information encoded in the detection

masks. The resulting method and the IVA method displaying the results, however, can be employed to any kind of data set.

The initial goal is to extract possible associations between spine shape and back pain characteristics. For this purpose, classification algorithms are combined with data visualization techniques. Then, Interactive Visual Analysis (IVA) highlights mutual dependencies between image-derived data and back pain-related variables. The focus lies on highlighting new correlations and triggering *hypotheses generation* rather than statistically validating complex epidemiological correlations. The contributions of the method presented in this section comprise:

- an IVA workflow for back pain analysis based on image-derived variables of 2,240 subjects,
- the identification of lumbar spine shape properties potentially associated with back pain,
- the detection of associations between image-based, socio-demographic and medical variables for *hypothesis generation*,
- the identification of the most important variables via classification methods using a novel *Decision Tree Quality Plot*.

5.1.1 *Data Preprocessing*

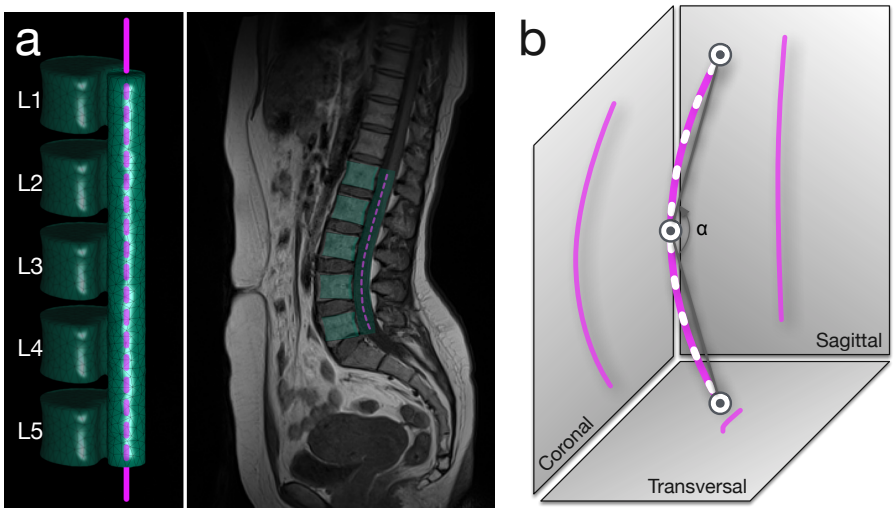


Figure 44: (a) Finite element model (FEM) of the lumbar spine (left), capturing the L1-L5 vertebrae and the lumbar spine canal (right). The purple dashed line describes the lumbar spine canal centerline with 92 points. (b) The weighted sum of *curvature* and *torsion* is extracted for all 92 points (white dashes) and the *curvature angle* (α) for each projection axis to assess their information gain. Image from [294].

All categorical variables are converted into binary *dummy variables*, indicating the presence or absence of a categorical variable manifestation. For example, a pain indicator variable ranging from *1 - no pain* to *4 - large pain* is transformed into four dichotomous variables to determine which manifestation can be described best using the image-based variables.

Foundation for the image data analysis are the finite element models derived through the method of Rak et al. [213]. The centerline representation, which was derived in Section 4.4, was applied to the models. The following metrics were calculated from the model (Fig. 44 b) using the Frenet frame [76]:

- *Mean Curvature* is defined as the average curvature of all points describing the centerline: $\sum_{i=1}^{I} \frac{curvature_i}{I}$. The mean curvature is referred to as *curvature*.
- *Mean Torsion* (how sharp a curve is twisting out of the curvature plane) is defined as the average torsion of all points describing the centerline: $\sum_{i=1}^{I} \frac{torsion_i}{I}$. The mean torsion is referred to as *torsion*.
- *Curvature angle* α is the angle defined by the middle point of the spine canal centerline as *vertex* and the line connecting middle point and top/bottom point as *sides*.

These metrics are also extracted in the sagittal, coronal and transversal projection of the model, yielding 9 image-derived variables. In the next section, the conducted experiments that assess the influence of the lumbar spine shape to lower back pain are presented.

5.1.2 *Experiments and Preliminary Results*

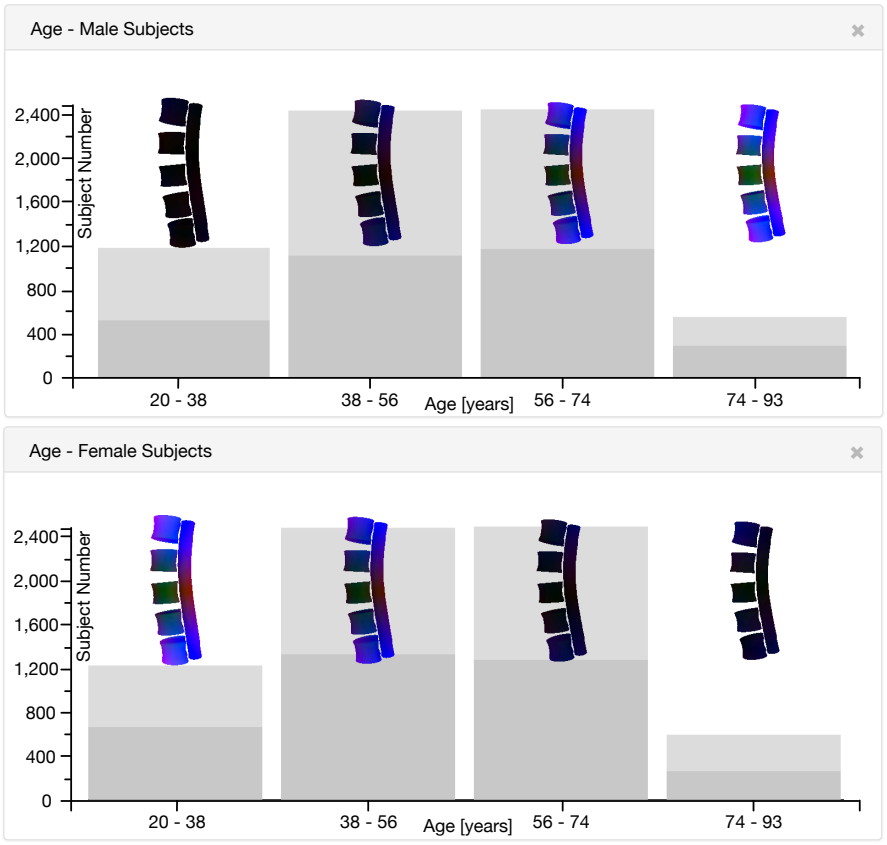


Figure 45: Correlation of *age* and *gender* regarding the lumbar spine size visualized with the bar chart augmented with the image data presented in Section 4.5. The bar chart shows subjects divided into four different *age* groups (x-axis) and their subject count (y-axis). Each bar contains the mean lumbar spine of the respective group. The shape color encodes the distance (red for x-axis, blue for y-axis, green for z-axis) to the overall male (top chart) or female (bottom chart) mean shape. The dark gray share of each bar encodes the portion of male (top chart) or female (bottom chart) subjects. Image from [294].

In this section, the image-derived variables are analyzed w.r.t. the dichotomous *back pain* indicator using a `generalized pairs plot` and all non-image variables by determining correlations. Spine shape is influenced by

several somatometric variables. Larger people (independent of gender) also have a longer spine with a straighter shape. Since men are on average taller than women and people of old age shrink due to *bone erosion*, *gender* and *age* are also risk factors (Fig. 45). Women have a higher life expectancy than men and hence a higher share in the old age group. Also, women are on average smaller than men, hence the larger shape similarity with older subjects. Large *body weight* increases the spine load, resulting in a bent shape. To assess their influence, they are taken into account when spine *curvature* and *torsion* is correlated with non-image variables. Since the *gender* encodes *body size*, subjects are divided into *body size* groups. Discretizing metric variables using quartiles avoids small outlier groups.

GENERALIZED PAIRS PLOT ANALYSIS As first experiment the shape variable was correlated with the dichotomous back pain indicator using a *generalized pairs plot* (recall Sec. 3.1.1). The metric image-derived variables are pairwise visualized using scatter plots on the left side of the matrix diagonal. The combination of the image variables with back pain is visualized as histogram in the last row and as box plot in the last column. The projections to the transversal planes attract attention as they have many outliers. The conclusion is that *curvature* is not as reliable on the transversal plane as it is on the other planes, which was also confirmed through a principal component analysis (see supplementary material at `ivapp15.dnsalias.com`). The *generalized pairs plot* shows similar distributions of subjects with or without back pain with respect to the shape variables in all sub-plots. The plot is discussed as part of the visualization of population study data using plot matrices in Section 5.3 and can be seen in Figure 53.

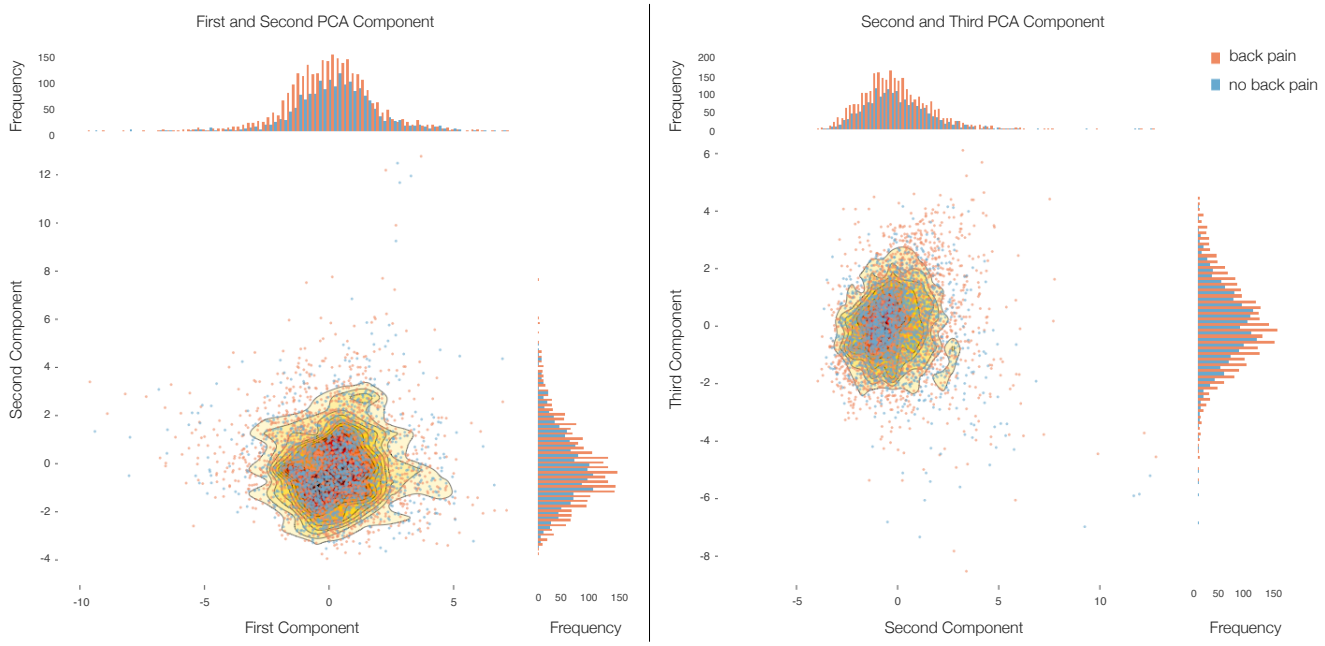


Figure 46: Scatter plot with augmented bar charts for combinations of the first three principal components of the image-derived features. The left plot shows the combination of the first and second principle component, the right plot shows the second and third component. No clear distinction can be made.

PRINCIPAL COMPONENT ANALYSIS The analysis of the principal components of the image-derived features support the observations made through the *generalized pairs plot* analysis. The first three principal components of the

features cover 75% of the variance in the nine image-derived features (recall Section 5.1.1). As seen in Figure 46, no clear distinction between back pain and no back pain can be found.

HETEROGENOUS CORRELATIONS The focus was then expanded on correlations of image-derived variables with all other non-image variables. Different correlation metrics depending on the type of the individual variables are used to derive correlations between all variables in the data set. The method uses the following correlation metrics for the different type combinations:

- *Pearson product-moment* for two continuous variables,
- *Polyserial* correlation for one continuous and one categorical variable, and
- *Polychoric* correlations for two categorical variables.

All values are scaled between 0 (no correlation) and 1 (perfect correlation). About 67 variables are too sparse for calculating correlations (less than , for example *treatment of diabetes* or *medication against high blood pressure* are omitted, since they are not statistically resilient. The resulting *contingency matrix* is displayed using a heat map, encoding correlation values with color brightness, with white for 0 and dark blue for 1. The contingency matrix is calculated for all size groups and searched for correlations between image- and non-image variables. The resulting contingency matrices show no strong correlation with image variables (see experiments page at `ivapp15.dnsalias.com`). Only weak correlations could be found for *mean curvature* with *gender* (0.42), *body size* (0.39) and *number of born children* (0.29). One surprising result was the small correlation of *torsion* with *Parkinson's disease* (0.24). Other than that, *torsion* correlated with almost no variables (values between 0 and 0.05).

Figure 47 shows a heat map representation of all numerical features of the data set derived through the Pearson correlation coefficient. It underlines the strong correlations between the image-derived metrics themselves. The heat map for all features is not shown, because it is hard to interpret without the support of interaction to identify the feature combination of each tile in the view. The observations derived through this analysis lead to the decision to incorporate classification techniques to assess the influence of the image-derived variables.

5.1.3 *Interactive Decision Tree Quality Plot Design*

As described before, correlation coefficients fail to infer back pain status based on lumbar spine canal *curvature* and *torsion*. The plot relies on predictive classification to obtain a complex rule set on how combinations of the image variables explain non-image variables. Decision trees are used to create predictive models. These models are built w.r.t. all input variables and capture more complex relationships than correlation coefficients. Leafs of a decision tree represent class labels, branches represent variable conjunctions leading to the class labels. Decision trees are easy to understand and to read. Too many branches impose *overfitting* to the data [175]. As a rule of thumb, we'll consider a tree with more than 10 rules overfitted.

THE C4.5/C5.0 ALGORITHM The C4.5 algorithm builds decision trees based on information entropy on a data set. Such a calculation requires a numeric or categorical target variable. C5.0 is developed to produce smaller decision trees than C4.5 and to improve the execution time. Here the `R` implementation of C5.0 [140] is used. Categorical attributes with more levels

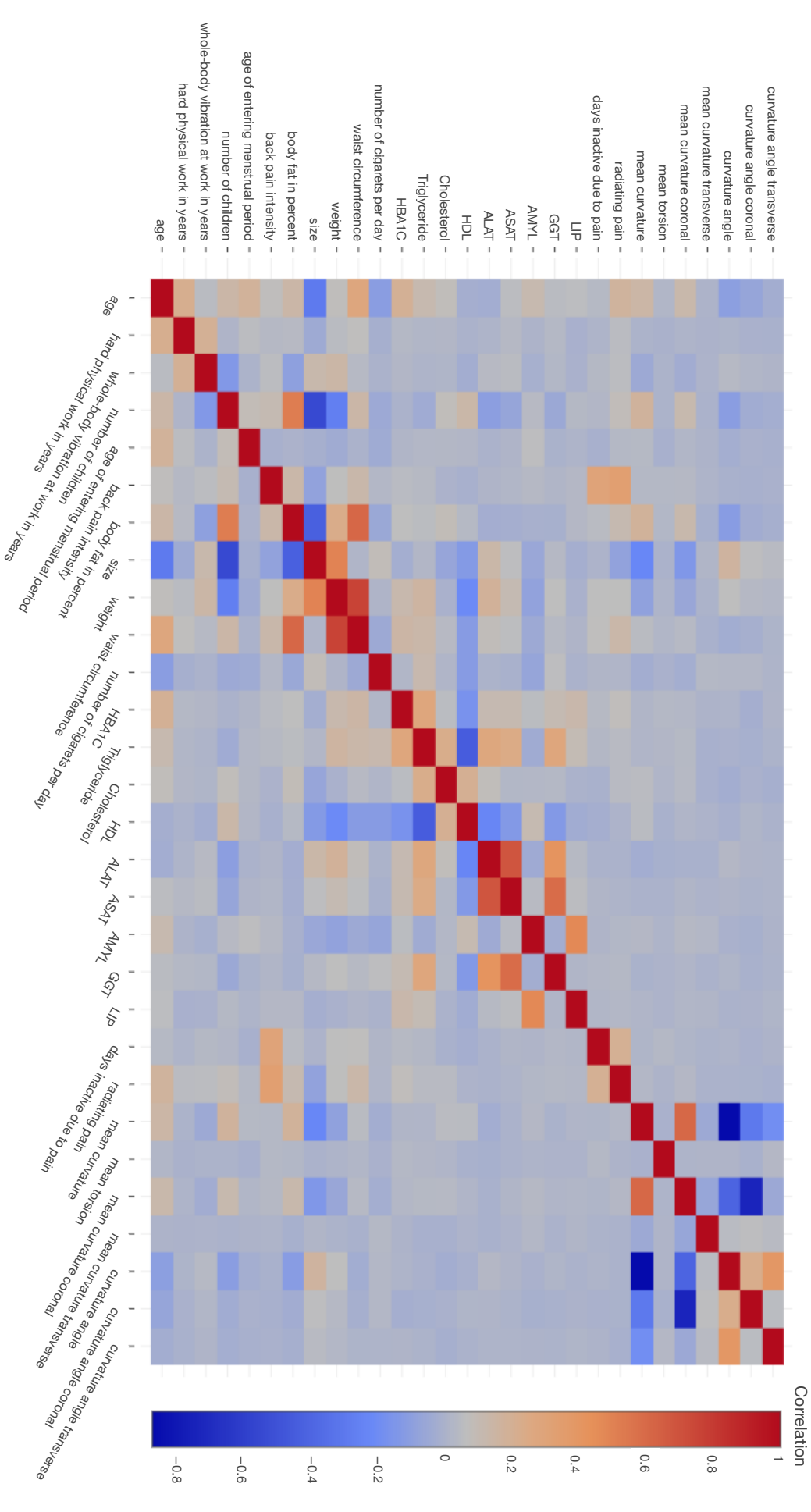


Figure 47: Heat map of Pearson's correlation coefficients between numerical features of the spine data set. The image-derived metrics correlate among each other and with body size and weight as well as the number of born children.

are biased with more information gain in a decision tree [56]. Creating a dummy variable by converting each manifestation of a categorical variable into a dichotomous variable bypasses this problem. In the following analysis, the focus lies on the complexity of decision trees and the classification accuracy.

CREATING DECISION TREES The decision tree has to be created for every non-image variable to analyze which one of them can be predicted by image-derived variables. Note that the target features do *not* have to be dichotomous in order to be classified using a decision tree. Since 134 non-image variables are available, the calculation yields the same amount of trees. Further subdivision, e.g., by quantiles of *body mass index*, increases the number to 536 trees. The results of the classification have to be abstracted to keep the mental effort of interpreting the data low. In other words, the results of such a large amount of decision trees have to be abstracted to be comprehensible.

DECISION TREE QUALITY PLOT The Visual Analytics mantra of *analyzing first*, *show the important* and *analyze further* (recall Sec. 3.2.1) acts as guideline for designing the plot. A first analysis step was performed by applying the classification algorithm to the data. The optimal classification uses a few rules to precisely predict the target variable. Therefore, *small trees* with a *low classification error* are desirable. The two measures form the axes for a *scatter plot* of the classification results. This *Decision Tree Quality Plot* is the central element for the interactive analysis of decision trees.

THE ERROR TERM Calculating the mean classification error is imprecise for non-uniform distributions. For example, if a variable indicating a disease is negative for 90% of the subjects and the classifier simply assigns all subjects to *not ill*, it yields a mean error of 10%, even though it is very imprecise. Based on this, a summary error based on the weighted mean is incorporated, which incorporates the discriminative power of each manifestation and is denoted as follows:

$$totalError = 1 - \sum_{m=1}^{M} \frac{correctlyClassified_m}{M \cdot N_m} \tag{9}$$

M represents the set of manifestations of each variable. N_m denotes the number of subjects in manifestation m. The error represents the share of incorrect classifications and denotes perfect classification with 0 and always wrong with 1. Only results below 0.5 are displayed, a value below 0.25 represents a good classification. It allows for comparability of error rates between variables with different manifestation count.

ATTRIBUTE MAPPING The Decision Tree Quality Plot axes are defined by *tree size* and the previously described *error metric*. This allows to visualize a multitude of classification results in one plot. Classification and comparison of variables for subject groups (e.g., male and female subjects) in one plot can be achieved by color coding group affiliation on the data points. Many variables are sparse, such as *medication of diabetes* or *reason of early retirement*. The classification algorithm may produce higher accuracy for variables with less subjects due to the small sample size, making these results less reliable. Therefore, a way to adjust the minimal number of subjects for each variable using a slider is provided. The initial value is empirically set to 100, marking a good trade-off between sparse variables and statistical informative value. Furthermore, the number of subjects associated with a variable is mapped to the diameter in the Decision Tree Quality Plot. This allows instant reliability

assessment of the result. A square root scale is applied for the tree size axis to highlight decision trees with few decision rules. Outlier results with large decision trees would otherwise distort a plot with linear scale.

DUMMY VARIABLES Dummy variables convert a categorical variable with multiple manifestations into several dichotomous variables. Each dichotomous variable encodes the presence of a manifestation. For example, a pain indicator variable ranging from *1 - no pain* to *4 - large pain* is subdivided into four dichotomous variables. One subject can only have one of these variables set to true. This is useful for the classification, because it allows to determine which manifestation of a variable can be predicted best using the image data variables.

DECISION TREE QUALITY PLOT INTERACTION The visualization provides a good overview of the classification results. *Details-on-demand* are displayed by clicking on an entry in the visualization, which then displays the corresponding decision tree in detail. This allows to sequentially analyze the classifications. Controls for adjusting the maximum classification error and minimum subject count for a variable are provided. This gives the user control to abstract or refine the displayed information. The subject subdivision is controlled by the selection of variables, such as *gender* or *employment status*. Metric variables, such as *body size*, are discretized using their quantiles. This allows users to assess the influence of a variable to the classification process.

IMPLEMENTATION All analyses are carried out using `R`, a widely used programming language for statistical calculations and visualizations. The interactive visualizations are realized using the `ggvis`[1] package. As opposed to the standard `R` plots, `ggvis` allows to adjust visualization variables using user interface controls, such as sliders. In order to make the train of thought comprehensible, `RMarkdown` is incorporated, which allows to create reports by combining `R` with the `Markdown` syntax. `R Shiny`[2] is incorporated to make the report available as dynamic web application. It allows users to combine the power of static `R Markdown` reports with dynamically parameterized `ggvis` plots. Furthermore, calculations based on a prior data selection can be redone within the report. The web-based approach allows users to quickly exchange results with the collaborating epidemiological experts. They can use the technique without installing any software. Exchanging the prototype becomes as easy as exchanging a hyperlink. The prototype is available at `ivapp15.dnsalias.com`.

5.1.4 *Results*

In this section, it is shown which non-image variables can be predicted using the 9 image-derived variables. Subject groups are created to assess the influence of variables affecting the lumbar spine shape. The groups are:

- All subjects,
- subdivision into *male* and *female*,
- subdivision by *Body Mass Index quantiles* ($\mathrm{BMI} = \frac{m}{l^2}$ where m is the *body mass* in kilogram and l is the *body size* in meter), yielding the groups (17, 24.7] (24.7, 27.4] (27.4, 30.5] (30.5, 48],
- subdivision by *size quantiles*, yielding the groups (139, 164], (164, 171], (171, 177], (177, 202].

1 Developed by RStudio, Inc; ggvis.rstudio.com
2 Developed by RStudio, Inc; shiny.rstudio.com

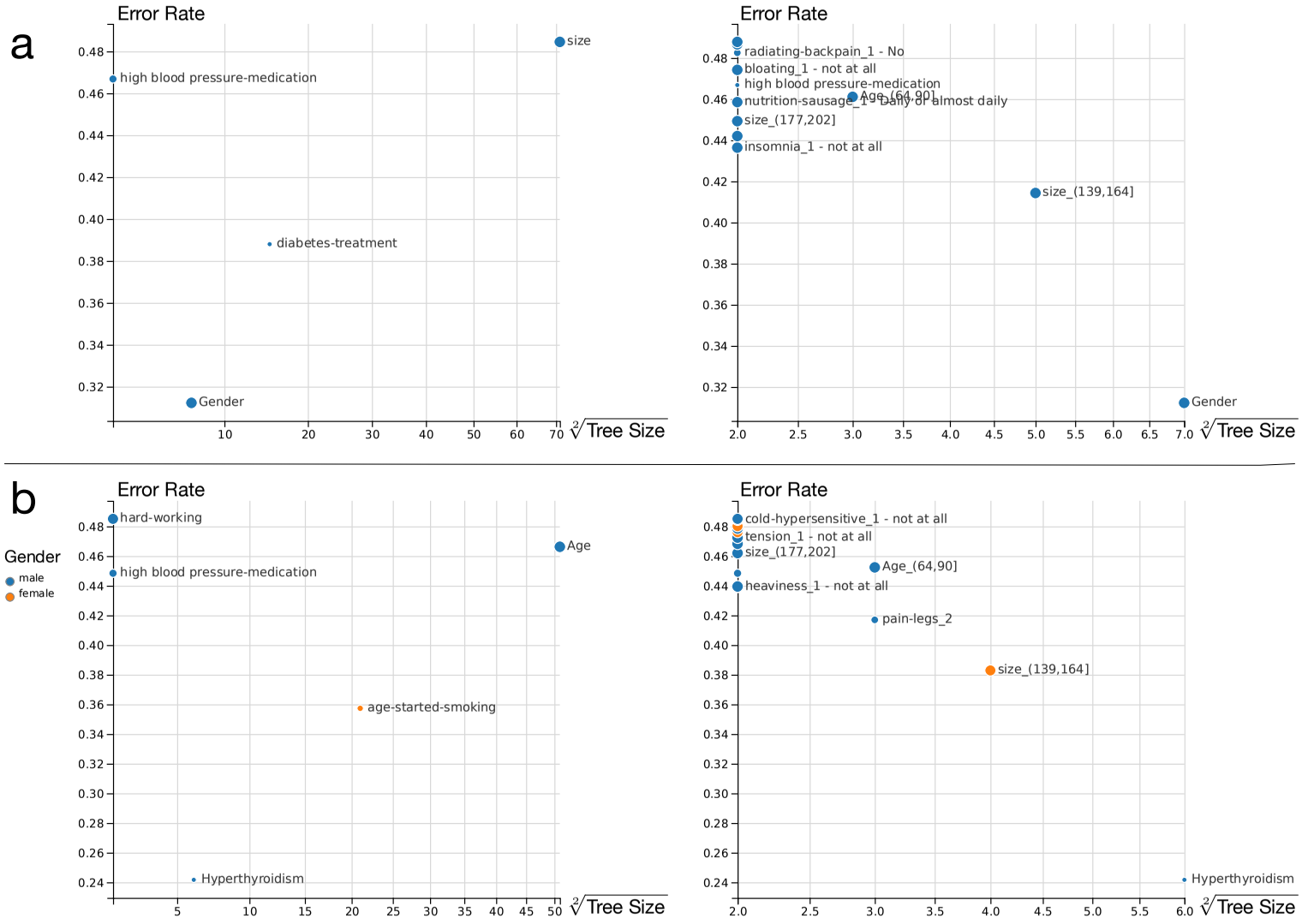


Figure 48: Decision Tree Quality Plot of classification results for all subjects and subjects divided by gender. The x-axis shows the number of decisions of the underlying model, the y-axis the classification error. The left scatter plot shows the results for all variables, either metrics expressed via their quantiles, or categorical. The right scatter plot displays the dummy variables derived from the original variables. Group affiliation of a data point is color-coded for (a) no group and (b) subdivision into *male* and *female* subjects (b). The number of subjects represented in a variable is denoted using the dot diameter. Only variables with an error below 0.5 are displayed. Results above this threshold are inaccurate. The interactive plot (see supplemental material at http://ivapp15.dnsalias.com/) has clickable data points, displaying the corresponding decision tree in a tool tip. Results of high interest are located close to the axis origins. Image adapted from [294].

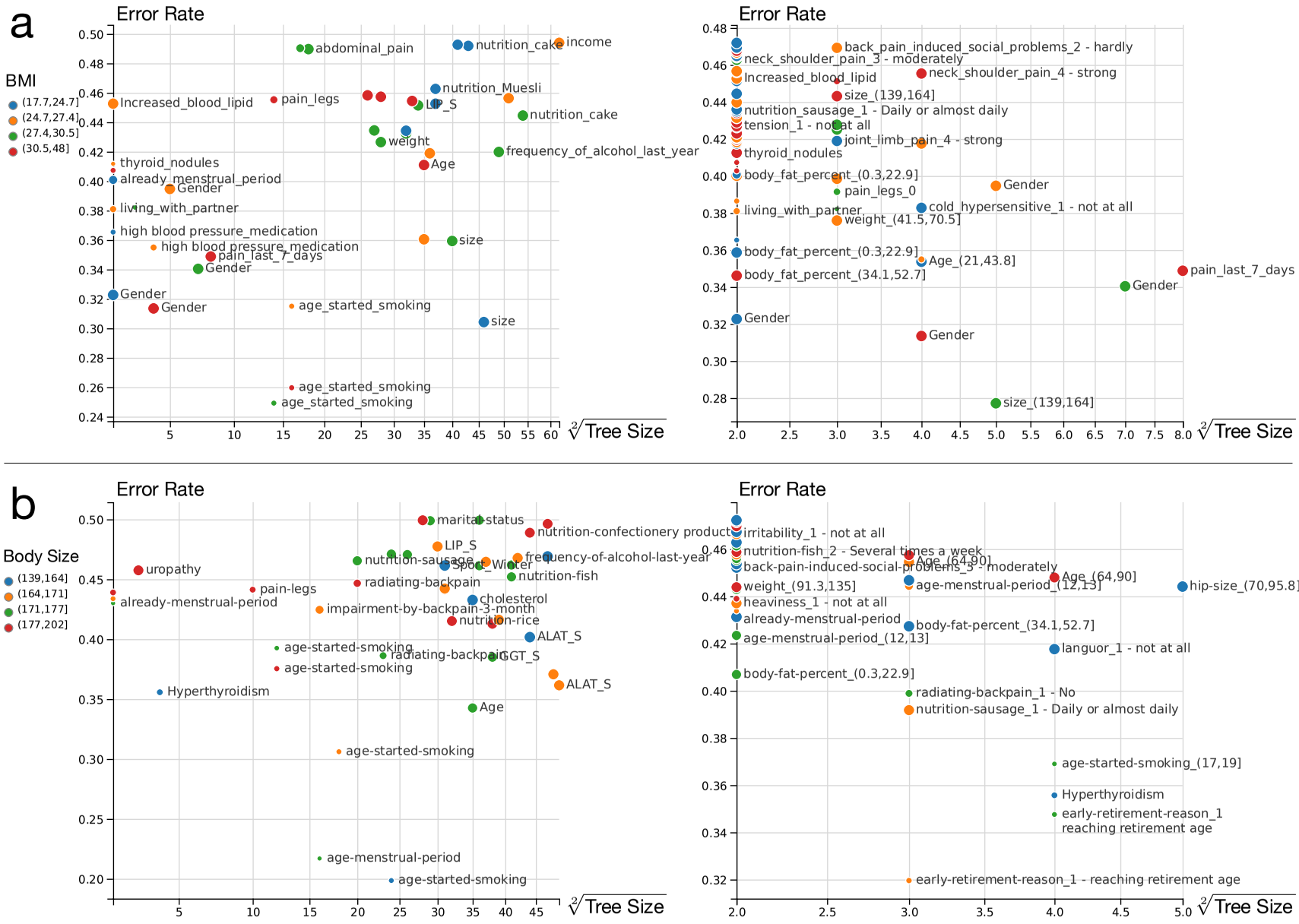


Figure 49: Analogous to Fig. 48, this figure shows the Decision Tree Quality Plot of classification results for subjects divided by quartiles of (a) *Body Mass Index* and (b) *body size*. Image adapted from [294].

Each group is plotted twice. The first plot shows all original variables, the second all categorical variables transformed into dichotomous *dummy variables*. The shown mutual dependencies aim to amplify *hypothesis generation*. Dedicated statistical analysis of these results of this work. The resulting plots can be seen in Figure 48.

ALL VARIABLES Results for all variables can be found in Fig. 48 (a). The majority of non-image variables cannot be automatically classified based on the image variables. This is reflected in the large amount of variables classified with an error above 0.6.

None of the pain indicators can be reliably classified using the image-based variables. The only variable reliably classified in this group is *gender*, which can be classified with an error of 0.31 using 7 rules and incorporates only *curvature*- and *curvature angle*-related variables. The distinction lies in the average difference in *body size* between *male* and *female* subjects. *Medication for high blood pressure* is classified for 1,058 subjects with an error of 0.47 solely based on *coronal mean curvature*. A high share of medicated subjects were correctly classified (796 of 1,058). The majority of non-medicated subjects are false-positive classified (262 of 1,058), yielding a poor quality of the classifier w.r.t. epidemiological research. The four *body size* groups could be classified with an error of 0.48, but the decision tree comprises 71 rules and imposes overfitting. The dummy variable analysis yields a result similar to the *blood pressure medication*. Variables, such as subject size 139-164 cm, between 64 and 90 years of *age* or *nutrition*-related variables are dominantly populated by one manifestation. The classifier neglects the other groups and yields an error below 0.5.

GENDER GROUPS Results for subjects divided by *gender* can be found in Fig. 48 (b). Classifications using groups divided by *gender* do not produce satisfying results. Only *hypothyroidism* could be classified for male subjects with an error of 0.24 for 110 subjects using the *mean curvature* and *curvature angle*. Since there are only 30 male subjects diagnosed with *hypothyroidism*, the statistical power of the result is reduced. The dummy variable analysis showed that female subjects of 139-164 cm *body size* could be discriminated using the *mean curvature* and *curvature angle*, with an error of 0.38.

BODY MASS INDEX GROUPS Results for subjects divided by *BMI* can be found in Fig. 49 (a). *Gender* could be classified for each *BMI* group using *mean curvature* and *curvature angle*. The error varies between 0.31 (*BMI* of $30.5-48$, 4 decision rules) to 0.39 (*BMI* of $24.7-27.4$, 5 decision rules). The starting age of smoking could be classified well with an error between 0.25 and 0.32 for all *BMI* groups, except for subjects with a *BMI* of $30.5-48$. The result is overfitted to the data due to tree sizes between 14 and 16. Some variables, such as *body size*, can be classified with an error of 0.3 to 0.36 using large decision trees with over 20 rules. Using mostly *mean curvature* and *curvature angle*, the *leg pain level* can be classified using 14 rules with an error of 0.46 for obese subjects (*BMI* higher than 30). This result also imposes overfitting. Subjects experiencing *pain in the last seven days* can also be classified for this group using the same variables with a tree consisting of 8 rules and an error of 0.35. Obese subjects are prone to *back* and *leg pain* due to a more stressed lumbar spine. The stress-induced spine deformation seems to directly influence the pain levels for these subjects. The dummy variable analysis shows many results using a decision tree with one rule based on *mean curvature* or *curvature angle* with an error between 0.35 and 0.47.

SIZE GROUPS Results for subjects divided by *body size* can be found in Fig. 49 (b). Many previously described results are influenced by *subject size*. Differences between *male* and *female* subjects can be explained by the average *body size* difference. For example, large subjects are already characterized by their rather straight spine. The question is whether the inter-group spine variability parameter is sufficient for predicting other variables or not. Dividing subjects into *body size* groups potentially highlights classifications not influenced by *body size*.

Large Decision Trees. Back pain-associated variables can be predicted for various *size* groups, but universal rules could not be extracted. *Radiating back pain* could be predicted with an error of 0.39 using 23 rules for subjects between $171 - 177$ cm *body size* with *torsion* and *mean curvature*. For subjects sized $177 - 202$ cm the error increases to 0.47 using 20 decision rules. There are several decision trees for laboratory values, e.g., *alanine aminotransferase* value (relevant for diagnosis of liver or gallbladder illness) in the blood can be predicted with an error of 0.4 ($139 - 164$ cm) to 0.36 ($164 - 171$ cm). Similar values can be observed for *cholesterol* or *age*. Due to the large decision trees, these results are not usable and impose overfitting to the data.

Small Decision Trees. The dummy variables show several variables predicted using only one decision rule with an error between 0.42 and 0.47. Most of these variables have a dominant manifestation and the classifier shows a low detection precision for the second manifestation. These variables include *nutrition*, *thyroid disorder* and *social problems induced by back pain*.

5.1.5 *Summary and Conclusion*

This section provided a comparative analysis method of decision trees independent of the variable manifestation count using a novel *Decision Tree Quality Plot*. The method is applied to gain insight into the predictive power of 9 image-derived variables for 134 non-image variables with focus on *back pain*. The analysis was performed for subject groups of *gender*, *BMI* and *body size* to assess their influence on the lumbar spine shape.

The methods presented herein may be applied to comprehensive epidemiological data sets to investigate mutual dependencies among variables and to generate hypotheses on potential associations and subgroups. These hypotheses, however, have to be substantiated by dedicated statistical analyses and replication in independent populations.

PREDICTIVE POWER OF IMAGE-DERIVED VARIABLES The presented results indicate that *torsion*, *curvature* and *curvature angle* of the lumbar spine at the presented precision are not sufficient to predict lumbar back pain in the SHIP data set. This method allows to assess their discriminative power, which is largely limited to separating male and female subjects, *nutrition* variables as well as different disease indicators. The C5.0 algorithm proved to be an effective tool for evaluating a set of derived metrics regarding their suitability to classify non-image variables. Over-fitting to the data indicated by complex decision trees has to be taken into account as well. The presented method only captures linear relationships between variables. To take more complex associations into account, methods such as regression analysis have to be incorporated.

APPLICABILITY Methods supporting *hypothesis generation* based on image information are new to the application domain. They are an addition to the standard epidemiological workflow, as they highlight new and possibly unknown relationships. Classification methods based on decision trees

have proven to be useful for assessing the discriminative power of a variable set. Their ability to consider variable combinations makes them more powerful than correlation coefficients calculated for each variable. This advantage comes with a much more complex output, the results are more challenging to assess and to abstract. The method to plot derived metrics and custom-tailored error measures proved to be effective. Huge result spaces could be navigated fast using the Decision Tree Quality Plot. Therefore, the method is applicable not only for deriving information based on image data, but on all *potential target variables*.

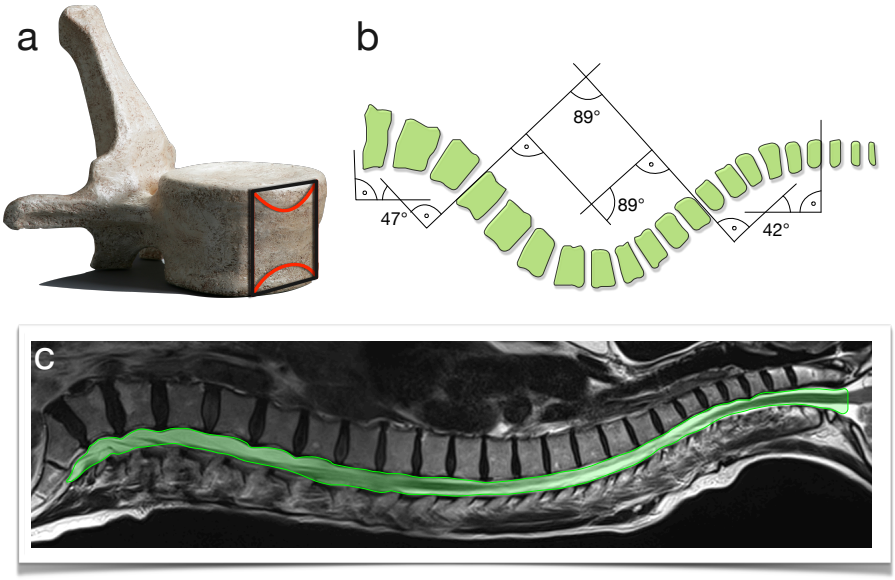


Figure 50: Potential features for covering detailed aspects of pathological spine deformations. (a) Dented vertebrae are a sign of heavy stress and bone absorption. (b, c) The spine canal shape can be used to characterize scoliosis (curvature sagittal) and lordosis (curvature coronal) as well as the *Cobb angles* [43]. (c) Spine canal thickness is associated with herniated disks.

PROPOSED FEATURES Pathological deformation of the lumbar spine is usually the last stage back pain. The deformation is associated with very strong lower back pain. Earlier signs of pathological change have to be captured in order to derive better predictive models.

A very early sign of pathological deformation is the *bone resorption in the center of a vertebra*. It changes from being block-shaped to be dented in the center (Fig. 50 a). This information can be obtained by segmenting the whole vertebra or the top and bottom point of each vertebra center. Another valuable variable would be the *spine canal thickness* (Fig. 50 c). Low spine canal thickness can be an indicator for an impending herniated disk. Both surface texture of the vertebra and thickness of the spine canal are used to diagnose herniated disks. The *overall spine canal shape* is also of interested, since scoliosis and lordosis can be characterized more precise by deriving the Cobb angles [43] from this shape (Fig 50 b, c).

OUTLOOK Combining the power of statistical analysis, visual analytics and classification techniques is essential for analyzing increasingly complex heterogenous population data. These methods do not aim to replace the traditional epidemiological workflow, but rather complement the weak points of standard statistical methods. The Decision Tree Quality Plot provides a novel way to gain insight into these complex data sets and amplifies *hypothesis generation*.

In the following section, additional clustering methods are presented, which are well suited for application in population study data.

5.2 CLUSTERING OF POPULATION STUDY DATA

This section is based on

> **Paul Klemm**, Lisa Frauenstein, David Perlich, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Clustering Socio-Demographic and Medical Attribute Data in Cohort Studies. In *Proceedings des Workshops Bildverarbeitung für die Medizin*, pages 180-185, 2014.

The implementation of the clustering techniques was carried out by Lisa Frauenstein and David Perlich as a scientific student project. Katrin Hegenscheid and Henry Völzke provided the data and the required epidemiological background knowledge. The concept was developed together in discussions with Lisa Frauenstein, David Perlich and Bernhard Preim.

In this section, the exploratory data analysis approach is enhanced by automatically generating subject groups using clustering algorithms based on non-image and image-derived data. The basis of this work is the lumbar back pain data set presented in Section 4.1. Instead of grouping subjects based on the shape of the spine, as presented in Section 4.4, the clustering is now focused on all non-spatial information about each subject. Deterministic cluster results are a major requirement to ensure statistical resilience. Clustering subjects aims to reveal undiscovered correlations.

This section covers a feasibility study of clustering techniques for the SHIP data set. The methods developed herein were *not* tested in analysis sessions with epidemiologists. The black box character of clustering techniques is not popular with epidemiologists, as they are primarily interested in characterizing relationships. The contributions comprise:

- Assessing three clustering methods (k-Prototypes, DBSCAN and hierarchical agglomerative clustering) for their suitability in population studies,
- Incorporating the clustering methods in a web-based Visual Analytics framework for browsing population study data.

5.2.1 *Clustering Workflow and Prototype*

In this section, a Visual Analytics prototype is described, which is used to analyze the clustering results, followed by a brief overview of the incorporated clustering methods.

VA PROTOTYPE To explore the clustering results, a Visual Analytics system was developed, which comprises multiple views for ordinal and metric variables and supports brushing and linking. The web-based application was implemented using HTML5, CSS and Javascript/jQuery with support of D3 [24]. The user can select a set of variables from a categorized list, similar to the method presented in Section 4.5, and add them onto the canvas area. The prototype of the system can be seen in Figure 51.

To trigger the clustering, the user selects either all parameters of a data set or a subset from a list. Due to missing values, the system immediately displays the number of subjects that are omitted in the clustering step given the current attribute selection. The selection of the clustering method and its parameters closes this process, which returns computed groups that are

Figure 51: The clustering result is embedded within the Visual Analytics framework. It displays 2,333 subjects of the `SHIP-2` cohort from the spine dataset and contains 179 features. A k-Prototypes clustering with $k = 3$ results in three color-coded clusters. All subjects with body weight above 120 kg are brushed using parallel coordinates and highlighted in the scatter plots with red circles. One subject of cluster two is selected in the list view, which increases its opacity in the parallel coordinates and its radius in the scatter plots. Image from [292].

rendered as seen in Figure 51. The user can re-run the clustering at any given time and can also add new plots to further investigate feature associations.

CLUSTERING METHODS Clustering methods divide the space spanned by data elements so that it maximizes the distance between groups and minimizes the within-groups variance [93]. Characteristic for population study data is that not every assessed subject has data for every attribute. The clustering process needs to account for missing data, as described in Section 4.1. This problem is tackled by displaying the number of omitted data elements upon the current attribute selection designated for clustering.

Measurement of Distance. Clustering heterogeneous data attributes at the same time requires distance measurements that consider different data types [114]. The similarity between numerical attributes is calculated using the Euclidean distance. Ordinal attribute values are compared in a binary fashion, having distance 0 when they are identical and distance 1 otherwise. The factor γ can be used to weight elements [114]. Three different clustering techniques were applied. Each clustering technique was chosen because it shows unique characteristics, which potentially makes them suitable for epidemiological analysis.

k-Means and k-Prototypes. The algorithms k-Means and k-Prototypes were chosen because they are well suited for numerical features due to their Euclidean distance function. Dividing the data into k clusters using randomly generated centroids, each data point is iteratively attached to its closest centroid. K-Prototypes [115] enhances k-Means to allow for the clustering of ordinal and scalar attributes using the previously described weighted distance. A centroid can additionally be described for each attribute by the most common value of its cluster. The random initialization of centroids renders the k-Prototypes clustering results non-deterministic. This is not suit-

able for epidemiological applications where reproducibility of all results is required. Therefore, the initial centroid positions are computed by placing centroids near values that are close to each other.

DBSCAN. The DBSCAN clustering algorithm was chosen because it potentially finds clustering results of arbitrary shape. *D*ensity-*B*ased *S*patial *C*lustering of *A*pplications with *N*oise computes clusters based on object density. Elements are density-connected when they are reachable by a chain of dense objects. Density-connected elements form a cluster. Outliers are objects that are not associated to a cluster via density. DBSCAN is steered by parameters that define the distance between neighbors (ϵ) and the number of neighbors that a "dense" element must comprise ($\mathtt{minPts}$). The method is independent of a predefined cluster number and accounts for outliers.

Hierarchical Agglomerative Clustering. This clustering technique was chosen, because it calculates clustering results for a large number of clusters and therefore scales well with a dynamic setting of cluster size. The stepwise aggregation of the closest elements into a cluster yields a dendrogram whose levels represent clusters. By varying the minimum similarity, the desired number of clusters is obtained. The method is known to be outlier-prone.

5.2.2 *Results*

The difficulty of comparing cluster results in this application domain is twofold. First, the accuracy of the result cannot be measured due to missing ground truth. Second, the presented clustering methods have different parameters, which have a strong impact on their results. The difference in the results is minimized by focusing on the same numerical and categorical parameters.

Table 4: Dice's coefficients for clustering results of k-Prototypes and DBSCAN.

Cluster Number	Algorithms	Dice's Coefficient
2	$\mathtt{k-Prototypes/DBSCAN}$ ($\epsilon = 1.3$)	0.634
	$\mathtt{k-Prototypes/DBSCAN}$ ($\epsilon = 1.4$)	0.655
	$\mathtt{k-Prototypes/DBSCAN}$ ($\epsilon = 1.5$)	0.657
3	$\mathtt{k-Prototypes/DBSCAN}$ ($\epsilon = 0.9$)	0.720
	$\mathtt{k-Prototypes/DBSCAN}$ ($\epsilon = 1.1$)	0.644
	$\mathtt{k-Prototypes/DBSCAN}$ ($\epsilon = 1.2$)	0.646
6	$\mathtt{k-Prototypes/DBSCAN}$ ($\epsilon = 1.0$)	0.406

K-Prototypes was tested in a range of two to ten clusters. The cluster sizes range from 94 to 487 subjects (from a total of 2333 subjects). No overly large or small clusters are computed.

DBSCAN's parameter $\mathtt{minPts}$ equals the minimum cluster size. Since epidemiologists are interested in larger groups of subjects, this value needs to be fairly high. Ester et al. [66] argue that the impact of $\mathtt{minPts}$ is little above a certain threshold. This value is set empirically to 50, which produces roughly size-balanced clusters. Parameter ϵ defines the size of an object's neighborhood. Set low, ϵ leads to many small outlier clusters, which is not desired. An ϵ-value between 0.6 and 0.8 classifies 1602 subjects as outliers and is therefore not reasonable. Parameter ϵ set to 0.9 to 1.2 results in balanced clusters.

Hierarchical Agglomerative Clustering creates very unbalanced trees for the data. Many clusters only contain one element. Complete-Linkage produced the best results in terms of cluster size, but still yields one large cluster

containing almost all subjects. Hence, this method was discarded for use on the data.

COMPARISON USING DICE'S COEFFICIENT Dice's coefficient [57] is incorporated to compare the clustering results under use of different parameters. It is defined as $\frac{2(A \bigcap B)}{|A|+|B|}$, where A and B are the clusters to compare, and $A \bigcap B$ is the amount of elements in A and B. Dice's coefficient is 0 for disjunct and 1 for identical clusters. Since the hierarchical agglomerative clustering results are not plausible, only k-Prototypes and DBSCAN are compared. The results for clusters with size 2, 3 and 6 for DBSCAN with corresponding k-Prototypes results can be found in Table 4. While Dice's coefficient for 2 to 3 clusters is close to 0.65, it is only at 0.4 for 6 clusters. Cluster results are similar, while there is a decreasing similarity for an increasing cluster number. This reflects the missing ground truth problem– these results are only an expression of similarity, not plausibility. The latter can only be determined in the context of epidemiological reasoning whether the groups represent meaningful correlations.

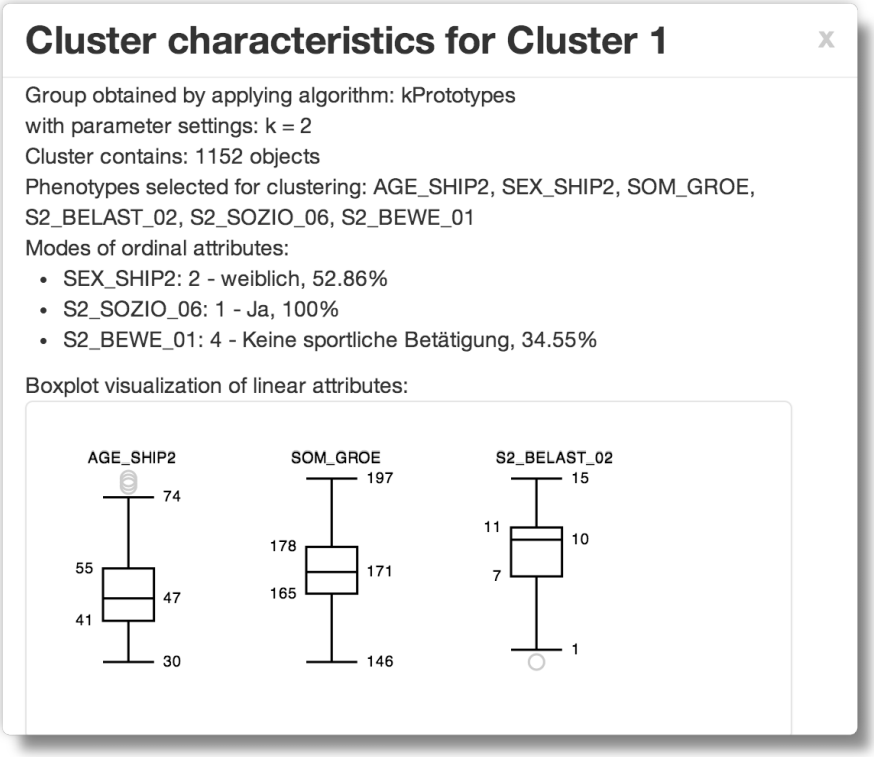


Figure 52: Information window for a clustering resulting from the k-Prototypes algorithm. The clustering parameters yield a reproducible clustering result. The distribution of metric parameters in the cluster is displayed using box plots. The most frequent value of each ordinal parameter is displayed using percentage statements. Image from [292].

VISUALIZATION OF CLUSTERING RESULTS Enhancing the Visual Analytics framework by clustering capabilities for automatic grouping was a key motivation for this work. Each group is rendered using a different color and can therefore be differentiated in the linked plots (Fig. 51). An additional information window is introduced, which contains statistical information associated to each cluster (Fig. 52).

5.2.3 *Summary and Conclusion*

This section focused on three methods for clustering epidemiological population study data to compute groups that capture data interactions. Linked to Visual Analytics systems, these methods provide an alternative way of gaining new insight into the complex interactions in these high-dimensional data sets. The methods k-Prototypes and DBSCAN are appropriate for the data. Hierarchical agglomerative clustering produced unbalanced cluster trees, yielding huge clusters containing almost all subjects and is therefore not suitable for the research. The clustering results are strongly dependent on the chosen variable types and the distance measure. Future extensions comprise better cluster group comparison to amplify hypothesis generation by highlighting influential parameters. Usability would benefit from automatic parameter designation using quality criteria. Missing data can be tackled with imputation [59]. For k-Prototypes, k could be derived by a knee function that plots the cluster number to a cluster quality measurement [227], as used in Section 4.4 and 4.5.

At the end, it falls to the user to validate the data for plausibility. A clustering-based automated grouping step can only highlight certain dependencies in the data set. It is no alternative to the classical epidemiological workflow, but rather an enhancement of the available tools, providing a different point of view.

The problem with clustering non-image data lies in the black box character of the algorithms. Epidemiologists want to characterize relationships that are obscured in the automatic clustering process. Clustering techniques, however, are well suited for an explorative analysis, where the clustering result can be used as input for a visual analysis, which searches for distinctive features.

5.3 PLOT MATRICES

This section provides an interactive plot matrix for use with population study data to assess their suitability for this application domain. Plot matrices, described in the related work under Section 3.2.2, are an efficient method for gaining insight into pairwise relationships of variables. Their structure is similar to a heat map (recall Fig. 47), except that the combination of variables is not encoded as a numerical value which is encoded through color, but as actual plot renderings. Figure 53 shows a *generalized pairs plot* calculated using the GGally package [229] for R to assess the influence of image-derived features of the lumbar spine on back pain. The plot contains many information about feature combination, such as correlation coefficients or scatter plots. The distinction between the groups *back pain* and *no back pain* is made apparent using bar charts, histograms as well as by color-coding the entries in the scatter plots and as individual correlation coefficients. The plot underlines the conclusion that no relationship between the extracted features and back pain can be identified. A generalized pairs plot in the experimental results found under the link `ivapp15.dnsalias.com` can be colored according to a target phenotype, such as back pain, age groups, gender. With this exception, the generated plots are static, hence, items can neither be highlighted nor brushed.

Generalized Plot Matrices (GPLOMs) [116] (recall Section 3.2.2) are suited for categorical and continuous variables. To reduce the complexity of supporting interaction between a large number of different visualization types, GPLOMs are restricted to scatter plots, heat maps and bar charts. GPLOMs, however, are restricted to brushing categories; quantitative variables cannot

be filtered at all. Additionally, only single categories can be filtered and highlighted at a time. Also, the proposed plots do not allow adding or removing variables. Variables can also not be hidden without applying a filter. This means that analysts are not able to pick only certain variables of interest. Focusing without filtering is therefore not possible. Scatter plots included in GPLOMs also often suffer from overplotting for a large number of entries. Since categorical features are mapped on color, each data point cannot simply be drawn semi-transparent, because it would yield a lot of different merged and potentially confusing colors.

5.3.1 *Enhanced GPLOMs*

The implementation of GPLOMs presented in this section does not contain all features presented by Im et al. [116]. For example, bendy highlights, textual search and the infobox (with kernel density estimation) are not implemented. The implementation aims to enhance filtering techniques to allow for applicability of this technique in an epidemiological context. The source of the prototype is available as open source repository.[3]

INTERFACE OVERVIEW The user interface of the enhanced GPLOM prototype is shown in the annotated screenshot seen in Figure 54. The GPLOM is a lower triangular matrix filled with three plot types: heat maps, histograms and scatter plots. Each row and column is labeled according to the variable it belongs to. The variable selection menu in the upper top right corner is used to select, add and remove variables. The current filter criteria are located on the left. Hovering the mouse over various elements shows additional information about the variables and the highlighted entries as tooltips and labels.

Heat maps are drawn for categorical variable pairs. They contain compartments for each combination of categories and color them according to the number of data points found in each. The darker the colors, the higher the frequencies. Hovering the mouse over a compartment shows a tool tip revealing what categories it belongs to and how many data points it contains. The color mapping (lightness) is scaled for each heat map and selection, which means that colors are not comparable between different heat maps or before and after a filter is applied.

Histograms are drawn for a categorical variable plotted against a quantitative variable. The original GPLOM implementation allows to select a defined aggregation function (min, max, average, sum, count). The prototype currently only supports sum aggregation. The bar's height is scaled between zero and the histogram's maximum. Comparing bars between histograms is therefore not possible. Hovering over a histogram reveals the categories each bar belongs to.

Scatter plots are used for two quantitative variables. Their bounding boxes are scaled to each variable's global minimum and maximum. This allows for comparisons between scatter plots that are on the same row or column. Each point is a black circle with a *radial transparency* gradient. If the semi-transparent points overlap, the resulting point is darker, which makes it possible for the analyst to identify overlaps and their magnitudes.

FILTERING AND HIGHLIGHTING Improved filtering and highlighting capabilities impose the major improvement over standard GPLOMs.

Show & hide variables. Variables can be introduced into the matrix using a selection menu. This simple yet effective enhancement allows to reduce the

3 https://github.com/rbyte/F-GPLOM

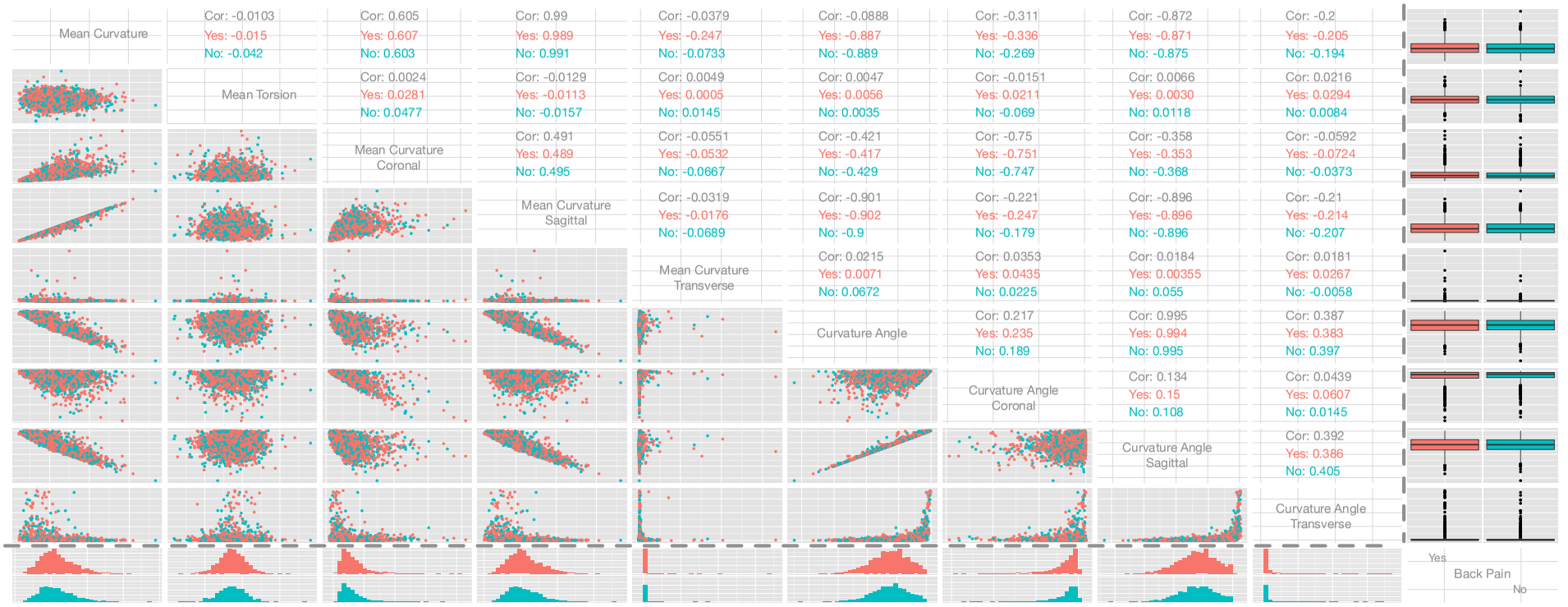


Figure 53: A `generalized pairs plot` of all image-derived variables colored by presence (red) or absence (turquoise) of back pain. Pairwise combinations of image-derived variables are visualized via scatter plots on the left of the matrix diagonal. Their correlation with back pain is denoted to the right of the matrix diagonal. The box plots (right) and histograms (bottom) display the distribution of each image variable encoded with back pain. No correlations with back pain can be identified in this plot.

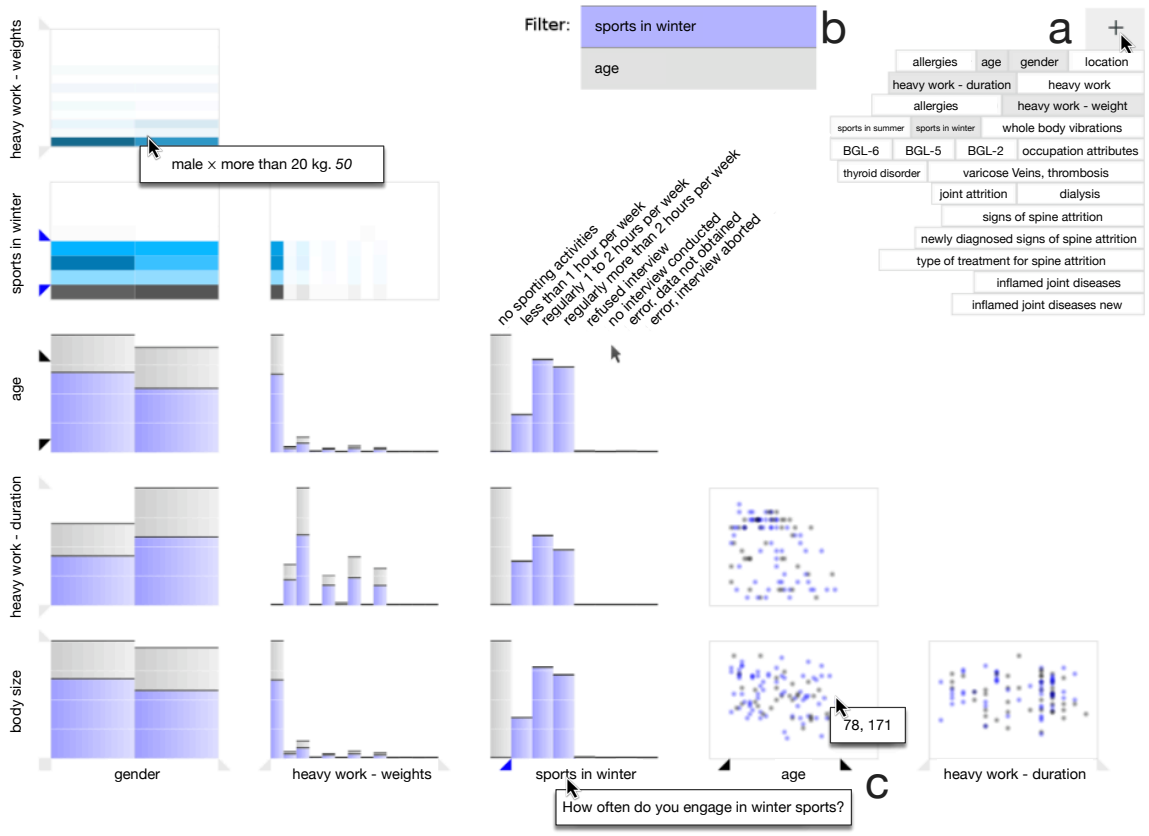


Figure 54: Annotated user interface of the enhanced GPLOM prototype. Five variables (*age*, *gender*, *heavy work - duration*, *heavy work - weight*, *sports in winter*) are selected using the variable selection box (a). A selection is indicated using the blue data items for the lifestyle variable *sports in winter*. The sliders are depicted using triangles. A slider turns from gray to black when it is used to filter the data. The filter status can be seen in the filter bar on the top (b). Only items inside the *age* filter are displayed in each plot (b, c). Data groups as well as individual points can be highlighted using tool tips. Image adapted from [83].

visual complexity of the GPLOMs, which get very cluttered and unreadable with a large number of displayed variables.

Range sliders are used to filter quantitative variables. The sliders allow the definition of an upper and lower boundary for each variable. In other words, they allow to *filter* the data. Values outside of the defined range are excluded. The sliders are also employed for categorical variables to allow for selection of multiple categories for filtering and highlighting. Sliders are provided for each row and column and are located next to the variable labels. All variables, except for the outer ones, are represented on both axes. Hence, a pair of sliders has a corresponding pair of siblings on the adjacent axis, except for the two variables. Pairs of sliders change synchronously. The relationship between rows and columns is visualized using so called bendy highlights in GPLOMs.

Associative highlighting is incorporated by dragging sliders and propagates the selected range over all plots, similar to brushing and linking. Figure 54 shows the variable *sports in winter* being highlighted. Sliders are gray in their default position, a range that includes everything. They turn blue if they are dragged, which visually hints their active state. Points in scatter plots are highlighted blue if they belong to the defined range. Histograms are analogously overlaid by blue bars. Heat maps highlight proportions blue with saturation encoding the proportion.

Two Filter Stages. Highlighted variables are added to the filter criteria list with a blue background colour. The list features two stages: *highlighted filters* and *applied filters* on variables. Highlighted filters can be applied by clicking a blue list item. The background turns gray and the sliders black to indicate the change. In the example in Figure 54, "age" is filtered. Data outside of this filter range are excluded entirely. *Sports in winter* is only highlighted in blue, but still included in all computations. This distinction between highlighted and applied filter criterion imposes two advantages:

1. Multiple criteria can be defined for highlighted selections. This allows analysts to perform fine-grained comparisons between the current selection (gray) and a specified subset (blue). The subset can now contain criteria on quantitative variables and multiple categories from categorical variables.

2. Applied filters can easily be highlighted by clicking the gray list items. Filtered variables are also not hidden automatically. This allows applied filters to be altered by adjusting the black sliders. In addition, context is preserved. Variables can always be included or excluded using the variable selection menu, but this does not affect the filter list, which keeps a record of the steps taken so far in narrowing down the dataset.

5.3.2 *Discussion*

The prototype still has many issues and problems. It does not implement all GPLOM features. For example, bendy highlights could make the relations between columns and rows clearer. A textual search can reduce time needed for a visual search and letting the user change the histogram aggregation function enables important changes of perspective on the data. GPLOMs are not custom-tailored for epidemiological studies. It was assumed that the core requirements can be generalized. This may *not* be the case.

If all variables of the lumbar spine dataset are viewed at once, the GPLOM is too complex to read and takes up large amounts of screen space. The implementation allows for selecting a variable subset. Therefore, no complete overview is provided.

DATASET-RELATED ISSUES The dataset contains **error codes**. They are used to include various meta information, for example, that a question was not answered. In categorical variables, those error codes are simply coded as additional categories. In quantitative variables, fields with error codes are ignored. Both approaches may be inappropriate. Possible solutions are handling error codes separately, ignoring error fields by default or grouping error codes into one category.

Binning of quantitative variables, such as age in years, imposes another problem. Those variables are coded as ordinal categorical, but this does not reflect their quantitative nature. Their high cardinality may make plots hard to read. The higher the cardinality, the smaller is the impact of overplotting. It may therefore be better to encode them as quantitative.

Another property of the dataset are the numerous **dichotomous variables**. Currently, the prototype codes them as nominal categorical variables with cardinality two. A possible improvement may be to use mosaic plots or violin plots instead.

FURTHER ENHANCEMENTS OF THE PLOTS An important drawback are the missing **axis labels** for each plot. Category names, ranges and numbers cannot be retrieved by checking labels and tooltips. Directly including this

information leads to cluttered views. Important key values, like domain boundaries for scatter plots, or the baseline and maximum of histograms should be included in future implementations.

Matrix Diagonal Histograms. Compared to the original generalized pairs plot, our GPLOM does not include a matrix diagonal for displaying a one-variable histogram. Its inclusion may improve the visualization. However, it is important to include visual clues to indicate that those histograms on the diagonal show something completely different than the ones that plot a categorical against a quantitative variable. This can be indicated using a different background color of the plot [263].

Variable Order. The order of variables and nominal categories is another issue worth considering further. Currently, the prototype orders them arbitrarily. It is useful to order variables by similarity, so that proximity reflects similarity. This requires a measure for similarity, such as correlation. Johansson et al. [124] proposed including correlation coefficients into each plot.

ENHANCED HEAT MAPS, SCATTER PLOTS AND HISTOGRAMS **Heat maps** encode the proportion of highlighted items in a compartment to saturation. This imposes a dense coding, since lightness is already used for overall compartment frequency. The main drawback is the difficulty to visually quantify color differences. Another problem is that blue saturation may have an impact on the perceived lightness of a color. In addition, the local scaling renders different heat maps incomparable. However, a per row or global scaling may hide (local) details. If the two variables have a low cardinality, mosaic plots may be a better choice. But this would degrade the advantages of the heat map alignment and render bendy highlights useless. Another suggestion for improving heat maps is to allow highlighting and filtering through hovering, clicking and dragging over compartments and areas.

Histograms are generally easy to read. The data density of the histogram is directly linked to the variable cardinality. This visual data density could be equalized by giving each histogram bar a fixed amount of space. This change would also affect the display of the heat map, which would have an even spacing. The problem with this approach may be that variables with a huge number of categories take up too much space. Limits on the space for each variable could solve this. Due to the local scaling of each histogram's bar height, multiple histograms are not directly comparable. This may be desirable on a per row basis. A general problem of histograms is that they hide the distribution of the data points for each category. This can only be alleviated by using a different plot, for example showing side-by-side the kernel density distributions using a line chart. Providing the analyst with the option to switch between multiple types of plots is worth considering.

The performance of the interface degrades with the number of points plotted in each **scatter plot**. To counter this effect, each scatter plot only shows a random selection of 100 points. This imposes the drawback that whenever the scatter plot is redrawn, the random selection changes, which causes flickering. Future implementations have to improve the scatter plot performance.

FILTERING AND HIGHLIGHTING Effects between variables can be complex. If those effects only occur under certain preconditions, filtering and highlighting may help to reveal the relevant subsets, if the defining criteria for such a subset are sufficient. The combination of different criteria, defined using ranges, allows to select fine-grained subsets. However, there are still limits. For example, selecting an area in a scatter plot is limited to a rectangle, no individual points can be selected and unions or exclusions are also not supported. All filtering is currently based on one-dimensional range inclusion and set intersection. More diverse options require enhanced user

interface utilities (see advanced filtering in Section 3.2.2). The authors of the GPLOM mentioned that associative highlighting lacked clear affordances and was rarely used. The sliders provide those affordances. However, that items in the filter list can be clicked is still not obvious. Another improvement of the associative highlighting could be allowing the analysts to set colors for each filter criterion and group criteria. Different groups could then be highlighted in different colors and intersections as a mixture of the two colors. A major problem occurs when a filter range is defined through dragging the sliders of a quantitative variable on the y-axis. Because the sliders are aligned with a row and the histogram's blue overlay bars are interactively responding to the sliders' movements, a confusing cause-effect connection is visible. This is due to the fact that the histogram bars hide the distribution of the data points inside each category. The slider domain and the histogram height domain are not the same.

5.3.3 *Summary and Conclusion.*

GPLOMs with enhanced filtering mechanisms allow for displaying heterogeneous population study data on a small scale. The herein presented enhanced method is prototypical; hence the many suggestions for enhancing it.

GPLOMs allow for a better overview than standard bivariate plots and can be used for an explorative analysis of a small set of variables. Whether the proposed interface is suitable is therefore a question left open for the epidemiologists to answer. A study could provide insights into whether the tool is preferred and how suitable it is for the task. In particular, it has to clarify whether the proposed additions actually make it easier to find patterns and understand the dataset or not. Various problems and drawbacks have been discussed and suggestions have been made for future research.

5.4 3D REGRESSION HEAT MAP

This section is based on the publication

> **Paul Klemm**, Kai Lawonn, Sylvia Glaßer, Uli Niemann, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. 3D Regression Heat Map Analysis of Population Study Data. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):81–90, 2016.

Katrin Hegenscheid and Henry Völzke provided the data and the epidemiological background knowledge as well as the underlying hypotheses. They also conducted in the evaluation of the method. The concept was developed in meetings with Kai Lawonn, Sylvia Glaßer and Bernhard Preim. Uli Niemann provided the idea of the correlation-based feature selection and helped with the implementation of this dimension reduction step. He also conducted the evaluation for the hepatic steatosis data set.

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 (see Section 4.5), are not tailored to a specific target feature. They mostly highlight correlations between features that are known to the domain expert (e.g., correlation between body size and spine shape). The regression analysis, which is familiar

to the domain experts, is incorporated in overview visualizations to support a hypothesis-free analysis or an analysis w.r.t. a specific disease or hypothesis. For this purpose, template regression formulas are provided, 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. The 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.

The difference of the presented approach in comparison with related work and prior methods is twofold.

1. The focus lies 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.

5.4.1 *3D Regression Heat Map*

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 Chapter 2). Relationships observed using such techniques are subject of detailed statistical testing by statisticians with background in epidemiology using statistical processors, such as `SPSS`.

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 [293]. 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 5.4.2.

Regression analysis is the statistical tool of choice for analyzing relationships in epidemiological data. 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.

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.

Since it is the goal to associate the regression analyses with an overview visualization, all possible combinations of (two or more) independent features describing a target are of interest. This is achieved by introducing dynamic variables X, Y and Z into the regression notation. The method replaces the dynamic variables with all features in the data set. In a data set with n (e.g., 100) features, the regression formula

$$Cancer \sim X + Y \tag{10}$$

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. A data set with 100 features and the formula $Z \sim X + Y$ yields 1,000,000 regression models. Assuming a 50 ms computation time for a regression analysis, the calculation lasts roughly 14 h. Therefore, the computational complexity needs to be reduced. An approach for this is presented in the following section.

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) [89] 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.

The CFS algorithm is applied 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. Since the CFS algorithm performs linear, it is also well suited for data with many features.

The number of features calculated by the CFS algorithm is dependent on the information entropy in the data. In the given epidemiological data sets, a number of 10 to 30 features is usually observed. 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 trade-off 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, interesting regression models are derived in a reasonable time span (seconds to minutes instead of hours). The next section shows ways of abstracting the results to make them visible.

Abstracting Regression Results

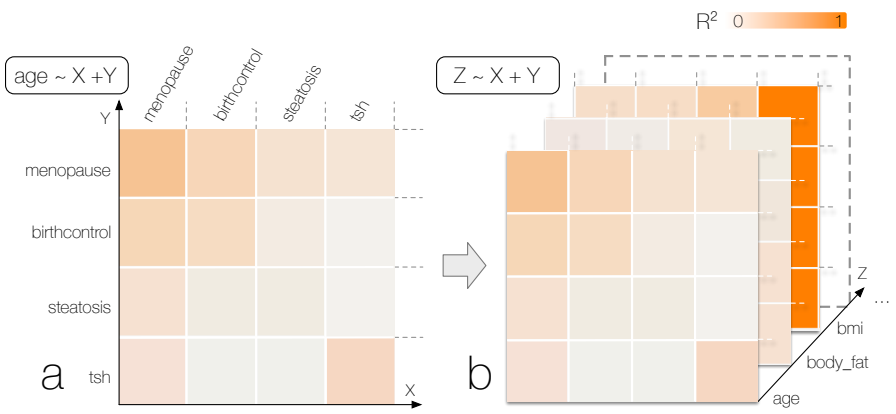


Figure 55: Overview visualization using a 2D heat map of the formula $Z \sim X + Y$, where Z assumes the feature age (a). The R^2 metrics extracted from the regression model are mapped to color saturation (a saturated color indicates a strong correlation). Now, Z is set to all features n and yields n 2D heat maps (b). These represent the slices in the *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. Image from [295].

The goal of an overview visualization is to provide a comprehensive view on the data (raw or using descriptive metrics [21]), which is easy to understand. As described in the previous work [293], 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 quality-of-fit of the resulting model is of high interest. This allows to infer the predictive quality of the independent features included in the model. The R^2, adjusted

R^2 and AIC (Akaike information criterion) metrics allow for this kind of assessment. The adjusted R^2 includes a penalty function for adding new independent features in the regression analysis. More independent features, even if they only contain noise, yield more potential information for creating the model and therefore should increase the R^2. The penalty function counteracts this effect by weighting the R^2 value with the number of independent features. This may even lead to negative R^2 values. This abstraction follows the design guidelines of hierarchical aggregation visualization proposed by Elmqvist and Fekete [64]. The results are abstracted in a way that they do not clutter the view with too many visual elements, allow for a visual summary of the underlying data, and can be discriminated using a simple visual representation. At the same time, the visualization stays interpretable, as details of each model can be accessed on demand.

2D (SLICE) VIEW Since R^2 is scaled between $[0, 1]$, it allows for comparison *between* regression models. The same 2D heat map can be applied by translating the R^2 values to *color saturation* (Fig. 55 a). 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, a slider input is provided, 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, the transfer function color mapping is inverted, assigning low AIC features to saturated colors. To include users unfamiliar with these metrics, the *Regression Heat Map* is set by default to show R^2 values.

3D VIEW Introducing Z creates a 3D heat map (Fig. 55 b). The selected metric (by 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, metrics of different regression methods are mapped to 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.

The goal is to create an overview visualization for a data set. Additionally, expert knowledge can be incorporated in 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*.

Analysis Workflow

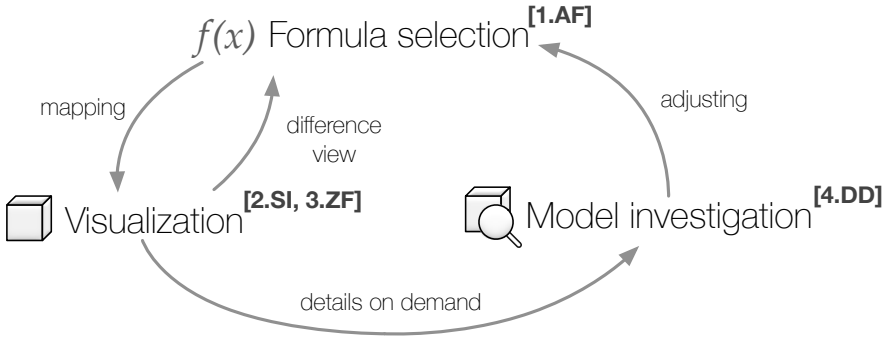


Figure 56: Different workflow types using the *3D Regression Heat Map*. [**1.AF**] The workflow starts by declaring a formula to specify a hypothesis, or to use a predefined formula for a hypothesis-free analysis. [**2.SI**] The *3D Regression Heat Map* is then visualized. The user has the option to either to adjust the formula, adjust the transfer function or to derive details-on-demand on models. [**3.ZF**] Insights into the data yield either an adjustment of the current formula or a selection of a difference view. The latter is used to compare *3D Regression Heat Maps*. [**4.DD**] Details about features using the 2D heat map representation yield insights and hypotheses about feature relations.

The *3D Regression Heat Map* is well suited for different workflow analysis techniques, based on the Visual Analytics (VA) Mantra of Keim et al. [126]:

1. **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.
2. **Show the important** [**2.SI**]. The 3D visualization acts as an overview of 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.
3. **Zoom, filter and analyze further** [**3.ZF**]. The slices of interest can then be analyzed using the 2D heat map of the slice.
4. **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).

The squared bracket abbreviation is incorporated for each step to denote the affiliation to the system design section later on. As shown in Fig. 56, 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 users 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, default formula are provided:

$Z \sim X + Y$. It represents all possible combinations of two independent features w.r.t. all features in the data set, since the features of interest are not known prior to the analysis. 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:

- $\mathtt{Cancer} \sim X + Y + Z$ 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*.
- $\mathtt{Cancer} \sim X + Y + Z + \mathtt{feature}_1 : \mathtt{feature}_2$ encodes more assumptions. This formula models the hypothesis of an interaction between $\mathtt{feature}_1$ and $\mathtt{feature}_2$ (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.
- $\mathtt{Cancer} \sim X + Y + Z$ subtracted with the regression metric from $\mathtt{Cancer} \sim \mathtt{Age}$ excludes the confounding effect that age has in view of the target $\mathtt{Cancer}$ 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 $Z \sim X + Y + \mathtt{Income}$).

5.4.2 *System Design*

The system is designed 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. Even small changes can be implemented 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

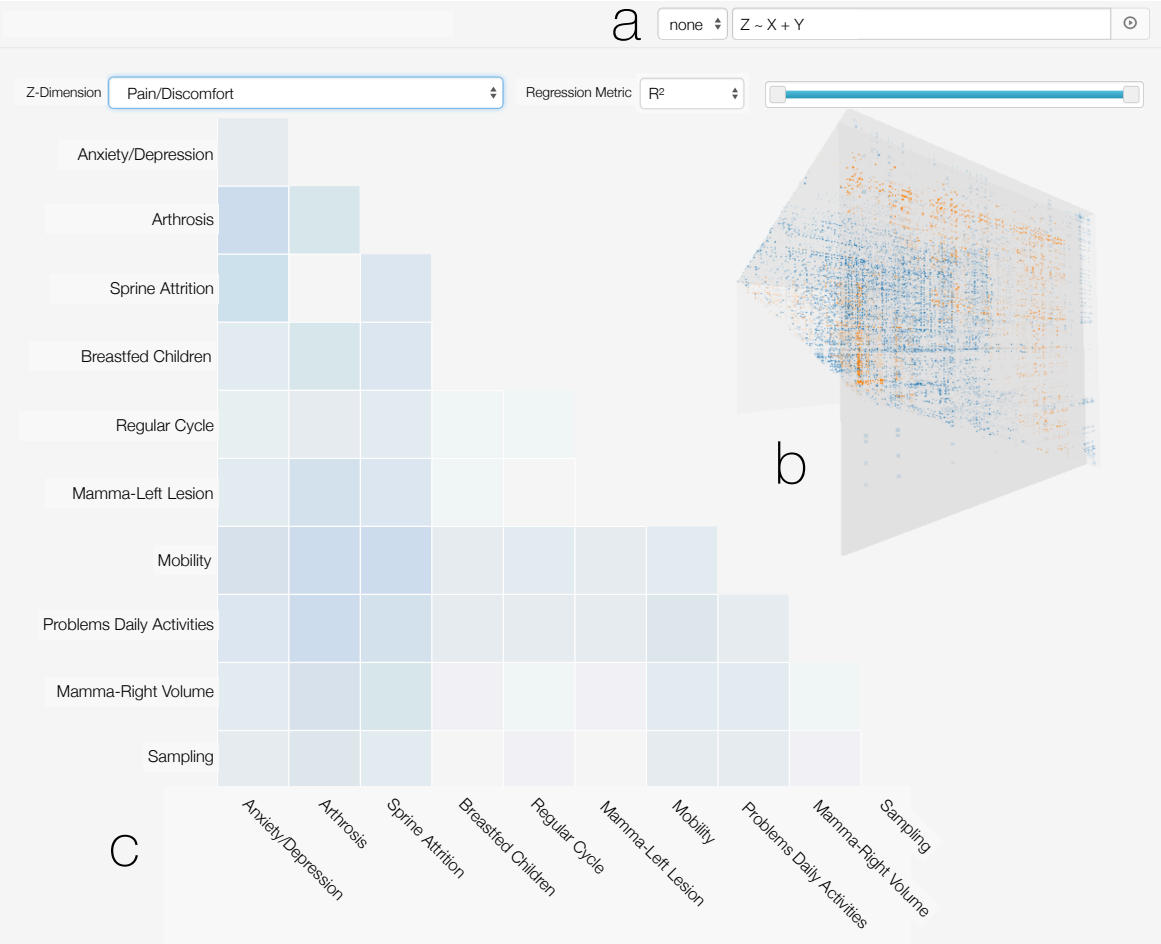


Figure 57: Breast density data set loaded into the prototype. (a) Using the formula input, the user specifies the dependent feature and calculation rules. (b) 3D heat map showing values above the matrix diagonal as overview. The values of the currently selected slice are mirrored and represented as orange data points on the slicing plane. (c) 2D heat map of the selected slice for feature *Pain/Discomfort*. Image from [295].

issues, such as the storage of confidential data, especially in the epidemiological context. Therefore, technical measures have to be incorporated to ensure a secure workflow.

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. The 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 [90], owe part of their success to their support of open file types. To allow other users even outside the epidemiological application domain to access the tool, standard ASCII-based CSV files are incorporated. 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.

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

Data security issues are raised by uploading data into an online service such as the 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. A SHA-256 hash is calculated 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. Autocompleting 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 drop-down 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 the regression metric for each cell.

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 to assess 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 55 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, half of the results can be discarded 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 drop-down 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 of 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 the 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. 57 b). 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. 59). 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. The goal is 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. 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 the 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. Two ways of achieving this are proposed.

- **The slicing metaphor from 3D volume data is applied.** In medical volume renderings, slicing views are common to view details on a selected plane in the scene. This technique for selecting 3D heat map slices is employed (e.g., by moving a plane via vertical mouse input while pressing the right mouse button). However, the whole 3D object is still displayed 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 drop-down 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 occlusion-free view. This allows to easier identify the current slice.

2D HEAT MAP SLICE VISUALIZATION [4.DD] The 2D heat map (Fig. 57 c) 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 [146].

5.4.3 *Implementation*

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

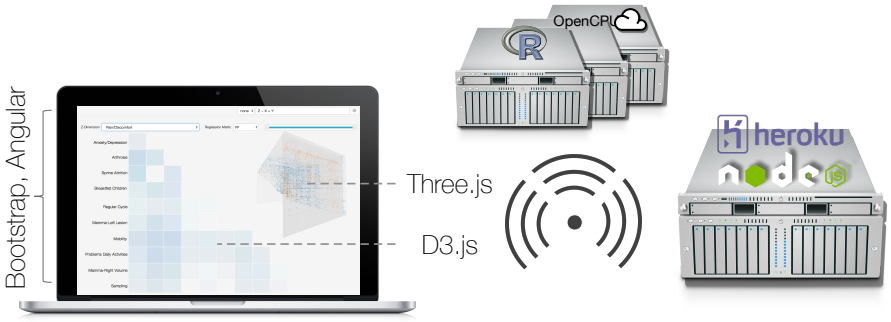


Figure 58: Overview of the technologies incorporated in the *3D Regression Heat Map* prototype. The front end (left) is realized with `HTML5/CSS3/Javascript` and different Javascript libraries, such as `Angular.js`, `Three.js` and `D3.js`. The web server (right) is written using `Node.js` and hosted 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. Image from [295].

FRONT END 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. The page layout is handled using `Twitter Bootstrap`, which also provides a rich set of user interface elements. The 2D heat map is implemented using the `D3.js` [24] information visualization library. It provides fast and easy methods for binding data to graphical elements. The 3D plot is created using the WebGL-based `Threejs` library. Different ways for displaying the cube were tested, including volume rendering, cube primitives for each data point and shader-based solutions. Open source volume rendering methods are available but do not satisfy the requirements. Creating a cube primitive for each data point resulted in non-interactive frame rates for data sets larger than 30 features (creating 30^3 cube primitives). Therefore, a shader-based solution was incorporated by rendering the cube as a sprite-based particle system, allowing to customize color and opacity of every data point. It is also the fastest tested solution.

BACK END 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`[4], a cloud application platform. The statistical computations are performed on the second structure. They rely on the statistical programming language `R`.[5] 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 [191]. This way, any computer which runs `R` can be turned into a statistical processor for the project. The back end functions necessary for all cube calculations are provided via an `R` package. It uses multi-core optimization to use all machine CPUs to speed up the calculation process. The server workload balances are managed by the front end code.

ACCESS AND SOURCE A running instance of the *3D Regression Heat Map* prototype can be found at `regressionheatmap.herokuapp.com`. The source

4 Owned by Salesforce.com, `heroku.com`
5 Open Source; `r-project.org`

for the prototype is freely available at Github.[6,7] 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.

5.4.4 *Application*

In this section, the application of the *3D Regression Heat Map* to two epidemiological data sets is described. The hepatic steatosis data set was analyzed using data mining algorithms, yielding risk groups, which are now analyzed further. Prior results from the analysis are reproduced as proof-of-concept of the 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. The method focuses on data exploration and knowledge extraction and requires a wide scope of sociodemographic, medical and lifestyle features.

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. [141] 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. *VDAR* is employed for analyzing the system.

PARTICIPANTS, SETUP AND PROCEDURE A web-based analysis is conducted 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. The analysis was conducted 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.

The Hepatic Steatosis Data Set

The data set used by Niemann et al. [186] to identify predictive features w.r.t. the reversible hepatic steatosis disorder is employed. The dichotomous

6 R-based back end:
github.com/paulklemm/regression-heatmap-r-package
7 Front End and Node.js Webserver:
github.com/paulklemm/regression-heatmap-prototype

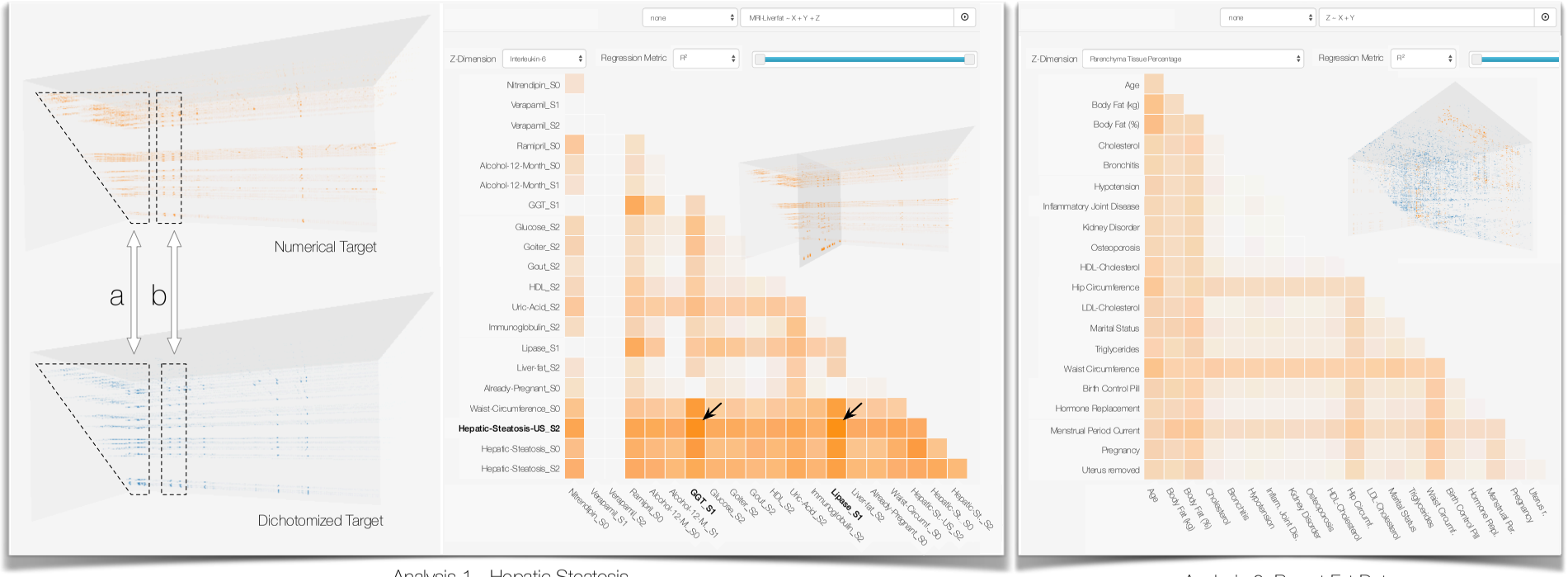


Figure 59: The analysis of the numerical and dichotomized target feature depicting liver fat values yields similar results (left). In (a), hotspots for somatometrical features with high correlations were found. High correlations were also found for features depicting *hepatic steatosis* (b). A high correlation 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*. Image from [295].

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.

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 Single Nucleotide Polymorphisms (DNA base pairs). Niemann et al. [186] 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, the analysis was performed on three populations: *males*, *females (all ages)* and *females older than 52 years*.

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% [169].

The data comprises 1,186 female subjects (368 from `SHIP-2`, 818 from `SHIP-TREND-0` cohort). 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 [100, 121].

The data of each cohort were presented as individual `SPSS` files. All features related to the mammography attributes were stored in an additional file. The `SPSS` data sets were converted 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.

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

Each analysis step is related to the VA Mantra (recall Sec. 5.4.1). The analysis goal was reproducing results with the herein presented visual analysis framework that are in accordance to the data mining-based results presented by Niemann et al. [186]. Therefore, *UN* started the [**1.AF**] step using the dichotomized MRI fat liver concentration and the formula $\texttt{mrt\_liverfat\_s2} \sim X + Y + Z$ for *male* subjects. The [**2.SI**] step using the 3D heat map locates hotspots at the end of the heat map (Fig. 59 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 steato-

sis 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. [186] were found, such as enzymes indicating liver dysfunctions, e.g., *aspartate aminotransferase*. Due to the difference between the regression model approach and the decision tree approach presented by Niemann et al. [186], a complete matching set of correlating features is not expected.

ANALYSIS OF NON-DISCRETIZED TARGET FEATURE Since the herein presented 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). One possibility is 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*, regulates the inflammation reaction of the body) correlation with liver fat values (R^2 of 0.8, see Fig. 59b). The correlation was high for both the dichotomized and continuous target feature. The literature described relations between *IL-6* and liver cancer [97] as well as chronic liver diseases [247]. For mice, strong effects of *IL-6* with hepatic steatosis were described [113]. The finding is subject to further analysis.

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 [169]. 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. 59 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 $\mathtt{Parenchyma\_Percentage} \sim X + Y + Z$ [**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.

FURTHER STATISTICAL ANALYSIS OF THE OBSERVED RELATIONSHIPS

The following analysis is available as open source and incorporates a `R Markdown` document.[8] Similar to the detailed statistical analysis conducted in Section 4.5.5, this paragraph aims to statistically evaluate the new hypotheses using standard statistical methods. The analysis is carried out using `R`. The data basis is the same data set as used in the 3D Regression Heat Map prototype, comprising of 1186 subjects with 231 features. The target is the continuous feature *parenchyma tissue percentage* and is referred to as *parenchyma tissue* in this section.

At first, the relationship between *parenchyma tissue* and *kidney disorder* is analyzed using an ANOVA, since the latter is a categorical feature. Even though there are differences in the resulting box plots (see the `R Markdown` document), the ANOVA yields a low F-value of 2.664 and a p-value of 0.103. Therefore, the alternative hypothesis, which suggests that there is a correlation between *parenchyma tissue* and *kidney disorder* has to be rejected.

The relationship between *parenchyma tissue* and *diabetes*, which was expected by *KH*, is confirmed by the ANOVA with an F-value of 12.6 and a p-value of 0.0003. Similarly, the expected relationship between *parenchyma tissue* and *hormone replacement therapy* is confirmed with a very high F-value of 43.58 and a low p-value of $6.14e^{-11}$. The classification of the mammography finding (*BI-RADS*) is defined as categorical feature for both breasts. *BI-RADS* [60] is a classification of the mammography results, ranging from "1 - no pathological findings" to "5 - highly suggestive of malignancy". There is also an additional level "6: Known biopsy - proven malignancy", but it is not included in the breast cancer data set. Therefore, two ANOVAs have to be conducted to assess the relationship with *parenchyma tissue*. The F-value for the left breast is 2.549 with a p-value of 0.0383, being barely under the 0.05 mark. The F-value for the right breast is 5.089 with a p-value of 0.0004. The differences between these results are noticeable. An explanation can be found in the plots depicted in Fig. 60 for the *BI-RADS* classification for both sides. There are less higher classifications for the left breast compared to the right one. Therefore, the difference may be explained only by chance, since the sample size is very small. Also Fig. 60 shows seemingly a decrease of the *parenchyma tissue share* for higher *BI-RADS* classifications.

The feature *lesion size* is divided into the three categories "none", "small focus (< 5 mm)" and "large focus(> 5 mm)". The ANOVA of *lesion size* with

8 **HTML version of the statistical analyses**
http://paulklemm.github.io/StatisticalReview/VAST15_Statistical_Review.html
Repository of the statistical analyses
https://github.com/paulklemm/StatisticalReview

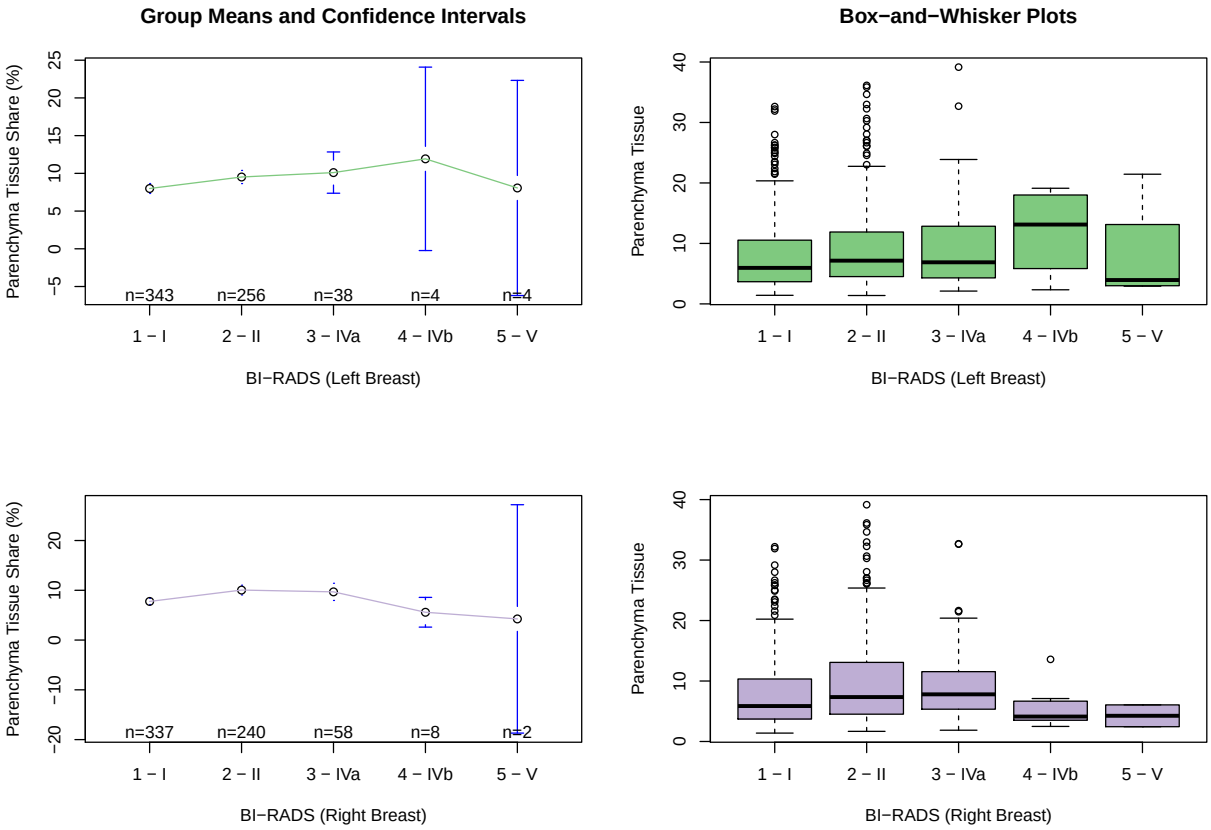


Figure 60: Plots for the *BI-RADS* five-level mammography finding feature against the *parenchyma tissue share*. The left plots comprise the mean share for each *BI-RADS* class as well as the confidence intervals and the number of women per class. The box plots on the right depict the distribution in each class. Note that there are less women in class IV and V for the left breast (46) compared to the right breast (68). Therefore, the distribution cannot be reliably described, yielding differences in the ANOVA result.

parenchyma tissue, however, yields no significant correlation with an F-value of 1.508 and a p-value of 0.222. Similarly, the ANOVA for *parenchyma tissue* and *spiral contraception* shows no correlation with a very low F-value of 0.226 and a p-value of 0.632. Therefore, a correlation between contraception, lesion size and parenchyma tissue is not supported by the data.

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 Results of complex statistical computations are mapped in 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 the herein presented 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 experts. 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 THE 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, the method presented in this section 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. The 3D heat map acts as mini-map visualization and has proven to be useful for this purpose, e.g., for highlighting differences rather than displaying absolute values (Fig. 59).

5.4.5 *Summary and Conclusion*

A technique for knowledge discovery in population study data sets with user-defined target features was presented. 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 users 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. The approach was successfully applied to find correlations in a hepatic steatosis as well as a breast density data set. The method was well received by the clinical partners.

One limitation of the proposed method is that the regression metrics always only capture a part of the underlying model. The analyst has to keep the aspects of the respective metric in mind to avoid false conclusions. 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, more regression types, which model different kinds of correlations can be introduced in the analysis. Another possibility is the extension of the 3D heat map to time-dependent data by expanding the difference heat map approach. All associated code is published as open source. Also, a freely accessible analysis platform open to heterogenous data types is provided. The goal is to open up knowledge discovery to a diverse group of domain experts to allow them to derive insight into their data.

Part III

CONCLUSION

6 SUMMARY & OUTLOOK

6.1 SUMMARY

Epidemiology aims to characterize health and disease conditions in defined populations. Insights into risk factors allow to characterize disease-specific high-risk groups [74]. 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 derive hypotheses from observations and research. The hypotheses are depicted using epidemiological features and are then statistically analyzed. Large-scale population studies collect large data sets to allow for queries with numerous diseases and hypotheses in mind. Analyzing medical image data as part of these population studies is challenging. The data has to be labeled and quantified (e.g., annotating the liver, knees or breast tissue) to allow for statistical tests of correlations with diseases. As the epidemiological data sets get larger, data-driven analysis approaches are required to utilize their complexity.

This thesis contributes a data-driven Interactive Visual Analysis approach for population study data and methods for the workflow application by combining new and existing visualizations with data mining techniques. The workflow is meant to be an enhancement of the classical epidemiological analysis pipeline. Confirmative analysis approaches, where existing hypotheses can be verified, as well as explorative, *hypothesis-free* analyses are supported. The contributions of this thesis are summarized in the following paragraphs.

CLASSIFICATION OF EXISTING VISUAL ANALYTICS AND INTERACTIVE VISUAL ANALYSIS METHODS A vast variety of Visual Analytics and Interactive Visual Analysis methods is available for numerous applications. This thesis aims to structure and categorize them to assess their suitability for population study data. Emphasis is put on analyzing medical image data, where structures of interests need to be segmented before they can be analyzed further. Different methods for creating shape variance models are described.

INTERACTIVE VISUAL ANALYSIS OF IMAGE-CENTRIC POPULATION STUDY DATA A workflow based on Visual Analytics and Interactive Visual Analysis for population study data is proposed. It incorporates different interaction complexity levels as well as the appropriate visualization type depending on the current analysis phase.

Hypothesis-based. The classical hypothesis-based confirmative analysis can be supported using Visual Analysis methods by providing fast and efficient ways of analyzing bivariate or more complex variable relationships using an integrated framework. Augmenting medical image data with non-image visualizations allows users to assess shape differences w.r.t. a target condition. It also allows users to analyze the influence of confounders, such as age or gender, to the structure of interest.

Hypothesis-free. For hypothesis-free analyses, overview visualizations are suitable to show hotspots in the data. Overview visualizations for shape-based analysis of medical image data can be carried out by applying clustering algorithms to derive shape groups. Augmenting overview visualizations of shape variances with non-image variables, such as binary disease indica-

tors, provide a simple yet powerful way of analyzing local and global shape influences.

DATA-DRIVEN ANALYSIS OF SOCIODEMOGRAPHIC, MEDICAL AND LIFESTYLE FACTORS Non-image hypothesis-based analyses are the focus of classical epidemiological analyses. Statisticians know a vast variety of well-established methods, such as regression analyses, which can be incorporated to analyze data variables w.r.t. diseases. These methods, however, fail for explorative analyses, where new relationships are derived *through* the data. Hence, this thesis focuses on combining the established methods, such as regression analyses, with clustering techniques and overview visualizations, to trigger hypothesis generation by observing new relationships. This is achieved using the 3D regression heat map, which allows epidemiologists to define relationships of interest using regression notations, which are then applied to all variable combinations to find the relationships. The Decision Tree Quality Plot incorporates clustering techniques to assess the predictive power of a set of variables towards all other variables in a data set. The Interactive Visual Analysis workflow replaces the "variable listing" step in the epidemiological pipeline, which follows after formulating a hypothesis, which is then statistically validated. This allows experts to find new relationships, which project back into the hypothesis formulation step. It also means that these new results still have to be investigated using standard statistical methods in order to be verified.

6.2 FUTURE WORK

There are many ways to extend the work presented in this thesis. The following sections cover selected aspects.

Collaborative Visual Analysis using Web-Based Technologies

Collaborative Visual Analysis between two domain experts (compare *pair analytics* in Sec. 3.2.3) allows to combine knowledge of different experts in joint analysis sessions. This does not only compensate the lack of knowledge from the field outside of the expert scope, but also triggers new ideas by communication between the experts. The interactive visual analysis methods presented in this thesis are all developed with this concept in mind and are implemented using web technologies. This allows for a fast exchange of software and enables online pair analysis sessions. Using Voice over IP and screen-sharing solutions, pair analytics sessions can be conducted with little setup times. As the number of experts and different locations increases, this solution becomes less attractive due to delay in voice transmission and lags.

The systems described in this thesis already use many advantages of web technologies. Heavy computations, such as the regression analyses in the 3D Regression Heat Map (recall Sec. 5.4), are outsourced to dedicated clusters of computers. The user's machine only needs to render the results. This approach also extends to server-side segmentation of images, as described by Jacinto et al. [122]. Responsive web-design even allows experts to use the system on mobile devices. Context menus, however, opened via mouse-over events, are not easy to facilitate. These, however, are possible with the advent of modern touch-pressure sensitive devices. Redesigning the visual analytics systems with these technologies in mind might increase the visibility of such techniques due to easy access and simple intuitive usability. In their taxonomy for visual analysis, Heer and Shneiderman [99] distinguish different steps in the analysis, namely *data & view specification*, *view manipulation* and *process & provenance*. They divide the latter in the following steps:

- *record* to investigate the analysis history,
- *annotate* to document findings,
- *share* to enable collaboration and
- *guide* users through the analysis.

The point of these four tasks is documenting the analysis process for either future analysis or sharing it with collaborators. They state that in order "*[...] to support the analysis life cycle fully, visual analytics tools should support social interaction*" [99]. This can be supported using a synchronous analysis either co-located at the same system or remote via web technologies by including multiple input devices or supporting multiple screens [120]. Vogt et al. [269] suggest large screen spaces to display many views and data entries using visual exploration tools for co-located pair analytics. For remote sessions, this may be achieved by incorporating additional means of communications, such as facilities to point to coordinates on the collaborator's screen. Social interactions for asynchronous analysis [120] can be achieved via exporting functions of views as images or whole datasets, e.g., using a bookmark feature [99]. Morton et al. [178] conclude that online visual analysis platforms, such as Tableau or Many Eyes[1], until now are only suitable to enhance the visibility of a method or a data set. They observed, "*[...] that authors tend to bring their own data and do not leverage the contributions of content from other authors*" [178]. Al-Hajj et al. [4] show that pair analytics sessions between subject matter experts and visual analytics experts work well for assessing injury information and that clinical experts value the analysis sessions.

Providing means for saving insights and states of the current analysis is a promising extension to the described methods. Mahyar et al. [162] describe a clear need for note taking as part of collaborative analyses. For analyses involving multiple experts, collaborative *and* personal note taking should be supported. The authors suggest a notebook model where individual notes can be compiled into a chronological history associated with system states.

Shrinivasan and van Wijk [242] propose the *knowledge view* for storing insights derived through the analysis using a mind map metaphor. Each insight can be stored using a short note on its content. The system saves the state of each note to resume to the analysis at these points. This does not only allow users to save analyses to resume them in a later step, but it also allows them to share states with other scientists for verification and comparison. Analysis sessions become more comprehensible, as the steps taken to derive a specific insight are recorded. This also allows experts to identify potential over-adaptation of expectations to the data set by applying too many dimension reductions or only analyzing subsets which support the current hypothesis. Extending the knowledge view of Shrinivasan and van Wijk and adapting it to the epidemiological application domain shows much potential. The knowledge view may also save figures, tables and other useful information about the data, which can be incorporated by epidemiologists to publish their findings.

Uncertainty Visualization

Epidemiological data is based on measures, simulations (e.g., simulate the spread of a contagious disease) or data derived by interviews. All modalities are associated with specific uncertainties. Measurements, for example, are prone to noise. Simulations are restricted by model assumptions. Interview question can be misunderstood or subjects may deliberately make false statements. Questions about alcohol or tobacco consumption, for example, may

1 Owned by IBM, www-01.ibm.com/software/analytics/many-eyes/

be inaccurate due to the fear of subjects of being judged. Biased answers prohibit the transfer of results of the population study to be transferred to the whole population. In the methods presented in this thesis, this fact is not considered. There are two reasons for that. First, the uncertainty has to be measured. This proves to be difficult, since there are no gold standard populations available that can be used as reference. Another possibility is including a binary flag, which marks potentially imprecise features. Second, if the uncertainty is measured, it has to be displayed, which substantially increases the plot complexity. Uncertainty adds a new dimension to the visualization [27]. Imagine, for example, a scatter plot of two numerical measures, where each subject in the population is represented using a dot and also includes confidence intervals of the accuracy for each feature. Even for a small number of subjects, this plot contains many visual clutter and is likely not helpful.

In recent years, however, uncertainty visualization gained importance [27, 206, 205]. As Brodlie et al. [27] point out, uncertainty can be really complex and consist of different descriptions. It might be defined as [27]:

- *probability density function*, where each data point is a random variate,
- *multivalue data*, where multiple values are derived for each data point,
- *bounded data*, where the value is inside finite bounds.

Brodlie et al. [27] provide different visualization ideas in their overview.

- *Juxtaposition* of plots displays the uncertainty in a separated plot, which reduces the visual clutter. The user, however, has to mentally map the corresponding data points. This can be supported by visual connections that are drawn on demand, for example using mouse-over events.
- *Overlaying* variance information imposes a similar approach as the juxtaposition, reducing the amount of visual clutter as much as possible. It can also be incorporated on demand. For example, inaccurate segmentation results can be overlaid with the variance at each data point on demand to assess the uncertainty. Aierts et al. [1] show that juxtapositioned plots are better suited for displaying uncertainty than toggling overlays.
- *Color* coding uncertainty can be incorporated for both non-image and image data. Uncertainty in surface meshes can be color-coded using a scale ranging from confident to uncertain. Data points in non-image plots can be colored using the same concept.
- *Animating* uses time as additional visualization dimension to continuously show the ranges of each data point. Lundström et al. [157] propose probablistic animation for medical image data, where classification uncertainty is displayed by incorporating a *sensitivity lens*. Animation, however, has to be incorporated with care and should be minimized as much as possible, since attention steering will become more complicated when there is much movement on the screen.
- *Sound* encodings for uncertainty information are proposed in several works in the 90s. To the knowledge of the author, there are no recent works of visualizations incorporating sound. Sound is a local feature as additional information on a selected data point and therefore it is difficult to properly include it into data visualization. The lack of research in this area, however, leaves much room for future work incorporating sound in uncertainty visualization.

The overall approach for uncertainty visualization should be displaying the information on demand. For example, when a correlation between features is displayed, the visualization needs to show the user that there is uncertainty in the data which may affect the conclusion. Therefore, the visualizations need to reflect the uncertainty in the data in a proper way to be helpful *and* truthful.

Uncertainty in epidemiology is also associated with clustering or classification techniques. The group affiliation of a subject can also be expressed using probabilities rather than binary statements. Fuzzy clustering and fuzzy classification approaches can be applied to determine the probabilities. It is also possible to set input ranges instead of single values as input parameters to determine the influence of the parameter changes and their range [248]. The different results can also be used to fit probability distributions to display uncertainty at a boundary of a cluster, e.g., through the boundary thickness. Alternatively, possible manifestations can be displayed directly. This uncertainty visualization type is referred to as Noodle Plot and is popular in meteorology, but also in climate research to showcase different scenarios. The ensemble data are derived from simulations with slight changes in the input parameters. These plots get cluttered very soon, since each simulation model, which is complex in itself, needs to be represented.

Time-Dependent Analysis

The methods presented in this thesis work for *one* acquisition cycle of a population study. Cohort studies, such as the SHIP, comprise multiple acquisition cycles of several years. The acquisition equipment and protocols for existing features remain the same to ensure comparability between cycles. Often, new features are introduced to the study to broaden its scope. For the SHIP, for example, MRI data are included in the third cycle 'SHIP-2'. Therefore, the majority of this work is focused on the 'SHIP-2' moment. Considering multiple points in time for the analysis imposes many challenges and opportunities for future work. The main questions are:

- How to *detect* subpopulations that differ in risk exposure over time? How to *monitor* the evolution of these subpopulations to predict their evolution and to identify the factors affecting this evolution?
- How to *explain* these subpopulations to the medical expert so that she or he can understand, explore and exploit the findings?

Incorporating medical image data for multiple moments w.r.t. diseases shows much potential. Instead of comparing whole differences between healthy and diseased subjects, the morphological changes between cycles can be analyzed and visualized. This, however, requires sophisticated detection and segmentation algorithms, which only capture morphological changes between the acquistion cycles. This is a very difficult task. The algorithm has to consider posture changes of the subject in the MRI as well as slight changes in the magnetic field. This becomes apparent for soft tissue, such as the liver, which may change even through slight posture changes, rendering the morphometric comparison difficult. It has to be clear which information represents morphological changes. For rigid structures, such as bones, the problem is less prominent. Changes in posture, however, can still have a strong influence on the metric measuring position and angles between structures, for example the lumbar spine canal shape.

Visualization techniques of medical image data for multiple time points may be augmented with plots of non-image visualization, similar to the method presented in Section 4.5. Standard plots, such as bar graphs, usually display one feature along the time scale in a two-axis visualization. Layered

area charts, as extension to line plots, can be used to compare data that share the same units [3]. Radially distributed line plots allow for comparing a number of univariate time-dependent features [257]. They are, however, hard to compare and the plot takes up a large amount of space. This incorporation of time-dependent image and non-image data is well suited for hypothesis-based analyses.

Clustering techniques for longitudinal data are a promising way to derive new hypotheses. This allows, for example, to assess differences in the evolution of healthy and diseased subjects. There are two ways to achieve this.

(i) Subjects are clustered at each discrete time step, yielding a group affiliation for each subject at each acquisition cycle. The features defining these clusters can then be further analyzed. Also, changes in the cluster affiliation between each time step can be analyzed. For example, clusters may split into several clusters, or a number of clusters merges into a bigger one. The features, which are responsible for these events, may indicate relevant changes w.r.t. the target disease.

(ii) Subjects are clustered along the time line. This yields one set of distinct clusters for the whole data set. The clustering does not incorporate total feature values, but differences (gradients) of them between each point in time. For example, subjects may be clustered because they highly increase the number of smoked cigarettes per day in between the acquisition cycles.

The two approaches require a distinct set of visualizations to comprehend the results. Analyzing the cluster transitions produced by clustering the subjects at each acquisition cycle (i) requires visualizations that highlight these transitions and the associated feature changes. One way to achieve this is by displaying a parallel coordinate per wave, where the cardinality of each cluster is represented using a bar. Subject transitions between the acquisition cycles can then be visualized using arcs between the bars, similar to parallel sets. The visualization of clusters with similar evolution paths (ii) has to provide means for comparing gradients of numerical features between clusters as well as differences in categorical features. The latter can be visualized using binary change flags, which indicate a change in the categorical features (e.g., transition from non-smoker to smoker).

Narrative Visualizations

A consequence of epidemiological findings is providing the public with updated information on a healthy lifestyle. Visualizations can help communicating the results to a broad audience. Large newspapers increasingly incorporate information visualizations in digital issues to provide readers with the opportunity to investigate the underlying data as supplementary material to the articles [72, 233]. The requirements for these *narrative visualizations* vastly differ from standard visualizations. The methods presented in this thesis are custom-built for epidemiologists, a very specific target group. They have an extensive knowledge of the medical conditions represented in the data and also a solid statistical background. Even with this background in mind, the visualizations and the incorporated data mining techniques have to be explained to the domain expert. The whole point of the pair analytics approach is that the expert with the domain knowledge has a visual analysis expert assisting the analysis by explaining phenomena, changing views and conducting interactions with the system. Narrative visualizations on the other side have to work for a broad audience, where one has to assume the worst case of very little experience both in epidemiology as well

as visual analysis. In order to be a valuable addition that supports the narrative, visualizations have to be clear, truthful and easy to learn. The use of optically pleasing colors that match the design of the digital platform as well as layout restrictions, imposes additional requirements which further reduce the design space for the visualizations.

In the 1920s, Otto Neurath aimed to enable citizens to participate in society and politics-related questions by educating them using pictographic statistics, so-called Isotypes [183]. In a pursue to create a universal pictorial language, Neurath encoded statistical information into easily understandable pictographics. The number of workers, for example, was mapped to minimalistic representations of humans. Combining the representation with other symbols, for example that of a factory, allowed to create contexts and encode different information. Neurath argued that numbers should not be provided in detail in favor of remembering the pictures instead of numbers. In modern narrative visualization, Isotypes are known as pictograms. The combination of pictograms allows to encode many features without requiring too much explanation. Designing proper pictograms and systems incorporating them, however, is still open research [72], even though there are rudimentary design suggestions available [233].

First approaches for providing results of epidemiological studies is the UK Longevity Explorer (Ubble) [79], which incorporates information of the UK Biobank study. Ganna and Ingelsson investigated 655 demographic, health and lifestyle features and associated the five-year death rate. The goal was to see how accurate a variable can predict the five-year death probability. Similar to the Decision Tree Quality Plot proposed in Section 5.1, the results are represented using an interactive scatter plot, which is available online.[2] In the accompanying text on the homepage the authors state that the service is intended both for people who want to investigate health-related issues and for scientists. A lay summary is also available on the homepage, which explains the displayed results without assuming a background in statistics and epidemiology. They also provide a risk calculator, which incorporates a couple of questions to detect similar subjects in the population and derive the risk of dying in the next five years.

A modern trend in the information age is self-quantification. People employ a wide variety of tools, such as fitness bands, tracking apps, smartphone journals, smart watches to quantify information about their life. The information consist of a wide variety of fields, e.g., fitness, nutrition, health status, mobility. An important aspect of collecting these information is evaluating and sharing them to develop and establish healthier habits. People like Nicholas Felton[3] experiment with new techniques to provide people with better insights into their data and correlating different entries. In order to employ this information for a large-scale comparison between multiple subjects, the analyses have to be strictly standardized to minimize acquisition biases. Otherwise, correlations are most likely showing different data acquisition habits rather than differences in their physique or habits.

Communicating health-related risk factors to people has to ensure the proper understanding of the information. The example of 23andMe, as mentioned in Section 3.3.5, shows that displaying information without the accompanying consultation of a physician can be dangerous and lead to questionable and dangerous decisions. This is an aspect that is often conveniently overlooked by visualization experts. A proper way to achieve this may be employing techniques used in journalism to tell "data stories" [233]. The main aspect here is the explaining text, which is accompanied with information visualizations. A good example is the New York Times article "*Tax Day: Are You Receiving a Marriage Penalty or Bonus?*", which allows to input

2 ubble.co.uk
3 felton.com

own information, but then steers attention to the accompanying text [50]. Other articles already have the character of web-apps, by employing many linked views that allow to assess many dimensions and relationships. A good example of this is the "*Is It Better to Rent or Buy?*" article, where users can assess the cost effectiveness of buying a home [25]. "Scientific storytelling" [159] tries to communicate scientific results to broad audiences. It requires a skillful selection of appropriate bits of data, accompanying explanations and intuitive visualizations that do not overwhelm the reader.

6.3 FUTURE POTENTIAL

As the previous section highlights in detail, there is a vast variety of future work to be done. Herein suggested is the focus on incorporating multiple time steps. Employing visual analytics methods for proper use of feature gradients along the time dimension shows great potential for deriving new hypotheses. It is also a promising way to describe risk factors. Risks can be described as changes in the medical condition, personal decisions and lifestyle factors. Employing Interactive Visual Analytics methods for the joint analysis of multiple time steps of medical image data with non-image features to highlight structural changes with disease indicators is a promising research area with little published work.

Communicating epidemiological results is key to help people to live a healthier life. This can be carried out using narrative visualizations. An even more promising way of accomplishing this is employing risk factors into personal assistance systems, which are integrated into wearables and smartphones. This way, people more consciously perceive the consequences of their lifestyle decisions, which may lead to behavioral changes to improve their health. Another related direction of future work is employing the information collected using these new wearables as additional input for the population study data. This requires a strict quality control of the involved machines, which may render this endeavor impossible. The approach, however, has to be evaluated. Alternatively, these data sources can compile a second, much larger control population, which can be used to cross check epidemiological findings as long as the related features are included in the self-quantification. But even then the data suffers from a selection bias, since self-quantification will arguably more likely be carried out by technology-affine people. Therefore, results in these populations have to be analyzed with care. The large numbers, however, may counteract this effect.

With population studies growing in both the number of participants and assessed features as well as additional data sources from social networks, self-quantification and other data sources, the need of proper analysis tools increases. People demand self-reflection based on their data. They want to know whether they live a healthy life or not. Clinicians want to assess risk factors for diseases and effective treatment plans to develop new or refined diagnoses and treatment methods. To allow for this, novel data representation and analysis techniques have to be employed. The purpose ranges from conducting explorative analyses, which may lead to new hypotheses, to communicating epidemiological results. Employing visualizations to allow for reasoning about the available data and guiding experts as well as the general public to the right direction is a promising way to go. This thesis gives an overview of the challenges in this area and provides techniques for analyzing the vast information space of large-scale population studies.

[1] Jeroen C. J. H. Aerts, Keith C. Clarke, and Alex D. Keuper. Testing Popular Visualization Techniques for Representing Model Uncertainty. *Cartography and Geographic Information Science*, 30(3):249–261, 2003. doi: 10.1559/152304003100011180. URL http://dx.doi.org/10.1559/152304003100011180.

[2] Hossein Ahmadi, Tarek Abdelzaher, Jiawei Han, Nam Pham, and Raghu 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.

[3] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of Time-Oriented Data*. Springer Science & Business Media, 2011. doi: 10.1007/978-0-85729-079-3. URL http://dx.doi.org/10.1007/978-0-85729-079-3.

[4] Samar Al-Hajj, Ian Pike, Bernhard Riecke, and Brian Fisher. Visual Analytics for Public Health: Supporting Knowledge Construction and Decision-Making. In *46th Hawaii International Conference on System Sciences*. IEEE Institute of Electrical & Electronics Engineers, 2013. doi: 10.1109/hicss.2013.599. URL http://dx.doi.org/10.1109/hicss.2013.599.

[5] Georgia Albuquerque, Martin Eisemann, Thomas Löwe, and Marcus A. Magnor. Hierarchical Brushing of High-Dimensional Data Sets Using Quality Metrics. In *Proc. of Vision, Modeling & Visualization*, pages 119–126, 2014. doi: 10.2312/vmv.20141284. URL http://dx.doi.org/10.2312/vmv.20141284.

[6] Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and M. Eduard Gröller. Reinventing the Contingency Wheel: Scalable Visual Analytics of Large Categorical Data. *IEEE Trans. on Visualization and Computer Graphics*, 18(12):2849–2858, 2012. doi: 10.1109/TVCG.2012.254. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2012.254.

[7] Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Helwig Hauser. Radial Sets: Interactive Visual Analysis of Large Overlapping Sets. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):2496–2505, 2013. doi: 10.1109/TVCG.2013.184. URL http://dx.doi.org/10.1109/TVCG.2013.184.

[8] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers. Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges. In *Eurographics Conference on Visualization - State of The Art Reports*, pages 1–21. Eurographics, 2014. doi: 10.2312/eurovisstar.20141170. URL http://publik.tuwien.ac.at/files/PubDat_228538.pdf. Vortrag: Eurographics Conference on Visualization, Swansea, UK; 2014-06-09 – 2014-06-13.

[9] Paolo Angelelli, Steffen Oeltze, Judit Haasz, Cagatay Turkay, Erlend Hodneland, Arvid Lundervold, Astri J. Lundervold, Bernhard Preim, and Helwig Hauser. Interactive Visual Analysis of Heterogeneous Cohort-Study Data. *IEEE Computer Graphics and Applications*, 34(5):70–82, 2014. doi: 10.1109/MCG.2014.40. URL http://dx.doi.org/10.1109/MCG.2014.40.

[10] George J. Annas and Sherman Elias. 23andMe and the FDA. *New England Journal of Medicine*, 370(11):985–988, 2014. doi: 10.1056/NEJMp1316367. URL http://dx.doi.org/10.1056/NEJMp1316367.

[11] Richard Arias-Hernández, Linda T. Kaastra, Tera Marie Green, and Brian D. Fisher. Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. In *44th Hawaii International International Conference on Systems Science Proceedings*, pages 1–10, 2011. doi: 10.1109/HICSS.2011.339.

[12] Ann Aschengrau and George Seage. *Essentials of Epidemiology in Public Health*. Jones & Bartlett Learning, 2008.

[13] Steven J. Atlas and Richard A. Deyo. Evaluating and Managing Acute low Back Pain in the Primary Care Setting. *Journal of General Internal Medicine*, 16(2):120–131, 2001. doi: 10.1111/j.1525-1497.2001.91141.x. URL http://dx.doi.org/10.1111/j.1525-1497.2001.91141.x.

[14] Michelle Q. Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. Guidelines for Using Multiple Views in Information Visualization. In *Advanced Visual Interfaces*, pages 110–119, 2000.

[15] Margaret E. Baron. A Note on the Historical Development of Logic Diagrams: Leibniz, Euler and Venn. *The Mathematical Gazette*, 53(384): 113–125, 1969. URL http://www.jstor.org/stable/3614533.

[16] Mauricio L. Barreto. The dot map as an Epidemiological Tool: a Case Study of Schistosoma Mansoni Infection in an Urban Setting. *International Journal of Epidemiology*, 22(4):731–741, 1993.

[17] Linda Beale, Juan Jose Abellan, Susan Hodgson, and Lars Jarup. Methodologic Issues and Approaches to Spatial Epidemiology. *Environmental Health Perspectives*, 116(8):1105–1110, 2008.

[18] Fabian Bendix, Robert Kosara, and Helwig Hauser. Parallel Sets: Visual Analysis of Categorical Data. In *IEEE Symposium on Information Visualization*, page 18, 2005. doi: 10.1109/INFOVIS.2005.27. URL http://doi.ieeecomputersociety.org/10.1109/INFOVIS.2005.27.

[19] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: the Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*, pages 319–326. Springer, 2008.

[20] Jacques Bertin. *Sémiologie Graphique: Les Diagrammes-Les réseaux-Les Cartes*. Gauthier-Villars, 1967.

[21] E. Bertini, Andrada Tatu, and Daniel Keim. Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. *IEEE Trans. on Visualization and Computer Graphics*, 17(12):2203–2212, 2011. doi: 10.1109/tvcg.2011.229. URL http://dx.doi.org/10.1109/tvcg.2011.229.

[22] Jorik Blaas, Charl P. Botha, and Frits H. Post. Interactive Visualization of Multi-Field Medical Data Using Linked Physical and Feature-Space Views. In *IEEE Symposium on Visualization*, pages 123–130, 2007. doi: 10.2312/VisSym/EuroVis07/123-130. URL http://dx.doi.org/10.2312/VisSym/EuroVis07/123-130.

[23] Ruth Bonita, Robert Beaglehole, and Tord Kjellström. *Basic Epidemiology*. World Health Organization, 2006.

[24] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[25] Mike Bostock, Shan Carter, and Archie Tse. Is It Better to Rent or Buy? http://www.nytimes.com/interactive/2014/upshot/buy-rent-calculator.html, 2015. [Online; accessed 31-July-2016].

[26] R. Brecheisen, B. Platel, B.M. Haar Romeny, and Vilanova A. Illustrative Uncertainty Visualization of DTI Fiber Pathways. *The Visual Computer*, 29(4):297–309, 2013.

[27] Ken Brodlie, Rodolfo Allendes Osorio, and Adriano Lopes. A Review of Uncertainty in Data Visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109. Springer Science & Business Media, 2012. doi: 10.1007/978-1-4471-2804-5_6. URL http://dx.doi.org/10.1007/978-1-4471-2804-5_6.

[28] Stef Busking. *Visualization of Variation and Variability*. PhD thesis, Delft University of Technology, 2014.

[29] Stef Busking, Charl P. Botha, and Frits H. Post. Dynamic Multi-View Exploration of Shape Spaces. *Computer Graphics Forum*, 29(3):973–982, 2010. doi: 10.1111/j.1467-8659.2009.01668.x. URL http://dx.doi.org/10.1111/j.1467-8659.2009.01668.x.

[30] Stef Busking, Charl P. Botha, Luca Ferrarini, Julien Milles, and Frits H. Post. Image-based Rendering of Intersecting Surfaces for Dynamic Comparative Visualization. *The Visual Computer*, 27(5):347–363, 2011. doi: 10.1007/s00371-010-0541-z. URL http://dx.doi.org/10.1007/s00371-010-0541-z.

[31] Jesus Caban, Penny Rheingans, and Terry S. Yoo. An Evaluation of Visualization Techniques to Illustrate Statistical Deformation Models. *Computer Graphics Forum*, 30(3):821–830, 2011. doi: 10.1111/j.1467-8659.2011.01931.x. URL http://dx.doi.org/10.1111/j.1467-8659.2011.01931.x.

[32] Georg Cantor. Beiträge zur Begründung der transfiniten Mengenlehre. *Mathematische Annalen*, 46(4):481–512, 1895.

[33] Nan Cao, David Gotz, Jimeng Sun, and Huamin Qu. DICON: Interactive Visual Analysis of Multidimensional Clusters. *IEEE Trans. on Visualization and Computer Graphics*, 17(12):2581–2590, 2011. doi: 10.1109/TVCG.2011.188. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2011.188.

[34] Daniel B. Carr, John F. Wallin, and D. Andrew Carr. Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. *Statistics in Medicine*, 19(17-18):2521–2538, 2000.

[35] BA Casazza. Diagnosis and Treatment of Acute low Back Pain. *American Family Physician*, 85(4):343–350, 2012.

[36] Yu-Hsuan Chan, Carlos D. Correa, and Kwan-Liu 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.

[37] Kai S. Chang. Canvas Parallel Coordinates - Shuffled Rendering. http://bl.ocks.org/syntagmatic/2420080, 2012. [Online; accessed 31-July-2016].

[38] Arnaud Chiolero. Big Data in Epidemiology: too big to Fail? *Epidemiology*, 24(6):938–939, 2013.

[39] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010*, pages 27–34, 2010. doi: 10.1109/VAST.2010.5652443. URL http://dx.doi.org/10.1109/VAST.2010.5652443.

[40] Yi-Yu Chou, Natasha Lepore, Christina Avedissian, Sarah K. Madsen, Neelroop Parikshak, Xue Hua, Leslie M. Shaw, John Q. Trojanowski, Michael W. Weiner, Arthur W. Toga, and Paul M. Thompson. Mapping Correlations Between Ventricular Expansion and CSF Amyloid And tau Biomarkers in 240 Subjects With Alzheimer's Disease, Mild Cognitive Impairment and Elderly Controls. *NeuroImage*, 46(2):394–410, 2009. doi: 10.1016/j.neuroimage.2009.02.015. URL http://dx.doi.org/10.1016/j.neuroimage.2009.02.015.

[41] Kenneth K.H. Chui, Julia B. Wenger, Steven A. Cohen, and Elena N. Naumova. Visual Analytics for Epidemiologists: Understanding the Interactions Between age, Time, and Disease With Multi-panel Graphs. *PLOS ONE*, 6(2):e14683, 2011.

[42] Herbert H. Clark. *Using Language*. Cambridge University Press (CUP), 1996. doi: 10.1017/cbo9780511620539. URL http://dx.doi.org/10.1017/cbo9780511620539.

[43] J.R. Cobb. Outline for the Study of Scoliosis. *Instructional Course Lectures*, 5:261–275, 1948.

[44] William G. Cochran. The $\chi 2$ test of goodness of fit. *The Annals of Mathematical Statistics*, pages 315–345, 1952.

[45] Christopher Collins, Gerald Penn, and M. Sheelagh T. Carpendale. Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations. *IEEE Trans. on Visualization and Computer Graphics*, 15(6):1009–1016, 2009. doi: 10.1109/TVCG.2009.122. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2009.122.

[46] German National Cohort GNC Consortium. The German National Cohort: Aims, Study Design and Organization. *European Journal of Epidemiology*, 29:371, 2014.

[47] Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PloS one*, 6(8): e23610, 2011.

[48] I. Corouge, S. Gouttard, and G. Gerig. Towards a Shape Model of White Matter Fiber Bundles Using Diffusion Tensor MRI. In *IEEE Symposium on Biomedical Imaging: Nano to Macro*, volume 1, pages 344–347, 2004. doi: 10.1109/ISBI.2004.1398545.

[49] German Ethics Council. The Future of Genetic Diagnosis–From Research to Clinical Practice. http://www.ethikrat.org/files/opinion-the-future-of-genetic-diagnosis.pdf, 2013. [Online; accessed 31-July-2016].

[50] Amanda Cox. Tax Day: Are You Receiving a Marriage Penalty or Bonus? http://www.nytimes.com/interactive/2015/04/16/upshot/marriage-penalty-couples-income.html, 2015. [Online; accessed 31-July-2016].

[51] D. R. Cox. Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1, 1984. doi: 10.2307/1403235. URL http://dx.doi.org/10.2307/1403235.

[52] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1946.

[53] Jason Davies. Parallel Sets: a Visualisation Technique for Multidimensional Categorical Data. https://www.jasondavies.com/parallel-sets/, 2015. [Online; accessed 31-July-2016].

[54] R.J.M. Dawson. The "Unusual Episode" Data Revisited. *Journal of Statistics Education*, 3(3), 1995. URL http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html.

[55] Maria Cristina Ferreira de Oliveira and Haim Levkowitz. From Visual Data Exploration to Visual Data Mining: A Survey. *IEEE Trans. on Visualization and Computer Graphics*, 9(3):378–394, 2003. doi: 10.1109/TVCG.2003.1207445. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2003.1207445.

[56] Houtao Deng, George Runger, and Eugene Tuv. Bias of importance measures for multi-valued attributes and solutions. In *Artificial Neural Networks and Machine Learning*, pages 293–300. Springer Science & Business Media, 2011. doi: 10.1007/978-3-642-21738-8_38. URL http://dx.doi.org/10.1007/978-3-642-21738-8_38.

[57] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297, 1945. doi: 10.2307/1932409. URL http://dx.doi.org/10.2307/1932409.

[58] Dubravko Dolic. *Statistik mit R: Einführung für Wirtschafts- und Sozialwissenschaftler*. Walter de Gruyter, 2004.

[59] A. Rogier T. Donders, Geert J.M.G. van der Heijden, Theo Stijnen, and Karel G.M. Moons. Review: a Gentle Introduction to Imputation of Missing Values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006. doi: 10.1016/j.jclinepi.2006.01.014. URL http://dx.doi.org/10.1016/j.jclinepi.2006.01.014.

[60] C.J. D'Orsi, E.A. Sickles, E.B. Mendelson, E.A. Morris, et al. ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. *Reston, VA: American College of Radiology*, 2013.

[61] Stephen G. Eick and ADVIZOR SULTIONS INC. ADVIZOR: A Technical Overview. *Visual Insights, Inc*, 1999.

[62] Stephen G. Eick and Graham J. Wills. High Interaction Graphics. *European Journal of Operational Research*, 81(3):445–459, 1995.

[63] Paul Elliott and Daniel Wartenberg. Spatial Epidemiology: Current Approaches and Future Challenges. *Environmental Health Perspectives*, pages 998–1006, 2004.

[64] N. Elmqvist and J.-D. Fekete. Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. *IEEE Trans. on Visualization and Computer Graphics*, 16(3):439–454, 2010. doi: 10.1109/tvcg.2009.84. URL http://dx.doi.org/10.1109/tvcg.2009.84.

[65] John W. Emerson, Walton A. Green, Barret Schloerke, Jason Crowley, Dianne Cook, Heike Hofmann, and Hadley Wickham. The Generalized Pairs Plot. *Journal of Computational and Graphical Statistics*, 22(1): 79–91, 2013.

[66] Martin Ester, Hans P. Kriegel, Jorg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of the Second International Conference on Knowledge Discovery and Data Mining.*, pages 226–231, 1996.

[67] B. S. Everitt and A. Skrondal. *The Cambridge Dictionary of Statistics*. Cambridge University Press (CUP), 2010. doi: 10.1017/cbo9780511779633. URL http://dx.doi.org/10.1017/cbo9780511779633.

[68] Maarten H. Everts, Henk Bekker, Jos B.T.M. Roerdink, and Tobias Isenberg. Depth-Dependent Halos: Illustrative Rendering of Dense Line Data. *IEEE Trans. on Visualization and Computer Graphics*, 15(6):1299–1306, 2009.

[69] Stephen J. Ferguson and Thomas Steffen. Biomechanics of the Aging Spine. *European Spine Journal*, 12(2):S97–S103, 2003.

[70] Luca Ferrarini, Hans Olofsen, Walter M. Palm, Mark A. van Buchem, Johan H. C. Reiber, and Faiza Admiraal-Behloul. GAMEs: Growing and Adaptive Meshes for Fully Automatic Shape Modeling And Analysis. *Medical Image Analysis*, 11(3):302–314, 2007. doi: 10.1016/j.media.2007.03.006. URL http://dx.doi.org/10.1016/j.media.2007.03.006.

[71] Luca Ferrarini, Giovanni B. Frisoni, Michela Pievani, Rossana Ganzola, Johan H. C. Reiber, and Julien Milles. Morphological Analysis in Epidemiological Studies Using Growing and Adaptive Meshes: Application to Subcortical Structures in AD. In *Proc. of IEEE Symposium on Biomedical Imaging*, pages 876–879, 2010. doi: 10.1109/ISBI.2010.5490126. URL http://dx.doi.org/10.1109/ISBI.2010.5490126.

[72] Ana Figueiras. Narrative Visualization: A Case Study of How to Incorporate Narrative Elements in Existing Visualizations. In *Proc. of Information Visualisation*. IEEE, 2014. doi: 10.1109/iv.2014.79. URL http://dx.doi.org/10.1109/iv.2014.79.

[73] Harvey V. Fineberg and Mary Elizabeth Wilson. Epidemic Science in Real Time. *Science*, 324(5930):987–987, 2009.

[74] Robert H. Fletcher, Suzanne W. Fletcher, and Grant S. Fletcher. *Clinical Epidemiology: the Essentials*. Lippincott Williams & Wilkins, 2012.

[75] Wolfgang Freiler, Kresimir Matkovic, and Helwig Hauser. Interactive Visual Analysis of Set-Typed Data. *IEEE Trans. on Visualization and Computer Graphics*, 14(6):1340–1347, 2008. doi: 10.1109/TVCG.2008.144. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2008.144.

[76] F. Frenet. Sur les courbes á double courbure. *Journal de Mathématiques Pures et Appliquées*, pages 437–447, 1852. URL http://eudml.org/doc/233946.

[77] Michael Friendly. Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data. *Journal of Computational and Graphical Statistics*, 8(3):373–395, 1999.

[78] Michael Friendly and Daniel Denis. The Early Origins and Development of the Scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2):103–130, 2005.

[79] Andrea Ganna and Erik Ingelsson. 5 Year Mortality Predictors in 498 103 UK Biobank Participants: a Prospective Population-based Study. *The Lancet*, 386(9993):533–540, 2015.

[80] Oliver Gloger, Jens Kühn, Adam Stanski, Henry Völzke, and Ralf Puls. A Fully Automatic Three-step Liver Segmentation Method on LDA-based Probability Maps for Multiple Contrast MR Images. *Magnetic Resonance Imaging*, 28(6):882–897, 2010.

[81] Oliver Gloger, Klaus Dietz Tönnies, Volkmar Liebscher, Bernd Kugelmann, Rene Laqua, and Henry Völzke. Prior Shape Level Set Segmentation on Multistep Generated Probability Maps of MR Datasets for Fully Automatic Kidney Parenchyma Volumetry. *IEEE Trans. on Medical Imaging*, 31(2):312–325, 2012.

[82] D. Gotz, A. Perer, and Z. Zhang. Iterative Refinement of Cohorts Using Visual Exploration and Data Analytics, 2014. URL https://www.google.com/patents/US20140108379. US Patent App. 13/650,786.

[83] Matthias Graf. Enhancement of GPLOM for use with Epidemiological Data. Technical report, Department of Simulation and Graphics, University of Magdeburg, 2014.

[84] Samuel Gratzl, Nils Gehlenborg, Alexander Lex, Hanspeter Pfister, and Marc Streit. Domino: Extracting, Comparing, and Manipulating Subsets Across Multiple Tabular Datasets. *IEEE Trans. Vis. Comput. Graph.*, 20(12):2023–2032, 2014. doi: 10.1109/TVCG.2014.2346260. URL http://dx.doi.org/10.1109/TVCG.2014.2346260.

[85] S. Greenland, J. Pearl, and J.M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

[86] Donna L. Gresh, Bernice E. Rogowitz, Raimond L. Winslow, David F. Scollan, and Christina K. Yung. WEAVE: a System for Visually Linking 3-d and Statistical Visualizations, Applied to Cardiac Simulation and Measurement Data. In *IEEE Visualization*, pages 489–492, 2000. doi: 10.1109/VISUAL.2000.885739. URL http://dx.doi.org/10.1109/VISUAL.2000.885739.

[87] Zhenyu Guo, Matthew O. Ward, and Elke A. Rundensteiner. Model Space Visualization for Multivariate Linear Trend Discovery. In *Proc. of IEEE VAST*, pages 75–82, 2009. doi: 10.1109/VAST.2009.5333431. URL http://dx.doi.org/10.1109/VAST.2009.5333431.

[88] Michael Hahsler and Sudheer Chelluboina. Visualizing Association Rules in Hierarchical Groups. In *42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011)*. The Interface Foundation of North America, June 2011.

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

[90] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: an Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[91] Kelli Ham. OpenRefine (version 2.5). http://openrefine. org. Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association*, 101(3):233, 2013.

[92] Anita Hamilton. 23andMe Invention of the Year 2008 in Time Magazine. http://content.time.com/time/magazine/pdf/best_invention_2008.pdf, 2008. [Online; accessed 31-July-2016].

[93] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques: Concepts and Techniques*. Elsevier, 2011.

[94] Ofer Harel, Enrique F. Schisterman, Albert Vexler, and Marcus D. Ruopp. Monitoring Quality Control: can we get Better Data? *Epidemiology*, 19(4):621, 2008.

[95] M. Harreby, J. Kjer, G. Hesselsøe, and K. Neergaard. Epidemiological Aspects and Risk Factors for low Back Pain in 38-year-old men and Women: a 25-year Prospective Cohort Study of 640 School Children. *European Spine Journal*, 5(5):312–318, 1996.

[96] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. Angular Brushing of Extended Parallel Coordinates. In *Symposium on Information Visualization*, pages 127–130, 2002. doi: 10.1109/INFVIS.2002.1173157. URL http://dx.doi.org/10.1109/INFVIS.2002.1173157.

[97] Guobin He, Debanjan Dhar, Hayato Nakagawa, Joan Font-Burgada, Hisanobu Ogata, Yuhong Jiang, Shabnam Shalapour, Ekihiro Seki, Shawn E. Yost, Kristen Jepsen, et al. Identification of Liver Cancer Progenitors Whose Malignant Progression Depends on Autocrine IL-6 Signaling. *Cell*, 155(2):384–396, 2013. doi: 10.1016/j.cell.2013.09.031. URL http://dx.doi.org/10.1016/j.cell.2013.09.031.

[98] Jeffrey Heer and Maneesh Agrawala. Design Considerations for Collaborative Visual Analytics. *Information Visualization*, 7(1):49–62, 2008. doi: 10.1057/palgrave.ivs.9500167.

[99] Jeffrey Heer and Ben Shneiderman. Interactive Dynamics for Visual Analysis. *Queue*, 10(2):30, 2012. doi: 10.1145/2133416.2146416. URL http://dx.doi.org/10.1145/2133416.2146416.

[100] K. Hegenscheid, J. Kühn, 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. doi: 10.1055/s-0028-1109510. URL http://dx.doi.org/10.1055/s-0028-1109510.

[101] Katrin Hegenscheid, Rebecca Seipel, Carsten O. Schmidt, Henry Völzke, Jens-Peter Kühn, Reiner Biffar, Heyo K. Kroemer, Norbert Hosten, and Ralf Puls. Potentially relevant incidental findings on research whole-body MRI in the general adult population: frequencies and management. *European Radiology*, 23(3):816–826, 2013.

[102] Julian Heinrich and Daniel Weiskopf. State of the art of Parallel Coordinates. *STAR Proceedings of Eurographics*, 2013:95–116, 2013.

[103] Harold V. Henderson and Paul F. Velleman. Building multiple regression models interactively. *Biometrics*, 37(2):391, 1981. doi: 10.2307/2530428. URL http://dx.doi.org/10.2307/2530428.

[104] Max Hermann, Anja C. Schunke, Thomas Schultz, and Reinhard Klein. A Visual Analytics Approach to Study Anatomic Covariation. In *IEEE Pacific Visualization Symposium*, pages 161–168, 2014. doi: 10.1109/PacificVis.2014.53. URL http://doi.ieeecomputersociety.org/10.1109/PacificVis.2014.53.

[105] Max Hermann, Anja C. Schunke, Thomas Schultz, and Reinhard Klein. Accurate Interactive Visualization of Large Deformations and Variability in Biomedical Image Ensembles. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):708–717, 2016. doi: 10.1109/tvcg.2015.2467198. URL http://dx.doi.org/10.1109/tvcg.2015.2467198.

[106] Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Mining Longitudinal Epidemiological Data to Understand a Reversible Disorder. In *Advances in Intelligent Data Analysis XIII - 13th International Symposium, IDA 2014, Leuven, Belgium, October 30 - November 1, 2014. Proceedings*, pages 120–130, 2014. doi: 10.1007/978-3-319-12571-8_11. URL http://dx.doi.org/10.1007/978-3-319-12571-8_11.

[107] Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Using Participant Similarity for the Classification of Epidemiological Data on Hepatic Steatosis. In *IEEE Symposium on Computer-Based Medical Systems*, pages 1–7, 2014. doi: 10.1109/CBMS.2014.28. URL http://dx.doi.org/10.1109/CBMS.2014.28.

[108] Jerry L. Hintze and Ray D. Nelson. Violin Plots: a box Plot-density Trace Synergism. *The American Statistician*, 52(2):181–184, 1998.

[109] Albert Hofman, Monique M. B. Breteler, Cornelia M. van Duijn, Harry L. A. Janssen, Gabriel P. Krestin, Ernst J. Kuipers, Bruno H. Ch. Stricker, Henning Tiemeier, Andre G. Uitterlinden, Johannes R. Vingerling, and Jacqueline C. M. Witteman. The Rotterdam Study: 2010 Objectives and Design Update. *European Journal of Epidemiology*, 24: 553–572, 2009.

[110] Heike Hofmann. Exploring Categorical Data: Interactive Mosaic Plots. *Metrika*, 51(1):11–26, 2000.

[111] Heike Hofmann, Arno Siebes, and Adalbert F. X. Wilhelm. Visualizing Association Rules With Interactive Mosaic Plots. In *Proc. of Knowledge Discovery and Data Mining*, pages 227–235, 2000. doi: 10.1145/347090.347133. URL http://doi.acm.org/10.1145/347090.347133.

[112] Markus Hofmann and Ralf Klinkenberg. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press, 2013.

[113] Feng Hong, Svetlana Radaeva, Hong-na Pan, Zhigang Tian, Richard Veech, and Bin Gao. Interleukin 6 Alleviates Hepatic Steatosis and Ischemia/Reperfusion Injury in Mice With Fatty Liver Disease. *Hepatology*, 40(4):933–941, 2004. doi: 10.1002/hep.1840400424. URL http://dx.doi.org/10.1002/hep.1840400424.

[114] Zhexue Huang. Clustering Large Data Sets With Mixed Numeric and Categorical Values. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21–34, 1997.

[115] Zhexue Huang. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998. doi: 10.1023/a:1009769707641. URL http://dx.doi.org/10.1023/a:1009769707641.

[116] Jean-Francois Im, Michael J. McGuffin, and Rock Leung. GPLOM: The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):2606–2614, 2013. doi: 10.1109/TVCG.2013.160. URL http://dx.doi.org/10.1109/TVCG.2013.160.

[117] Square Inc. Crossfilter: Fast Multidimensional Filtering for Coordinated Views. https://square.github.io/crossfilter/, 2015. [Online; accessed 31-July-2016].

[118] Alfred Inselberg and Bernard Dimsdale. Parallel Coordinates. In *Human-Machine Interactive Systems*, pages 199–233. Springer, 1991.

[119] John P.A. Ioannidis. Why Most Published Research Findings are False. *Chance*, 18(4):40–47, 2005.

[120] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, Kwan-Liu Ma, and H. Hagen. Collaborative Visualization: Definition, Challenges, and Research Agenda. *Information Visualization*, 10(4):310–326, 2011. doi: 10.1177/1473871611412817. URL http://dx.doi.org/10.1177/1473871611412817.

[121] Tatyana Ivanovska, René Laqua, Lei Wang, Volkmar Liebscher, Henry Völzke, and Katrin Hegenscheid. A Level Set Based Framework for Quantitative Evaluation of Breast Tissue Density from MRI Data. *PLoS ONE*, 9(11):e112709, 2014. doi: 10.1371/journal.pone.0112709. URL http://dx.doi.org/10.1371/journal.pone.0112709.

[122] Hector Jacinto, Razmig Kéchichian, Michel Desvignes, Rémy Prost, and Sébastien Valette. A web Interface for 3D Visualization and Interactive Segmentation of Medical Images. In *Proc. of 3D Web Technology*. Association for Computing Machinery (ACM), 2012. doi: 10.1145/2338714.2338722. URL http://dx.doi.org/10.1145/2338714.2338722.

[123] A. Jackson. Quantitative MRI of the Brain: Measuring Changes Caused by Disease. *The British Journal of Radiology*, 78(925):87–87, 2005. doi: 10.1259/bjr.78.925.780087a. URL http://dx.doi.org/10.1259/bjr.78.925.780087a.

[124] Sara Johansson and Jimmy Johansson. Visual Analysis of Mixed Data Sets Using Interactive Quantification. *SIGKDD Explorations*, 11(2):29–38, 2009. doi: 10.1145/1809400.1809406. URL http://doi.acm.org/10.1145/1809400.1809406.

[125] Wolfgang Kabsch. A Solution for the Best Rotation to Relate two Sets of Vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[126] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual Analytics: Scope and Challenges. In *Lecture Notes in Computer Science*, pages 76–90. Springer Science & Business Media, 2008. doi: 10.1007/978-3-540-71080-6_6. URL http://dx.doi.org/10.1007/978-3-540-71080-6_6.

[127] Daniel A. Keim, Florian Mansmann, and Jim Thomas. Visual Analytics: how much Visualization and how much Analytics? *SIGKDD Explorations*, 11(2):5–8, 2009. doi: 10.1145/1809400.1809403. URL http://doi.acm.org/10.1145/1809400.1809403.

[128] Daniel A. Keim, Jörn Kohlhammer, Geoffrey P. Ellis, and Florian Mansmann. *Mastering the Information age - Solving Problems With Visual Analytics*. Eurographics Association, 2010. ISBN 978-3-905673-77-7. URL http://diglib.eg.org/EG/Publications/bookstore/Data/pe_vismaster2010.htm.

[129] Ken Kelley and Kristopher J. Preacher. On effect size. *Psychological Methods*, 17(2):137–152, 2012. doi: 10.1037/a0028086. URL http://dx.doi.org/10.1037/a0028086.

[130] M.G. Kendall, A. Stuart, J.K. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics, Classical Inference and the Linear Model (Kendall's Library of Statistics) (Volume 2A)*. Arnold, 1999.

[131] Betty R. Kirkwood et al. *Essentials of Medical Statistics*. Blackwell Scientific Publications, 1988.

[132] Stefan Klein, Marco Loog, Fedde van der Lijn, Tom den Heijer, Alexander Hammers, Marleen de Bruijne, Aad van der Lugt, Robert P. W. Duin, Monique M. B. Breteler, and Wiro J. Niessen. Early Diagnosis of Dementia Based on Intersubject Whole-brain Dissimilarities. In *Proc. of IEEE Biomedical Imaging*, pages 249–252, 2010. doi: 10.1109/ISBI.2010.5490366. URL http://dx.doi.org/10.1109/ISBI.2010.5490366.

[133] Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P. W. Pluim. elastix: A Toolbox for Intensity-based Medical Image Registration. *IEEE Trans. on Med. Imaging*, 29(1):196–205, 2010. doi: 10.1109/TMI.2009.2035616. URL http://dx.doi.org/10.1109/TMI.2009.2035616.

[134] Teuvo Kohonen. Self-organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43(1):59–69, 1982. doi: 10.1007/bf00337288. URL http://dx.doi.org/10.1007/bf00337288.

[135] Zoltan Konyha, Alan Lez, Kresimir Matkovic, Mario Jelovic, and Helwig Hauser. Interactive Visual Analysis of Families of Curves Using Data Aggregation and Derivation. In *Proc. of Knowledge Management and Knowledge Technologies*, page 24, 2012. doi: 10.1145/2362456.2362487. URL http://doi.acm.org/10.1145/2362456.2362487.

[136] Robert Kosara, Fabian Bendix, and Helwig Hauser. Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data. *IEEE Trans. on Visualization and Computer Graphics*, 12(4):558–568, 2006. doi: 10.1109/TVCG.2006.76. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2006.76.

[137] Josua Krause, Adam Perer, and Harry Stavropoulos. Supporting Iterative Cohort Construction with Visual Temporal Queries. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):91–100, 2016. doi: 10.1109/tvcg.2015.2467622. URL http://dx.doi.org/10.1109/tvcg.2015.2467622.

[138] Martin Krzywinski and Naomi Altman. Points of Significance: Visualizing Samples With box Plots. *Nature Methods*, 11(2):119–120, 2014.

[139] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. Circos: an Information Aesthetic for Comparative Genomics. *Genome Research*, 19(9):1639–1645, 2009.

[140] Max Kuhn, Steve Weston, Nathan Coulter, and Mark Culp. C code for C5.0 by R. Quinlan. *C50: C5.0 Decision Trees and Rule-Based Models*, 2015. URL http://CRAN.R-project.org/package=C50. R package version 0.1.0-24.

[141] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Trans. on Visualization and Computer*

Graphics, 18(9):1520–1536, 2012. doi: 10.1109/TVCG.2011.279. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2011.279.

[142] Hans Lamecker, Martin Seebass, Thomas Lange, Hans-Christian Hege, and Peter Deuflhard. Visualization of the Variability of 3D Statistical Shape Models by Animation. *Medicine Meets Virtual Reality 12: Building a Better You: the Next Tools for Medical Education, Diagnosis, and Care*, 98: 190, 2004.

[143] Claudia Lamina, Gisela Sturm, Barbara Kollerits, and Florian Kronenberg. Visualizing Interaction Effects: a Proposal for Presentation and Interpretation. *Journal of Clinical Epidemiology*, 65(8):855–862, 2012.

[144] Tim Lammarsch, Wolfgang Aigner, Alessio Bertone, Silvia Miksch, and Alexander Rind. Towards a Concept how the Structure of Time can Support the Visual Analytics Process. In *Proc. of the Int. Workshop Visual Analytics*, pages 9–12, 2011.

[145] Bernd Landauer and Hansjörg Hofer. Radial Sets Web Demo. http://www.cvast.tuwien.ac.at/RadialSets/rsDemo/demo.html, 2015. [Online; accessed 31-July-2016].

[146] James M. Landwehr, Daryl Pregibon, and Anne C. Shoemaker. Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association*, 79(385):61–71, 1984. doi: 10.1080/01621459.1984.10477062. URL http://dx.doi.org/10.1080/01621459.1984.10477062.

[147] Morin Lang-Tapia, Vanesa España-Romero, Juan Anelo, and Manuel J. Castillo. Differences on Spinal Curvature in Standing Position by Gender, Age and Weight Status Using a Noninvasive Method. *Journal of Applied Biomechanics*, 27(2), 2011.

[148] John M. Last, International Epidemiological Association, et al. *A Dictionary of Epidemiology*, volume 4. Oxford University Press, 2001.

[149] Jeffrey T. Leek and Roger D. Peng. Statistics: P values are just the tip of the iceberg. *Nature*, 520(7549):612–612, 2015.

[150] Steff Lewis and Mike Clarke. Forest Plots: Trying to see the Wood and the Trees. *British Medical Journal*, 322(7300):1479, 2001.

[151] Alexander Lex, Marc Streit, Ernst Kruijff, and Dieter Schmalstieg. Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context. In *IEEE Pacific Visualization Symposium*, pages 57–64, 2010. doi: 10.1109/PACIFICVIS.2010.5429609. URL http://dx.doi.org/10.1109/PACIFICVIS.2010.5429609.

[152] Alexander Lex, Marc Streit, Hans-Jörg Schulz, Christian Partl, Dieter Schmalstieg, Peter J. Park, and Nils Gehlenborg. StratomeX: Visual Analysis of Large-scale Heterogeneous Genomics Data For Cancer Subtype Characterization. *Comput. Graph. Forum*, 31(3):1175–1184, 2012. doi: 10.1111/j.1467-8659.2012.03110.x. URL http://dx.doi.org/10.1111/j.1467-8659.2012.03110.x.

[153] Wei Liao, Huafu Chen, Qin Yang, and Xu Lei. Analysis of fMRI Data Using Improved Self-Organizing Mapping and Spatio-Temporal Metric Hierarchical Clustering. *IEEE Trans. on Medical Imaging*, 27(10): 1472–1483, 2008.

[154] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. *imMens*: Real-time Visual Querying of Big Data. *Computer Graphics Forum*, 32(3):421–430, 2013. doi: 10.1111/cgf.12129. URL http://dx.doi.org/10.1111/cgf.12129.

[155] Yarden Livnat, Per Gesteland, Jose Benuzillo, Warren Pettey, Dan Bolton, Frank Drews, Heidi Kramer, and Matthew Samore. Epinome-a Novel Workbench for Epidemic Investigation and Analysis of Search Strategies in Public Health Practice. In *AMIA Annual Symposium Proceedings*, volume 2010, page 647. American Medical Informatics Association, 2010.

[156] Yarden Livnat, T. Rhyne, and Matthew Samore. Epinome: a Visual-analytics Workbench for Epidemiology Data. *IEEE Computer Graphics and Applications*, 32(2):89–95, 2012.

[157] Claes Lundström, Patric Ljung, Anders Persson, and Anders Ynnerman. Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation. *IEEE Trans. on Visualization and Computer Graphics*, 13(6):1648–1655, 2007. doi: 10.1109/tvcg.2007.70518. URL http://dx.doi.org/10.1109/tvcg.2007.70518.

[158] Stephen Kaggwa Lwanga, Stanley Lemeshow, et al. Sample Size Determination in Health Studies: a Practical Manual. *Geneva: World Health Organization*. http://apps.who.int/iris/handle/10665/40062, 1991. [Online; accessed 31-July-2016].

[159] Kwan-Liu Ma, I. Liao, J. Frazier, H. Hauser, and H.N. Kostis. Scientific Storytelling Using Visualization. *IEEE Computer Graphics and Applications*, 32(1):12–19, 2012. doi: 10.1109/mcg.2012.24. URL http://dx.doi.org/10.1109/mcg.2012.24.

[160] Jock Mackinlay. Automating the Design of Graphical Presentations of Relational Information. *ACM Transactions On Graphics*, 5(2):110–141, 1986.

[161] Jock D. Mackinlay, Pat Hanrahan, and Chris Stolte. Show Me: Automatic Presentation for Visual Analysis. *IEEE Trans. on Visualization and Computer Graphics*, 13(6):1137–1144, 2007. doi: 10.1109/TVCG.2007.70594. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2007.70594.

[162] N. Mahyar, A. Sarvghad, and M. Tory. Note-taking in Co-located Collaborative Visual Analytics: Analysis of an Observational Study. *Information Visualization*, 11(3):190–204, 2012. doi: 10.1177/1473871611433713. URL http://dx.doi.org/10.1177/1473871611433713.

[163] Marian Majchrzycki, P.M. Mrozikiewicz, Piotr Kocur, Joanna Bartkowiak-Wieczorek, Marcin Hoffmann, W. Stryła, Agnieszka Seremak-Mrozikiewicz, and E. Grześkowiak. Low Back Pain in Pregnant Women. *Ginekologia Polska*, 81(11):851–855, 2010.

[164] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. An Evaluation of Visual Analytics Approaches to Comparing Cohorts of Event Sequences. In *EHRVis Workshop on Visualizing Electronic Health Record Data*, volume 14, pages 1–6, 2014.

[165] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. In *Proc. of International Conference on Intelligent User Interfaces*, pages 38–49, 2015. doi: 10.1145/2678025.2701407. URL http://doi.acm.org/10.1145/2678025.2701407.

[166] Nisha J. Manek and A. J. MacGregor. Epidemiology of Back Disorders: Prevalence, Risk Factors, and Prognosis. *Current Opinion in Rheumatology*, 17(2):134–140, 2005.

[167] Madhav V. Marathe and Anil Kumar S. Vullikanti. Computational Epidemiology. *Communications of the ACM*, 56(7):88–96, 2013. doi: 10.1145/2483852.2483871. URL http://doi.acm.org/10.1145/2483852.2483871.

[168] Kresimir Matkovic, Wolfgang Freiler, Denis Gracanin, and Helwig Hauser. ComVis: A Coordinated Multiple Views System for Prototyping New Visualization Technology. In *Proc. of Information Visualisation*. IEEE, 2008. doi: 10.1109/iv.2008.87. URL http://dx.doi.org/10.1109/iv.2008.87.

[169] Valerie A McCormack and Isabel 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. doi: 10.1158/1055-9965.epi-06-0034. URL http://dx.doi.org/10.1158/1055-9965.epi-06-0034.

[170] Wouter Meulemans, Nathalie Henry Riche, Bettina Speckmann, Basak Alper, and Tim Dwyer. KelpFusion: A Hybrid Set Visualization Technique. *IEEE Trans. on Visualization and Computer Graphics*, 19 (11):1846–1858, 2013. doi: 10.1109/TVCG.2013.76. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2013.76.

[171] David Meyer, Achim Zeileis, and Kurt Hornik. The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. *Journal of Statistical Software*, 17(3), 2006. doi: 10.18637/jss.v017.i03. URL http://dx.doi.org/10.18637/jss.v017.i03.

[172] Joerg Meyer, Jim Thomas, Stephan Diehl, Brian D. Fisher, and Daniel A. Keim. From Visualization to Visually Enabled Reasoning. In *Scientific Visualization: Advanced Concepts*, pages 227–245. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2010. doi: 10.4230/DFU.SciViz.2010.227. URL http://dx.doi.org/10.4230/DFU.SciViz.2010.227.

[173] Michael I. Miller. Computational Anatomy: Shape, Growth, and Atrophy Comparison via Diffeomorphisms. *NeuroImage*, 23:S19–S33, 2004.

[174] Michael N. Mitchell. *A Visual Guide to Stata Graphics*. Stata Press, 2008.

[175] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Boston, MA, 1997.

[176] B. Moberts, A. Vilanova, and J.J. van Wijk. Evaluation of Fiber Clustering Methods for Diffusion Tensor Imaging. In *IEEE Visualization*, pages 65–72, 2005.

[177] Megan Monroe, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman, Jeff Millstein, and Sigfried Gold. Exploring Point and Interval Event Patterns: Display Methods and Interactive Visual Query. Technical Report HCIL-2012-06, University of Maryland, 2012.

[178] Kristi Morton, Magdalena Balazinska, Dan Grossman, Robert Kosara, Jock Mackinlay, and Alon Halevy. A Measurement Study of Two Web-based Collaborative Visual Analytics Systems. Technical report, University of Washington, 2012. Technical Report UW-CSE-12-08-01.

[179] RF Mould. An investigation of the variations in normal liver shape. *The British Journal of Radiology*, 45(536):586–590, 1972.

[180] Thomas Mühlbacher and Harald Piringer. A Partition-based Framework for Building and Validating Regression Models. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):1962–1971, 2013. doi: 10.1109/TVCG.2013.125. URL http://dx.doi.org/10.1109/TVCG.2013.125.

[181] Steven J. Murdoch. Graph Redesign in R. https://www.cl.cam.ac.uk/~sjm217/projects/graphics/, 2015. [Online; accessed 31-July-2016].

[182] N. J. D. Nagelkerke. A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78(3):691–692, 1991. doi: 10.1093/biomet/78.3.691. URL http://dx.doi.org/10.1093/biomet/78.3.691.

[183] Otto Neurath. *International Picture Language; the First Rules of Isotype: With Isotype Pictures*. K. Paul, Trench, Trubner & Company, 1936.

[184] Uli Niemann, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Interactive Medical Miner: Interactively Exploring Subpopulations in Epidemiological Datasets. In *Machine Learning and Knowledge Discovery in Databases*, pages 460–463. Springer, 2014.

[185] Uli Niemann, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Subpopulation Discovery in Epidemiological Data with Subspace Clustering. *Foundations of Computing and Decision Sciences (FCDS)*, 39 (4):271–300, 2014.

[186] Uli Niemann, Henry Völzke, Jens-Peter Kühn, and Myra 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.

[187] Uli Niemann, Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Can We Classify the Participants of a Longitudinal Epidemiological Study from Their Previous Evolution? In *IEEE Symposium on Computer-Based Medical Systems*, pages 121–126, 2015. doi: 10.1109/CBMS.2015.12. URL http://dx.doi.org/10.1109/CBMS.2015.12.

[188] Steffen Oeltze, Helmut Doleisch, Helwig Hauser, Philipp Muigg, and Bernhard Preim. Interactive Visual Analysis of Perfusion Data. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1392–1399, 2007. doi: 10.1109/TVCG.2007.70569. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2007.70569.

[189] Steffen Oeltze, Helmut Doleisch, Helwig Hauser, Philipp Muigg, and Bernhard Preim. Interactive Visual Analysis of Perfusion Data. *IEEE Trans. on Visualization and Computer Graphics*, 13(6):1392–1399, 2007. doi: 10.1109/tvcg.2007.70569. URL http://dx.doi.org/10.1109/tvcg.2007.70569.

[190] Steffen Oeltze, Helwig Hauser, and Johannes Kehrer. Interactive Visual Analysis of Scientific Data, 2013. Half Day Tutorial at IEEE VIS, Seattle, WA, U.S.

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

[192] H.G. Pagendarm and Frits H. Post. Comparative Visualization - Approaches and Examples. In *Visualization in Scientific Computing*. Springer Verlag, 1995. 95–108.

[193] Tan Pang-Ning, Michael Steinbach, Vipin Kumar, et al. Introduction to Data Mining. In *Library of Congress*, page 74, 2006.

[194] Michael Papadakis, Georgios Sapkas, Elias C. Papadopoulos, and Pavlos Katonis. Pathophysiology and Biomechanics of the Aging Spine. *The Open Orthopaedics Journal*, 5:335, 2011.

[195] Christian Partl, Denis Kalkofen, Alexander Lex, Karl Kashofer, Marc Streit, and Dieter Schmalstieg. enRoute: Dynamic Path Extraction from Biological Pathway Maps For In-depth Experimental Data Analysis. In *IEEE Symposium on Biological Data Visualization*, pages 107–114, 2012. doi: 10.1109/BioVis.2012.6378600. URL http://dx.doi.org/10.1109/BioVis.2012.6378600.

[196] L. J. Paterson and J. O. Rawlings. Applied Regression Analysis: A Research Tool. *Biometrics*, 46(1):281, 1990. doi: 10.2307/2531656. URL http://dx.doi.org/10.2307/2531656.

[197] Neil Pearce. Classification of Epidemiological Study Designs. *International Journal of Epidemiology*, 41(2):393–397, 2012.

[198] Neil Pearce and Franco Merletti. Complexity, Simplicity, and Epidemiology. *International Journal of Epidemiology*, 35(3):515–519, 2006.

[199] Karl Pearson. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London.*, pages 343–414, 1895.

[200] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[201] Adam Perer and David Gotz. Data-driven Exploration of Care Plans for Patients. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 439–444. ACM, 2013.

[202] S. E. Petersen, P. M. Matthews, F. Bamberg, and et al. Imaging in Population Science: Cardiovascular Magnetic Resonance in 100,000 Participants of UK Biobank - Rationale, Challenges And Approaches. *Journal of Cardiovascular Magnetic Resonance*, 28:15–46, 2013.

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

[204] Lukasz Piwek. Tufte in R. http://motioninsocial.com/tufte/, 2015. [Online; accessed 31-July-2016].

[205] Kai Pothkow and Hans-Christian Hege. Positional Uncertainty of Isocontours: Condition Analysis and Probabilistic Measures. *IEEE Trans. on Visualization and Computer Graphics*, 17(10):1393–1406, 2011. doi: 10.1109/tvcg.2010.247. URL http://dx.doi.org/10.1109/tvcg.2010.247.

[206] Kristin Potter, Andrew Wilson, Peer-Timo Bremer, Dean Williams, Charles Doutriaux, Valerio Pas, and Chris R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *IEEE Trans. on Data Mining Workshops*. Institute of Electrical & Electronics Engineers (IEEE), 2009. doi: 10.1109/icdmw.2009.55. URL http://dx.doi.org/10.1109/icdmw.2009.55.

[207] Bernhard Preim and Charl P. Botha. *Visual Computing for Medicine: Theory, Algorithms, and Applications*. Morgan Kaufmann Publishers Inc., 2013.

[208] Bernhard Preim, Paul Klemm, Helwig Hauser, Katrin Hegenscheid, Steffen Oeltze, Klaus Toennies, and Henry Völzke. *Visualization in Medicine and Life Sciences III*, chapter Visual Analytics of Image-Centric Cohort Studies in Epidemiology. Springer, 2015.

[209] Kai Puolamäki, Panagiotis Papapetrou, and Jefrey Lijffijt. Visually controllable data mining methods. In *IEEE Conference on Data Mining Workshops*, pages 409–417, 2010. doi: 10.1109/ICDMW.2010.141. URL http://dx.doi.org/10.1109/ICDMW.2010.141.

[210] Ad Quetelet. Recherches sur le poids de l'homme aux différens âges. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 7:1, 1832.

[211] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[212] R.G. Raidou, U.A. van der Heide, C.V. Dinh, G. Ghobadi, J.F. Kallehauge, M. Breeuwer, and A. Vilanova. Visual Analytics for the Exploration of Tumor Tissue Characterization. *Computer Graphics Forum*, 34 (3):11–20, jun 2015. doi: 10.1111/cgf.12613. URL http://dx.doi.org/10.1111/cgf.12613.

[213] Marko Rak, Karin Engel, and Klaus Tönnies. Closed-form hierarchical finite element models for part-based object detection. In *VMV 2013 - Vision, Modeling, Visualization*, pages 137–144, 2013.

[214] Ramana Rao and Stuart K Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+ Context Visualization for Tabular Information. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 318–322. ACM, 1994.

[215] Babette Regierer, Valeria Zazzu, Ralf Sudbrak, Alexander Kühn, Hans Lehrach, ITFoM Consortium, et al. Future of Medicine: Models in Predictive Diagnostics and Personalized Medicine. In *Molecular Diagnostics*, pages 15–33. Springer, 2013.

[216] Mohsen Rezaeian, Graham Dunn, Selwyn St Leger, and Louis Appleby. Geographical Epidemiology, Spatial Analysis and Geographical Information Systems: a Multidisciplinary Glossary. *Journal of Epidemiology and Community Health*, 61(2):98–102, 2007.

[217] Alan S. Rigby. Statistical Methods in Epidemiology: I. Statistical Errors in Hypothesis Testing. *Disability & Rehabilitation*, 20(4):121–126, 1998.

[218] F. Ritter, T. Boskamp, A. Homeyer, H. Laue, M. Schwier, F. Link, and H. Peitgen. Medical Image Analysis. *IEEE Pulse*, 2(6):60–70, 2011. doi: 10.1109/mpul.2011.942929. URL http://dx.doi.org/10.1109/mpul.2011.942929.

[219] James M. Robins, Miguel Angel Hernan, and Babette Brumback. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[220] Anthony C. Robinson. Geovisualization and Epidemiology: a General Design Framework. In *Proceedings of the 22nd International Cartographic Conference*, pages 9–16, 2005.

[221] A. Romero-Corral, V. K. Somers, J. Sierra-Johnson, Y. Korenfeld, S. Boarin, J. Korinek, M. D. Jensen, G. Parati, and F. Lopez-Jimenez. Normal weight obesity: A risk factor for cardiometabolic dysregulation and cardiovascular mortality. *European Heart Journal*, 31(6):737–746, 2009. doi: 10.1093/eurheartj/ehp487. URL http://dx.doi.org/10.1093/eurheartj/ehp487.

[222] Francisco S. Roque, Peter B. Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Søeby, Søren Bredkjær, Anders Juul, Thomas Werge, Lars J. Jensen, and Søren Brunak. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Comput Biol*, 7(8):e1002141, 08 2011. doi: 10.1371/journal.pcbi.1002141. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1002141.

[223] Hans Rosling, Rönnlund A. Rosling, and Ola Rosling. New Software Brings Statistics Beyond the Eye. *Statistics, Knowledge and Policy: Key Indicators to Inform Decision Making*, pages 522–530, 2005.

[224] Emma Ross. Unapproved Drug use: Compassionate or Cause for Concern? *The Lancet Neurology*, 8(2):136–137, 2009.

[225] Daniel Rueckert, Alejandro F. Frangi, and Julia A. Schnabel. Automatic Construction of 3-d Statistical Deformation Models of the Brain Using Nonrigid Registration. *IEEE Trans. on Medical Imaging*, 22(8): 1014–1025, 2003.

[226] Daniel Rueckert, Alejandro F. Frangi, and Julia A. Schnabel. Automatic Construction of 3-d Statistical Deformation Models of the Brain Using Nonrigid Registration. *IEEE Trans. on Medical Imaging*, 22(8): 1014–1025, 2003.

[227] S. Salvador and P. Chan. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In *Proc. of Tools with Artificial Intelligence*, pages 576 – 584, 2004.

[228] S. Schipf, S. Knüppel, J. Hardt, and A. Stang. Directed Acyclic Graphs (DAGs)-Die Anwendung kausaler Graphen in der Epidemiologie. *Das Gesundheitswesen*, 73(12):888–892, 2011.

[229] Barret Schloerke, Jason Crowley, Di Cook, Heike Hofmann, Hadley Wickham, Francois Briatte, Moritz Marbach, and Edwin Thoen. *GGally: Extension to ggplot2.*, 2014. URL http://CRAN.R-project.org/package=GGally. R package version 0.5.0.

[230] J. Schroeder, H. Schaar, and K. Mattes. Spinal Alignment in low Back Pain Patients and Age-related Side Effects: a Multivariate Cross-sectional Analysis of Video Rasterstereography Back Shape Reconstruction Data. *European Spine Journal*, 22(9):1979–1985, 2013. doi: 10.1007/s00586-013-2787-4. URL http://dx.doi.org/10.1007/s00586-013-2787-4.

[231] Hans-Jörg Schulz, Mathias John, Andrea Unger, and Heidrun Schumann. Visual Analysis of Bipartite Biological Networks. In *Proc. of Visual Computing for Biomedicine*, pages 135–142, 2008. doi: 10.2312/VCBM/VCBM08/135-142. URL http://dx.doi.org/10.2312/VCBM/VCBM08/135-142.

[232] Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. Visual Parameter Space Analysis: A Conceptual Framework. *IEEE Trans. on Visualization and Computer Graphics*, 20(12):

2161–2170, 2014. doi: 10.1109/TVCG.2014.2346321. URL http://dx.doi.org/10.1109/TVCG.2014.2346321.

[233] E. Segel and J. Heer. Narrative Visualization: Telling Stories with Data. *IEEE Trans. on Visualization and Computer Graphics*, 16(6):1139–1148, 2010. doi: 10.1109/tvcg.2010.179. URL http://dx.doi.org/10.1109/tvcg.2010.179.

[234] Yoones A. Sekhavat and Orland Hoeber. Visualizing Association Rules Using Linked Matrix, Graph, and Detail Views. *International Journal of Intelligence Science*, 3:34–49, 2013. doi: 10.4236/ijis.2013.31A005.

[235] Jinwook Seo and Ben Shneiderman. Interactively Exploring Hierarchical Clustering Results. *IEEE Computer*, 35(7):80–86, 2002. doi: 10.1109/MC.2002.1016905. URL http://doi.ieeecomputersociety.org/10.1109/MC.2002.1016905.

[236] Jinwook Seo and Ben Shneiderman. A Rank-by-feature Framework for Interactive Exploration of Multidimensional Data. *Information Visualization*, 4(2):96–113, 2005. doi: 10.1057/palgrave.ivs.9500091. URL http://dx.doi.org/10.1057/palgrave.ivs.9500091.

[237] J. P. Shaffer. Multiple Hypothesis Testing. *Annual Review of Psychology*, 46(1):561–584, 1995. doi: 10.1146/annurev.ps.46.020195.003021. URL http://dx.doi.org/10.1146/annurev.ps.46.020195.003021.

[238] Rahman Shiri, Jaro Karppinen, Päivi Leino-Arjas, Svetlana Solovieva, Helena Varonen, Eija Kalso, Olavi Ukkola, and Eira Viikari-Juntura. Cardiovascular and Lifestyle Risk Factors in Lumbar Radicular Pain or Clinically Defined Sciatica: a Systematic Review. *European Spine Journal*, 16(12):2043–2054, 2007. doi: 10.1007/s00586-007-0362-6. URL http://dx.doi.org/10.1007/s00586-007-0362-6.

[239] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. of IEEE Symposium on Visual Languages*, pages 336–343, 1996. doi: 10.1109/VL.1996.545307. URL http://dx.doi.org/10.1109/VL.1996.545307.

[240] Ben Shneiderman. *Leonardo's Laptop: Human Needs and the New Computing Technologies*. MIT Press, 2003.

[241] Ian Shrier and Robert W. Platt. Reducing Bias Through Directed Acyclic Graphs. *BMC Medical Research Methodology*, 8(1):70, 2008.

[242] Yedendra Babu Shrinivasan and Jarke J. van Wijk. Supporting the Analytical Reasoning Process in Information Visualization. In *Proc. of Human Factors in Computing Systems*, 2008. doi: 10.1145/1357054.1357247. URL http://dx.doi.org/10.1145/1357054.1357247.

[243] Terry A. Slocum. *Thematic Cartography and Geovisualization*. Prentice Hall, 2009.

[244] Karline Soetaert. *Plot3D: Plotting Multi-Dimensional Data*, 2014. URL http://CRAN.R-project.org/package=plot3D. R package version 1.0-2.

[245] Blausen.com Staff. Quantitative MRI of the Brain: Measuring Changes Caused by Disease. *Wikiversity Journal of Medicine*, 1(2), 2015. doi: 10.15347/wjm/2014.010. URL http://dx.doi.org/10.15347/wjm/2014.010.

[246] Martijn D. Steenwijk, Julien Milles, M.A. Buchem, J.H. Reiber, and Charl P. Botha. Integrated Visual Analysis for Heterogeneous Datasets in Cohort Studies. In *IEEE VisWeek Workshop on Visual Analytics in Health Care*, volume 3, 2010.

[247] Konrad L. Streetz, Frank Tacke, Ludger Leifeld, Torsten Wüstefeld, Andrea Graw, Christian Klein, Kenjii Kamino, Ulrich Spengler, Hans Kreipe, Stefan Kubicka, et al. Interleukin 6/Gp130-dependent Pathways are Protective During Chronic Liver Diseases. *Hepatology*, 38 (1):218–229, 2003. doi: 10.1053/jhep.2003.50268. URL http://dx.doi.org/10.1053/jhep.2003.50268.

[248] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[249] Marc Streit, Alexander Lex, Michael Kalkusch, Kurt Zatloukal, and Dieter Schmalstieg. Caleydo: Connecting Pathways and Gene Expression. *Bioinformatics*, 25(20):2760–2761, 2009. doi: 10.1093/bioinformatics/btp432. URL http://dx.doi.org/10.1093/bioinformatics/btp432.

[250] Marc Streit, Alexander Lex, Samuel Gratzl, Christian Partl, Dieter Schmalstieg, Hanspeter Pfister, Peter J. Park, and Nils Gehlenborg. Guided Visual Exploration of Genomic Stratifications in Cancer. *Nature Methods*, 11(9):884–885, 2014.

[251] Martin Styner, Ipek Oguz, Shun Xu, Christian Brechbühler, Dimitrios Pantazis, James J. Levitt, Martha E. Shenton, and Guido Gerig. Framework for the Statistical Shape Analysis of Brain Structures Using SPHARM-PDM. *The Insight Journal*, 4(1071):242, 2006.

[252] Melanie Swan. Emerging Patient-driven Health Care Models: an Examination of Health Social Networks, Consumer Personalized Medicine and Quantified Self-tracking. *International Journal of Environmental Research and Public Health*, 6(2):492–525, 2009.

[253] National Lung Screening Trial Research Team et al. The National Lung Screening Trial: Overview and Study Design. *Radiology*, 258 (1):243–253, 2011. doi: 10.1148/radiol.10091808. URL http://dx.doi.org/10.1148/radiol.10091808.

[254] Sarah Thew, Alistair Sutcliffe, Rob Procter, Oscar de Bruijn, John McNaught, Colin C. Venters, and Iain Buchan. Requirements Engineering for E-science: Experiences in Epidemiology. *IEEE Software*, 26(1):80–87, jan 2009. doi: 10.1109/ms.2009.19. URL http://dx.doi.org/10.1109/ms.2009.19.

[255] James J. Thomas and Kristin A. Cook. *Illuminating the Path: the Research and Development Agenda for Visual Analytics*. IEEE Computer Society Press, 2005.

[256] Sengwee Toh and Richard Platt. Is Size the Next big Thing in Epidemiology? *Epidemiology*, 24(3):349–351, 2013.

[257] Christian Tominski, James Abello, and Heidrun Schumann. Axes-based Visualizations With Radial Layouts. In *Proc. of Applied Computing*. ACM, 2004. doi: 10.1145/967900.968153. URL http://dx.doi.org/10.1145/967900.968153.

[258] Bulent B. Tucer, Bektas Murat Yalcin, Ahmet Ozturk, Mustafa Mumtaz Mazicioglu, Yusuf Yilmaz, and Metehan Kaya. Risk Factors for low Back Pain and its Relation With Pain Related Disability and Depression in a Turkish Sample. *Turkish Neurosurgery*, 19(4):327–332, 2009.

[259] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1983. ISBN 0-9613921-0-X.

[260] Edward R. Tufte. *Envisioning Information*. Graphics Press, 1990.

[261] John W. Tukey and Paul A. Tukey. Computer Graphics and Exploratory Data Analysis: an Introduction. *The Collected Works of John W. Tukey: Graphics: 1965-1985*, 5:419, 1988.

[262] Cagatay Turkay, Július Parulek, Nathalie Reuter, and Helwig Hauser. Interactive visual analysis of temporal cluster structures. *Comput. Graph. Forum*, 30(3):711–720, 2011. doi: 10.1111/j.1467-8659.2011.01920.x. URL http://dx.doi.org/10.1111/j.1467-8659.2011.01920.x.

[263] Cagatay Turkay, Arvid Lundervold, Astri Johansen Lundervold, and Helwig Hauser. Hypothesis Generation by Interactive Visual Exploration of Heterogeneous Medical Data. In *Proc. of Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 1–12, 2013. doi: 10.1007/978-3-642-39146-0_1. URL http://dx.doi.org/10.1007/978-3-642-39146-0_1.

[264] Cagatay Turkay, Alexander Lex, Marc Streit, Hanspeter Pfister, and Helwig Hauser. Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX. *IEEE Computer Graphics and Applications*, 34(2):38–47, 2014. doi: 10.1109/MCG.2014.1. URL http://doi.ieeecomputersociety.org/10.1109/MCG.2014.1.

[265] Stephen D Turner. qqman: An R Package for Visualizing GWAS Results Using QQ and Manhattan Plots. *bioRxiv*, page 005165, 2014.

[266] Stef van den Elzen and Jarke J. van Wijk. Small Multiples, Large Singles: A New Approach for Visual Data Exploration. *Computer Graphics Forum*, 32(3):191–200, 2013. doi: 10.1111/cgf.12106. URL http://dx.doi.org/10.1111/cgf.12106.

[267] Maurits van Tulder, Bart Koes, and Claire Bombardier. Low Back Pain. *Best Practice & Research Clinical Rheumatology*, 16(5):761–775, 2002. ISSN 1521-6942. doi: 10.1053/berh.2002.0267.

[268] Anja Victor, Amelie Elsäßer, Gerhard Hommel, and Maria Blettner. Judging a Plethora of P-values: how to Contend With the Problem of Multiple Testing-part 10 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*, 107(4):50, 2010.

[269] Katherine Vogt, Lauren Bradel, Christopher Andrews, Chris North, Alex Endert, and Duke Hutchings. Co-located Collaborative Sensemaking on a Large High-Resolution Display with Multiple Input Devices. In *Human-Computer Interaction*, pages 589–604. Springer Science & Business Media, 2011. doi: 10.1007/978-3-642-23771-3_44. URL http://dx.doi.org/10.1007/978-3-642-23771-3_44.

[270] Henry Völzke, D. Alte, C. Schmidt, et al. Cohort Profile: The Study of Health in Pomerania. *International Journal of Epidemiology*, 40(2):294–307, 2010. doi: 10.1093/ije/dyp394. URL http://dx.doi.org/10.1093/ije/dyp394.

[271] Taowei David Wang, Catherine Plaisant, Ben Shneiderman, Neil Spring, David Roseman, Greg Marchand, Vikramjit Mukherjee, and Mark S. Smith. Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison. *IEEE Trans. on Visualization and Computer Graphics*, 15(6):1049–1056, 2009. doi: 10.1109/TVCG.2009.187. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2009.187.

[272] Chris Weaver. Cross-filtered Views for Multidimensional Visual Analysis. *IEEE Trans. on Visualization and Computer Graphics*, 16 (2):192–204, 2010. doi: 10.1109/TVCG.2009.94. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2009.94.

[273] Gunther H. Weber and Helwig Hauser. Interactive Visual Exploration and Analysis. In *Scientific Visualization*, pages 161–173. Springer, 2014.

[274] Rick Weiss and Lisa-Joy Zgorski. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments. *Office of Science and Technology Policy Executive Office of the President*, 2012.

[275] Tracey L. Weissgerber, Natasa M. Milic, Stacey J. Winham, and Vesna D. Garovic. Beyond bar and line graphs: Time for a new data presentation paradigm. *PLoS Biol*, 13(4):e1002128, 04 2015. doi: 10.1371/journal.pbio.1002128. URL http://dx.doi.org/10.1371%2Fjournal.pbio.1002128.

[276] Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2015. URL http://CRAN.R-project.org/package=dplyr. R package version 0.4.3.

[277] Paul Wicks, Timothy E. Vaughan, Michael P. Massagli, and James Heywood. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, 29(5):411–414, 2011.

[278] Leland Wilkinson, Anushka Anand, and Robert L. Grossman. High-Dimensional Visual Analytics: Interactive Exploration Guided by Pairwise Views of Point Distributions. *IEEE Trans. on Visualization and Computer Graphics*, 12(6):1363–1372, 2006. doi: 10.1109/TVCG.2006.94. URL http://doi.ieeecomputersociety.org/10.1109/TVCG.2006.94.

[279] Pak Chung Wong. Visual data mining. *IEEE Computer Graphics and Applications*, 19(5):20–21, 1999. doi: 10.1109/MCG.1999.788794. URL http://doi.ieeecomputersociety.org/10.1109/MCG.1999.788794.

[280] Krist Wongsuphasawat and David Gotz. Outflow: Visualizing Patient Flow by Symptoms and Outcome. In *IEEE VisWeek Workshop on Visual Analytics in Healthcare*, 2011.

[281] World Health Organization. *Obesity: Preventing and Managing the Global Epidemic*. Number 894 in WHO Technical Report Series. World Health Organization, 2000.

[282] Li Yang. Visualizing Frequent Itemsets, Association Rules, and Sequential Patterns in Parallel Coordinates. In *Proc. of the Conference on Computational Science and Its Applications: Part I*, pages 21–30. Springer-Verlag, 2003.

[283] Hongfeng Yu, Chaoli Wang, Ching-Kuang Shene, and Jacqueline H. Chen. Hierarchical Streamline Bundles. *IEEE Trans. on Visualization and Computer Graphics*, 18(8):1353–67, 2012.

[284] Athanassios Zagouras. L-method for Computing the Optimal Number of Clusters. https://de.mathworks.com/matlabcentral/fileexchange/38771-l-method, 2014. [Online; accessed 31-July-2016].

[285] Zhiyuan Zhang, David Gotz, and Adam Perer. Interactive Visual Patient Cohort Analysis. In *Proc. of IEEE VisWeek Workshop on Visual Analytics in Health Care*, pages 14–15, 2012.

[286] Zhiyuan Zhang, David Gotz, and Adam Perer. Iterative Cohort Analysis and Exploration. *Information Visualization*, pages 289–307, 2015. doi: 10.1177/1473871614526077. URL http://dx.doi.org/10.1177/1473871614526077.

[287] Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. MatrixWave: Visual Comparison of Event Sequence Data. In *Proc. of Conference on Human Factors in Computing Systems, 2015*, pages 259–268, 2015. doi: 10.1145/2702123.2702419. URL http://doi.acm.org/10.1145/2702123.2702419.

[288] Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, and Baoquan Chen. Visual Clustering in Parallel Coordinates. *Computer Graphics Forum*, 27(3):1047–1054, 2008. doi: 10.1111/j.1467-8659.2008.01241.x. URL http://dx.doi.org/10.1111/j.1467-8659.2008.01241.x.

[289] G.A. Zielhuis. Biobanking for Epidemiology. *Public Health*, 126(3):214–216, 2012. doi: 10.1016/j.puhe.2011.12.007. URL http://dx.doi.org/10.1016/j.puhe.2011.12.007.

[290] Paul Klemm, Steffen Oeltze, Katrin Hegenscheid, Henry Völzke, Klaus D. Tönnies, and Bernhard Preim. Visualization and exploration of shape variance for the analysis of cohort study data. In *Proc. of the Vision, Modeling, and Visualization Workshop*, pages 221–222, 2012. doi: 10.2312/PE/VMV/VMV12/221-222. URL http://dx.doi.org/10.2312/PE/VMV/VMV12/221-222.

[291] Paul Klemm, Kai Lawonn, Marko Rak, Bernhard Preim, Klaus D. Tönnies, Katrin Hegenscheid, Henry Völzke, and Steffen Oeltze. Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data. In *Proc. of the Vision, Modeling, and Visualization Workshop*, pages 121–128, 2013. doi: 10.2312/PE.VMV.VMV13.121-128. URL http://dx.doi.org/10.2312/PE.VMV.VMV13.121-128.

[292] Paul Klemm, Lisa Frauenstein, David Perlich, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Clustering Socio-Demographic and Medical Attribute Data in Cohort Studies. In *Proceedings des Workshops Bildverarbeitung für die Medizin*, pages 180–185, 2014. doi: 10.1007/978-3-642-54111-7_36. URL http://dx.doi.org/10.1007/978-3-642-54111-7_36.

[293] Paul Klemm, Steffen Oeltze-Jafra, Kai Lawonn, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Trans. on Visualization and Computer Graphics*, 20(12):1673–1682, 2014. doi: 10.1109/TVCG.2014.2346591. URL http://dx.doi.org/10.1109/TVCG.2014.2346591.

[294] Paul Klemm, Sylvia Glaßer, Kai Lawonn, Marko Rak, Henry Völzke, Katrin Hegenscheid, and Bernhard Preim. Interactive Visual Analysis of Lumbar Back Pain - What the Lumbar Spine Tells About Your Life. In *Proc. of Information Visualization Theory and Applications*, pages 85–92, 2015. doi: 10.5220/0005235500850092. URL http://dx.doi.org/10.5220/0005235500850092.

[295] Paul Klemm, Kai Lawonn, Sylvia Glaßer, Uli Niemann, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. 3D Regression Heat Map Analysis of Population Study Data. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):81–90, 2016. doi: 10.1109/tvcg.2015.2468291. URL http://dx.doi.org/10.1109/tvcg.2015.2468291.

[296] Bernhard Preim, Paul Klemm, Helwig Hauser, Katrin Hegenscheid, Steffen Oeltze, Klaus Toennies, and Henry Völzke. Visualization in Medicine and Life Sciences III. chapter Visual Analytics of Image-Centric Cohort Studies in Epidemiology, pages 221–248. Springer, 2016.

[297] Klaus D. Tönnies, Oliver Gloger, Marko Rak, Charlotte Winkler, Paul Klemm, Bernhard Preim, and Henry Völzke. Image Analysis in Epidemiological Applications. *it - Information Technology*, 57(1):22–29, 2015. doi: 10.1515/itit-2014-1071. URL http://dx.doi.org/10.1515/itit-2014-1071.