Integrated Dual Analysis of Quantitative and Qualitative High-Dimensional Data

Juliane Müller, Laura Garrison, Philipp Ulbrich, Stefanie Schreiber, Stefan Bruckner, Helwig Hauser, Steffen Oeltze-Jafra

Abstract—The Dual Analysis framework is a powerful enabling technology for the exploration of high dimensional quantitative data by treating data dimensions as first-class objects that can be explored in tandem with data values. In this work, we extend the Dual Analysis framework through the *joint* treatment of quantitative (numerical) and qualitative (categorical) dimensions. Computing common measures for all dimensions allows us to visualize both quantitative and qualitative dimensions in the same view. This enables a natural *joint* treatment of mixed data during interactive visual exploration and analysis. Several measures of variation for nominal qualitative data can also be applied to ordinal qualitative and quantitative data. For example, instead of measuring variability from a mean or median, other measures assess inter-data variation or average variation from a mode. In this work, we demonstrate how these measures can be integrated into the Dual Analysis framework to explore and generate hypotheses about high-dimensional mixed data. A medical case study using clinical routine data of patients suffering from Cerebral Small Vessel Disease (CSVD), conducted with a senior neurologist and a medical student, shows that a joint Dual Analysis approach for quantitative and qualitative data can rapidly lead to new insights based on which new hypotheses may be generated.

Index Terms—Dual Analysis approach, High-dimensional data, Mixed data, Mixed statistical analysis

1 Introduction

Visual analytics provides valuable mixed-initiative approaches, where computational data analysis complements interactive visual exploration. This enables the study of rich datasets from a large variety of fields, including data-driven sciences as well as big data applications in business and society. Such a mixed-initiative approach augments powerful statistical tools with interactive visualization to permit an iterative analysis for rapid hypothesis generation. Approaches addressing mixed data, however, are rare and visual analysis of such data remains a major challenge of increasing relevance. Many existing approaches treat quantitative (numerical) and qualitative (categorical) data separately, where qualitative data are used only for variable grouping [1], [2], [3]. Since most descriptive statistics are not applicable to both data types, an integrated analysis of all dimensions or items is difficult. This then prevents the user from exploring the entire space of dimensional relations in hypothesis generation.

The Dual Analysis framework [4] is an advanced method for high-dimensional data analysis that provides a hypothesis-generative and correlation exploratory approach through interactive visual analysis. It allows for simultaneous investigation of data items and dimensions, achieved by (1) plotting summary statistics of dimensions, (2) equipping the plots with brushing facilities, and then (3) linking these to corresponding item plots. This Dual Analysis concept is especially useful for correlation analysis in so-called "wide and shallow" datasets, such as clinical routine data, which are characterized by many dimensions (columns in a table), few observations (rows), and often, a high frequency of missing values. This dataset type is difficult to analyze using standard statistical approaches. Although imputation methods can assist in correlation analysis when data are missing, the results are strongly affected by the proportion of missingness [5]. Therefore, visual approaches, such as the Dual Analysis approach, are a valuable complement in hypothesis generation for this type of data.

A critical limitation of the current Dual Analysis approach is that qualitative data can only be employed for selection purposes due to the lack of joint statistics for both quantitative and qualitative data. Using quantitative statistical measures without accounting for the qualitative dimensions may overlook interesting and relevant patterns and relationships. The lack of joint descriptive statistical measures for both quantitative and qualitative data in this approach therefore may only provide a partial picture of the relevant relationships and patterns in a given dataset. Consider for instance a synthetic dataset containing four quantitative [height, weight, waist circumference, BMI], one date [birthdate], and six qualitative [education level, workout frequency, smoking, gender, eye color, cardiac risk] dimensions. Some of these measures are interrelated irrespective of their specific data type. For example,

- *Juliane Müller is with Dept. of Neurology, Otto von Guericke University Magdeburg, Germany. E-mail: juliane.mueller@med.ovgu.de*
- *Laura Garrison is with Dept. of Informatics & Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland Univ. Hospital, University of Bergen, Norway. E-mail: laura.garrison@uib.no*
- *Philipp Ulbrich is with Dept. of Neurology & the Center for Behavioral Brain Sciences, Otto von Guericke University Magdeburg, Germany. E-mail: ph.ulbrich@web.de*
- *Stefanie Schreiber is with Dept. of Neurology & the Center for Behavioral Brain Sciences, Otto von Guericke University Magdeburg, Germany. E-mail: stefanie.schreiber@med.ovgu.de*
- *Stefan Bruckner is with Dept. of Informatics & Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland Univ. Hospital, University of Bergen, Norway. E-mail: stefan.bruckner@uib.no*
- *Helwig Hauser is with Dept. of Informatics & Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland Univ. Hospital, University of Bergen, Norway. E-mail: helwig.hauser@uib.no*
- *Steffen Oeltze-Jafra is with Dept. of Neurology & the Center for Behavioral Brain Sciences, Otto von Guericke University Magdeburg, Germany. E-mail: steffen.oeltze-jafra@med.ovgu.de*

Manuscript received April 01, 2020; accepted January 28, 2021.

978-1-5386-5541-2/18/$31.00

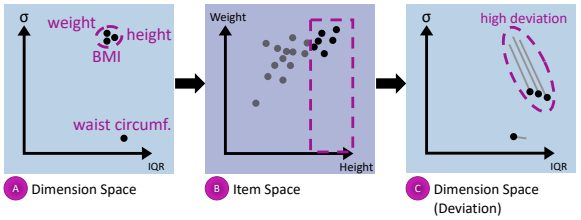

Fig. 1. The Dual Analysis approach. Users treat *quantitative* dimensions (A) and items (B) as first order objects. Selecting a subset of patients taller than 1.8m (B), we note the changes in dimension statistics given by deviation lines pointing from the original value in descriptive statistics for the whole dataset to the measure resulting from subset selection in the deviation plot (C). In this context, high deviations, as given for weight, height, and BMI, are indicators for a possible correlation with the subset selector, in this case: height.

height (quantitative measure) is correlated with gender (qualitative measure), which would be missed in the original Dual Analysis approach when investigating possible correlations with height.

Statistical tools such as SPSS [6] and PSPP [7] are commonly used for complex data analysis. While such tools support numerous computation methods and rudimentary visualization capabilities, they lack interactivity. Interactive exploration of various data subsets, however, aids in the identification of patterns and correlations. This style of interaction necessitates a naturally iterative approach that is difficult to accomplish with standard statistical packages. Approaches that employ interaction techniques for iterative exploration, without limitation by data type, provide the user an opportunity to follow a line of inquiry and then rapidly change tracks as new, interesting patterns become visible.

With this paper, our key contribution is the introduction of an integrated approach to hypothesis generation by the joint treatment of quantitative and qualitative data in the Dual Analysis framework. We build our method on the existing Dual Analysis approach by Turkay et al. [4] to enable a more free form approach to hypothesis generation. Our main contributions include:

- An *integrated visual analysis approach* for both quantitative and qualitative data in the Dual Analysis model.
- *Validation* of our integrated quantitative and qualitative data analysis approach in a *case study with real-world medical cohort data* recorded in clinical routine.

The source code of our interactive visual approach is publicly available online (https://github.com/JulianeMu/IntegratedDualAnalysisAproach_MDA).

2 Background: Dual Analysis Approach

The Dual Analysis method, developed by Turkay et al. [4], [8], is a visual approach allowing for a simultaneous investigation of data items and data dimensions in two linked spaces, *dimension space* and *item space* (Fig. 1). In dimension space, each visual mark represents a dimension of the data (column of the dataset), such as *height* or *weight*, whereas in item space, each visual mark represents a data item, e.g., a row of the dataset. Since individual dimensions are comprised of different data types with different data ranges, they cannot be easily compared unless presented by descriptive statistics, e.g., *Interquartile Range* (IQR) and *standard deviation* (σ). The Dual Analysis approach adopts this method for presenting dimensions, and is thus able to treat both items and dimensions as first-order objects for a simultaneous investigation of the whole dataset [4], [8]. This treatment assists in obtaining a better overview on the distribution of the data. For example, in our synthetic dataset we can see that height, weight, and BMI form a cluster, due to similar *standard deviations* and *IQRs*, whereas waist circumference is an outlier in this context (Fig. 1, A).

With a possible hypothesis in mind, the user can quickly and iteratively check their assumed correlations by selecting a subset of interest and investigating the related changes among the descriptive statistics between the whole dataset and selected subset. In this context, high deviations within the descriptive statistics indicate a possible correlation between the dimensions, with respect to the selection. For this purpose, Turkay et al. [4], [8] integrated a deviation plot presenting the changes in the statistical computations resulting from subset selections (Fig. 1, (C)). This allows for investigating clusters of dimensions behaving similarly in the data subset. In the synthetic dataset, for example, the possible correlation between *weight*, *height*, and *BMI* (all quantitative measures) becomes rapidly visible when selecting a subset of persons taller than 1.8m (Fig. 1, (C)). This is emphasized when investigating the resulting high change in IQR and σ in the deviations plot.

A key limitation of this approach, however, is in its unbalanced treatment of mixed data. Qualitative dimensions are only used to define data subsets, or are converted to numeric representations, if they are used at all for hypothesis generation. No visual emphasis of changes in descriptive statistics for qualitative data is provided. Consequently, a possible correlation between BMI (quantitative measure) and workout frequency (qualitative measure) would have been missed, e.g., when selecting only persons with a BMI > 25.

3 Related Work

There is an extensive body of work detailing methods for interactive visual analysis of high-dimensional data; Kehrer and Hauser [9] provide a thorough review of these methods. Coordinated multiple views with brushing and linking mechanisms have consistently proven useful in retrieving patterns and relationships from the data. Widely-used applications, for instance Tableau [10], exemplify such interaction methods. Computational methods such as aggregation, clustering, and overview statistics have also been frequently used for studying high-dimensional datasets on a regular basis [9], [11].

Visual Analysis of Mixed Data. Kehrer et al. [12] provide a variety of statistical moments to explore data outlyingness as well as relationships between data items. In a medical context, Angelelli et al. [13] presented a data-cube model which used qualitative attributes as dimensional filters for quantitative measures. These and other visual and computational analysis methods aid in coping with high dimensionality, but have focused on quantitative variable analysis in their approach. Qualitative data are often visualized as subset selectors for quantitative data – an approach often used is that of pivotization, which manifests as trellis-style displays as seen in Tableau or Polaris [2]. Kehrer et al. [1] describe and use this technique in scatterplots, function plots, and map visualizations. SeekAView similarly uses qualitative data primarily as subset selectors within subspaces to reduce the dimensionality of mixed high-dimensional data [14]. A second common method of analysis for mixed data focuses on investigating the associations between quantitative and qualitative data by correlation analysis, such as in causality analysis [15], [16], or through the use of correlation maps [17].

Statistical Analysis of Qualitative Data. Descriptive statistics tools commonly focus on the analysis of qualitative data via frequencies, proportions, or marginal distributions. Central tendency

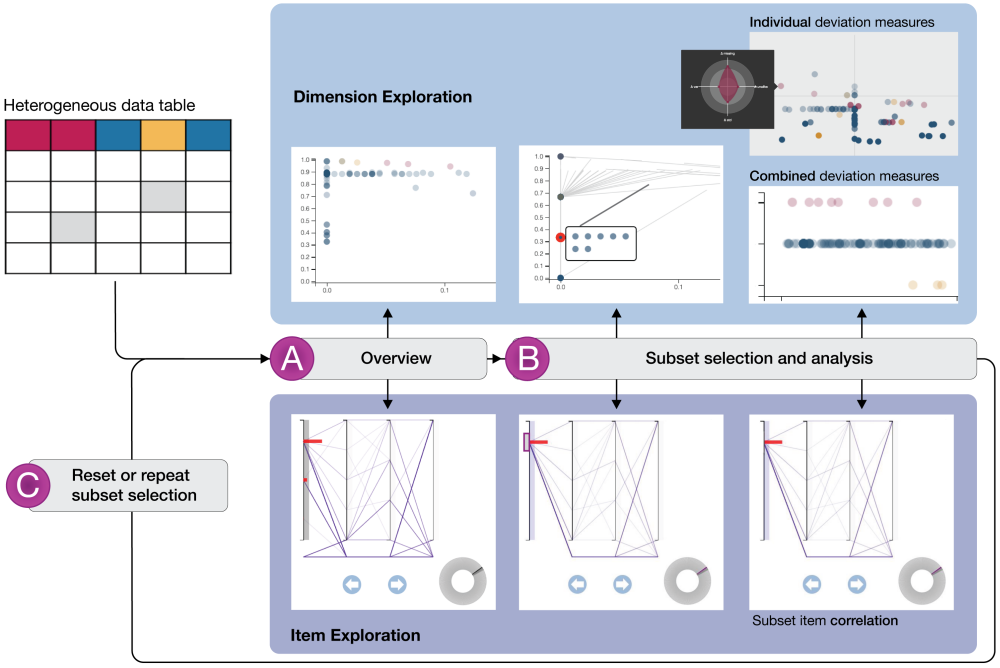

Fig. 2. Conceptual workflow for hypothesis generation in the Integrated Dual Analysis framework. We represent quantitative data as blue, qualitative data as red, and dates (considered for the sake of analysis as quantitative data) as yellow. Beginning with tabular data as input, users are provided with an overview of quantitative and qualitative data in linked dimension and item views using scatter- and parallel coordinates plots, respectively (A). Subsequent item exploration allows for item selection. These selections may be interactively observed in the dimensions plot (B). Deeper exploration of joint descriptive statistical measures of variability and missingness for quantitative and qualitative data can be seen in specialized deviation plots: deviation overview and combined deviation measures. Combined deviation measures, sorted by data type, present the average of all descriptive statistical measures calculated in our approach. In this context, a higher deviation is an indicator for a possible correlation with the applied subset selection. The correlation type for the data subset can then be investigated in the item view. At any stage in the process subset selections may be updated, or the user may return to the beginning of the process to initiate a new exploration path (C).

measures beyond the mode are often meaningless in nominal data, but variation methods offer a handful of analysis options. We focus our utilization of statistical measures on nominal data methods since these are the greatest limiting factor of all data types. Applicable measures of variation for nominal data use the concept of diversity rather than variation from a central tendency measure. These diversity measures subscribe to two basic principles: (1) variation increases as relative probabilities become increasingly equal, and (2) variation typically is normalized over the range of $[0, 1]$ [18]. The simplest approach, variation ratio, determines the proportion of the number of items outside of the mode against the total number of items [19]. *Normed entropy (H*)* [20] and the *index of qualitative variation (IQV)* [18] are continuous functions of all probabilities, but their results can embellish the variation in the data. The *coefficient of nominal variation (CNV)* is a measure that extends the robustness of *IQV* and *normalized entropy* proposed by Kvålseth [21]. The *coefficient of unalikeability* is a relatively new method that measures how often values differ in pairwise comparison within a dimension [22]. In contrast to the previously mentioned methods, it can measure the variation of both quantitative and qualitative data without the need for binning the former. We utilize this measure in our approach. Wilcox introduced several similar measures of qualitative variation; *VarNC* and *StDev* indices are analogous to quantitative measures of *variance* and *standard deviation*, which we utilize alongside the previously mentioned *coefficient of unalikeability* [23]. Additional closely-related measures of variation are domain-specific. For example, Simpson's diversity index [24] in ecology and social sciences measures population diversity. Since such measures utilize slight variations of the same formulae, we do not include these.

Visual Analysis Approaches for Qualitative and Mixed Data. While some works have explored the integration of nominal-focused descriptive statistics, their solutions have been primarily limited to domain-specific scenarios. One approach by Pearlman et al. [25] represents variation as diversity using glyphs. However, their technique does not scale well for many-attribute datasets, which are the target of our approach. Simpson's diversity index, a measure similar to the *coefficient of unalikeability*, is utilized by Mackin and Patterson [26] in analyzing diversity of opinions in large social media networks. However, this is designed solely for network data. Diversity Maps by Pham et al. [27] use normalized entropy as a diversity measure for the ecology domain to visualize large-scale multivariate data with a heatmap-based approach. An extension to this approach by Wee [28], the adaptive diversity table, implements a broader set of interactions and color schemes. Each of these approaches are domain-specific, and emphasize only one core statistical measure (diversity). Our approach integrates several variation measures while using a different set of visual encodings. Finally, Blade Graph by Kobayashi et al. [29] measures the frequency-based variation from a central tendency measure to create color-coded area plots based on the calculated magnitude of variation. While this approach allows comparison of subset distributions, it is limited to quantitative data. Our Integrated Dual Analysis approach leverages statistical measures applicable to both quantitative and qualitative data, and is generalized for analysis of any domain.

4 Integrated Dual Analysis Approach

Real world data is often complex and mixed. Advanced visual analytics tools, like the existing Dual Analysis approach, provide

only limited means for analyzing these data. Filtering quantitative data items with qualitative data subset selections is possible, but correlations between quantitative and qualitative data and vice versa cannot be investigated. Instead of separately analyzing these data, our introduction of qualitative data to a common statistical analysis amounts to a significant extension of the analysis space by a novel freeform approach to hypothesis generation. Our encoding for qualitative data includes only nominal data; numeric ordinal data, since they are more closely related to numerical data in terms of orderability, are grouped with quantitative data. We provide a conceptual overview of our approach in Fig. 2, where mixed data in tabular form initiate the analysis pipeline. Users are then provided with an **overview** of the data for simultaneous dimension and item exploration, where quantitative and qualitative data are visualized together (Fig. 2 (A)). **Subset selection and analysis** (Fig. 2 (B)) allows the user to interactively select a data subset and to track the resulting changes in descriptive statistics for all dimensions. Visual feedback on individual and combined deviation measures for each dimension in dimension plot as well as for the data items in the item plot allow for identifying possible correlations with the subset selector. These steps may be **reset or repeated** (Fig. 2 (C)).

4.1 Measuring Variability

One of our aims in the Integrated Dual Analysis approach is the implementation of statistical measures that describe both quantitative and qualitative data. Intradimensional variability is a key measure for understanding data spread or differences. This provides implications on data quality and the nature of a population. For instance, a patient cohort comprised of only smokers will paint a very different picture of health than the entire population. Variability may also be measured from different perspectives: it can describe the variation around a measure of central tendency, such as the mean or the mode, or it can describe overall data variability by pairwise differences between data items – the latter is known as diversity [22]. Both types of variability can be equally applied to quantitative and qualitative data. For variation around the central tendency we use *standard deviation* and *variance* with their corresponding qualitative data analogs, i.e., *stDev* and *variance analog (VA)*. To measure diversity we use the *coefficient of unalikeability*, which is universal for both data types. For consistency in the following measure descriptions we use n to represent the sample size, while k indicates the number of categories.

4.1.1 Diversity as a Measure of Variability

The *coefficient of unalikeability* (u) is a relatively new statistical measure that offers a natural perspective on variability. Rather than measuring variability from a measure of central tendency, unalikeability instead measures variability *within* the data. This does so by representing the proportion of possible individual data pairings x_i and x_j that are unalike for a finite number of observations n [22]. Presented as a proportional value ranging from 0 to 1, this measure provides insights on *intradimensional* diversity. Formally, the *coefficient of unalikeability* is defined as:

$$u = \frac{\sum_{i \neq j} c(x_i, x_j)}{n^2 - n} \tag{1}$$

where

$$c(x_i, x_j) = \begin{cases} 1, & x_i \text{ is not alike } x_j \\ 0, & x_i \text{ is alike } x_j. \end{cases} \tag{2}$$

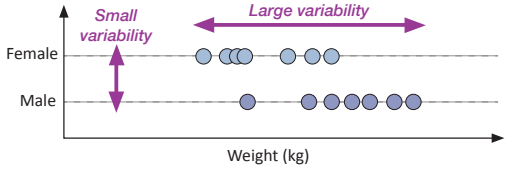

Fig. 3. Illustrative two-dimensional dataset opposing weight and gender to demonstrate the *coefficient of unalikeability*, a measure of dimensional diversity that examines pairwise differences between data items. While unalikeability is high for *body weight* (assuming a small threshold ϵ), the same measure of variation is low when computed for *gender*. This indicates, sensibly, that the data are more diverse, i.e., show more variation in the body weight dimension while gender is homogeneous by comparison, since there are only two options to choose from in this dimension.

An increased u indicates a diverse dataset. A value of 1 indicates that all values are unique, while a value closer to 0 means that all values are more similar. Fig. 3 demonstrates an example plotting two dimensions, weight and gender. We can see that u is high, indicating large variability, when computed for the body weight dimension. However, for the gender dimension it is much lower, which suggests a smaller variability. This tells us that our data are more diverse, or are more unique, in the weight dimension while the gender dimension is comparatively homogeneous. This makes sense, since weight as a quantitative variable can be expressed as any value within an expected range, while gender usually has only two options, male or female.

As demonstrated in Fig. 3, we can see that quantitative data have a tendency to show higher unalikeability relative to their qualitative counterparts. Although it may be desirable in some cases to allow this natural tendency to show in analysis, it may also be useful to allow a looser interpretation of "unalike" for quantitative data. This provides closer alignment with qualitative data measurements. In the comparative example with weight and gender, a less specific interpretation of unalikeness for weight would result in more similar measures of unalikeability for both dimensions. In larger, more complex datasets, such a loose interpretation could act as a means for noise reduction in quantitative dimensions, and emphasize only meaningful variation. While several solutions are possible, we introduce a simple user-adjustable threshold-based definition as a means to introduce flexibility in interpreting pairwise equivalence between items in quantitative dimensions (Eq. 3). The threshold is initially set to five percent of the range within the variable. We derived this number empirically in finding a reasonable trade-off between noise removal and unalikeability representation.

$$c(x_i, x_j) = \begin{cases} 1, & |x_i - x_j| > \epsilon \\ 0, & |x_i - x_j| \leq \epsilon \end{cases} \tag{3}$$

4.1.2 Variation Around the Mean/Mode

In addition to measuring the diversity in a dataset, it is also often useful to calculate the spread of data from a measure of central tendency. Variation around the center point(s) of a dataset provides a different perspective than that given by the *coefficient of unalikeability*; in this instance, we can understand the shape, or width, of the data. In the context of waist circumference values in a patient cohort, we can use central tendency variation to tell us how close a given measurement is to the mean waist circumference reading. A low variation measure informs us that the cohort waist circumference readings are tightly packed around the mean,

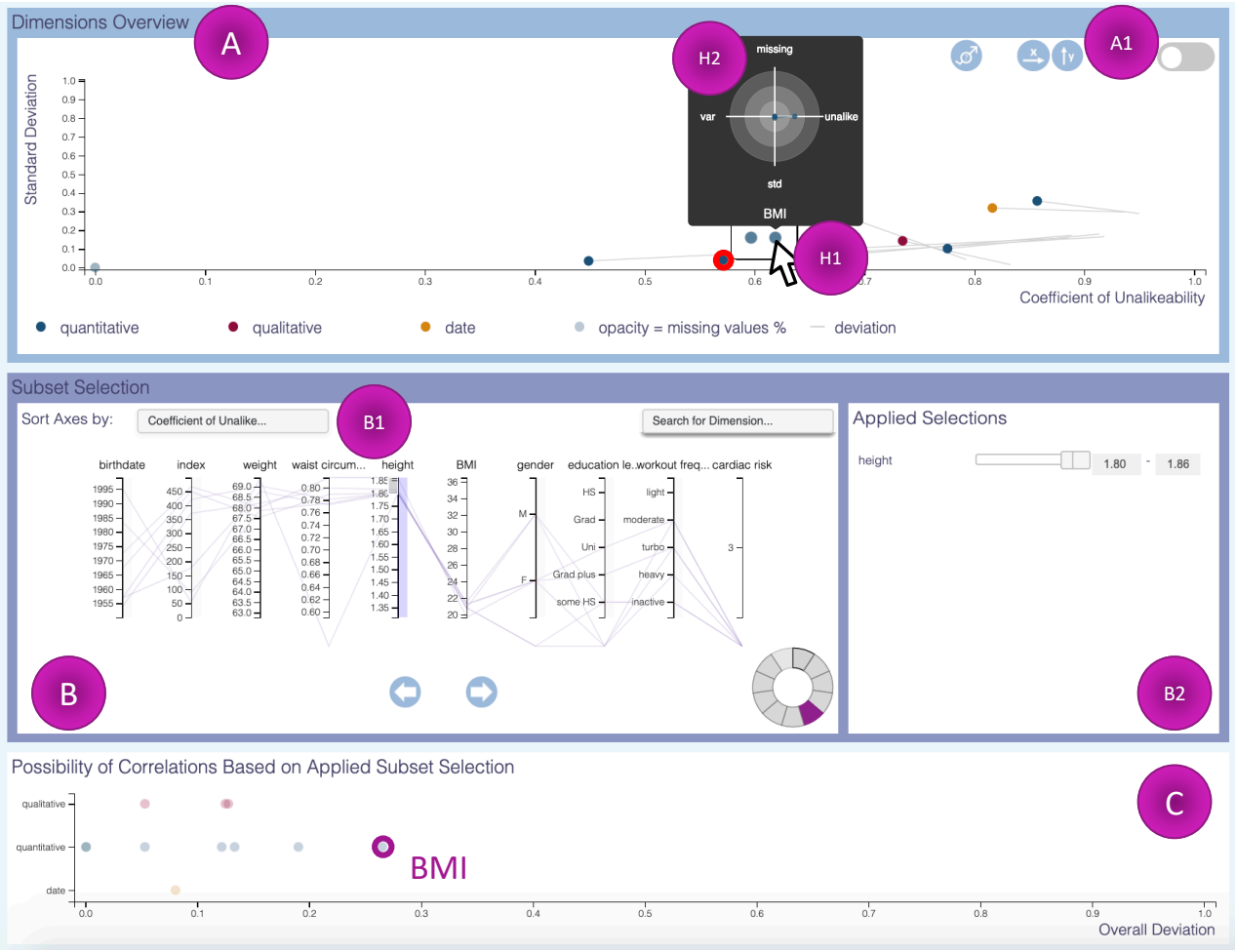

Fig. 4. Prototype demonstrating the Integrated Dual Analysis method, a visual approach for the joint exploratory analysis of quantitative and qualitative data using comparable descriptive statistics. Such treatment permits visual integration for exploratory analysis at both the dimension (A, C) and item (B) levels. Brushing and linking between these levels allows users to select dimensions of interest in dimension space and emphasize the selection in item space, to gather items of interest in an adapted parallel coordinates plot (B1), to store and adjust these selections (B2), and to observe the deviations in statistics and missingness calculations between the entire dataset and the selection. Buttons assist in dimension investigation using different statistical measures or exclusion of missing values (A1). Light grey lines indicate the magnitude and direction of the deviations in the dimensions overview (A). The Possibility of Correlations plot (combined deviation measures) presents an aggregate view of all statistics and missingness information supporting in the investigation of outliers (C) with possibly correlated dimensions, e.g., BMI. Tooltips ease visualization of overplotted regions (H1) as well as a more detailed presentation of the individual statistics for a hovered dimension (H2). This information may be used to inform scatterplot axes choices for more detailed examination of dimension behaviors in (A).

which can indicate either a naturally homogeneous population or a poor sampling method. *Standard deviation* and *variance* are the most familiar measures in this category for quantitative data, and are utilized in our approach as well. They describe variation around the mean, where 0 indicates completely homogeneous data. Unfortunately, such measures are not immediately applicable for qualitative data. Nominal data have no intrinsic ordering, while ordinal data are described as discretized values or categories. However, analogous measures for *standard deviation* and *variance* for qualitative data, known in part as the Wilcox Indices [23], exist as *stDev* and *variance analog (VA)*. In our approach, we use the inverse of these measures to correspond value-wise with their quantitative counterparts, where 0 is demonstrative of an identical set of data items. The *VA* measure is an analogous measure to *variance* when the data are normalized to $[0, 1]$ [30], and is robust to uncertainty when sampling variability is low. The inverse of *VA* is defined in Eq. 4, where f_i is the frequency of the i^{th} category.

$$VA = \frac{\sum_{i=1}^{k} \left(f_i - \frac{n}{k}\right)^2}{\frac{n^2(k-1)}{k}} \tag{4}$$

The analogous qualitative measure of *standard deviation* for normalized data is defined as *stDev*, which can be represented as the square root of the *Mean Difference Analog (MDA)*, another of the Wilcox Indices [23], when analysis includes nominal data. *MDA* is given as *"the average of the differences of all the possible pairs of variate-values, taken regardless of sign"* [31]. Like *VA*, this measure exhibits a low sampling variability, and so is also robust to uncertainty. The equation for the inverse *stDev* is:

$$stDev = \sqrt{\frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (f_i - f_j)^2}{n^2(k-1)}} \tag{5}$$

where f_i and f_j are the frequencies of the i^{th} and j^{th} categories in the sample, respectively.

Fig. 8 compares variability and central tendency in a scatterplot. Given the relatedness of *standard deviation* and *variance* we see the expected parabolic shape, but our interest and questions lie in the groupings of the dimension points: *"Do we see clusters of dimensions that have similar variability? Is it typical for these dimensions to have this degree of variability? Do these*

dimensions with similar variability measures remain together as we subset the data?" For instance, in this patient population, do we expect to see, e.g., *"diagnosis group, exhibiting high variability?"* This could indicate a diverse patient population with numerous diseases, or an inconsistent method of data entry in the case of free form text. Discoveries like this, using variability from the perspective of diversity or variability from central tendency, can lead to new interesting hypotheses to explore in more detail.

4.2 Understanding Modality

Modes can be used as reliable measures of central tendency for both quantitative and qualitative data types. Furthermore, we can learn about the shape of the data distribution through an analysis of the modes of a distribution. For example, a multimodal distribution of blood pressure data in a patient cohort would be an unexpected and highly interesting finding: *"What could be the reason for a cluster of values that are different from the "normal" healthy result?"* We apply Kernel density estimation (KDE) for quantitative data, while for qualitative data we use thresholding for high-frequency categories. KDE $\widehat{f}_h(x)$ is a popular statistical method for estimating the underlying distribution function of a given data sample. Given a set of n univariate samples consisting of $(x1, \ldots, xn)$ real numbers, we can estimate the shape and density of this function (Eq. 6):

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big) \tag{6}$$

where K is a non-negative kernel function, integrating to 1, while h is a smoothing parameter, often called the bandwidth of KDE. The choice of bandwidth h is strategic and absolutely critical (while different choices of the kernel function usually do not have a strong influence). An undersmoothed KDE will overemphasize the individual data points, while oversmoothing flattens the distribution and obscures meaningful variations. The specifics on choosing h are beyond the scope of this paper, but can be explored in more detail in works specific to KDE, such as those by Chiu and by Florek and Hauser [32], [33].

Our approach uses the Freedman-Diaconis estimator [34] to calculate the bandwidth for the KDE. This method makes use of the *Interquartile Range* (IQR), rather than the *standard deviation* as applied in Silverman's rule of thumb [35], so it is more robust to skew and variation within the data. The bandwidth is calculated from the number of observations n in the data sample x:

$$h = 2\frac{IQR(x)}{\sqrt[3]{n}} \tag{7}$$

The number of modes present within each KDE dimension can then be calculated by finding the number of local maxima. In qualitative distributions, we can use a straightforward method to determine the mode of nominal and ordinal data. We do this by a simple check of the most frequent category, minus a small user-adjustable percentage, e.g., 10%, to set the threshold of qualitative observations that fall above this value.

Fig. 7 shows a graphical example of how modes are used in the Integrated Dual Analysis approach. Dimensions with a high number of modes can provide another indicator for a diverse dimension distribution. In these outlying cases, the user may then investigate if this outlyingness results from the description of the dimension itself, i.e., *"Is it expected that liquor examination would have a multimodal distribution? Is that normal for the recording modality, and for this population sample?"*

4.3 Visual Representation

We base our visual approach primarily on position and hue (pre-attentive stimuli) to encode data elements and relationships in both dimension and item space. We do this as a means to reduce complexity of the targeted data (Fig. 4) [36]. We plot all quantitative and qualitative data jointly in both spaces. Joint plotting along a common axis emphasizes ease of comparison for all data items; position along a common axis has been shown in prior graphical perception studies as the most effective visual channel [36].

Dimension Exploration. In dimension space we use linked scatterplots. These effectively use common axis positioning while allowing an easy visualization of pairwise correlations for large datasets. Scatterplot axes may be exchanged between different statistical measures. We additionally allow plotting of data by their proportion of missing items in each dimension. For this view we chose individual scatterplots over scatterplot matrices (SPLOM). While SPLOMs allow simultaneous visualization of multiple measures, we found them to be visually overwhelming and difficult to brush over specific regions of interest with limited screen space. Our approach utilizes three different scatterplots (Fig. 4 (A and C), Fig. 6), each of which serves a specific role in hypothesis generation. All three are connected by brushing and linking. This allows users to select the appropriate measures to place on each axis while preserving maximum screen space. To improve the visual differentiation of dimension points, we apply focus and context-based hovering techniques, inspired by *interactive lenses*, for regions characterized by overplotting issues [37]. Instead of a force-directed jitter plot distortion, we use an unsorted grid-based layout for glyph positioning. This allows for easier distinction between dimensions (Fig. 4, (H1)). The related dimensions are emphasized with a red circumscription.

We incorporate radar plots displayed while hovering over a dimension point to represent detailed statistical measures of the dimension (Fig. 4 (H2)). Our use of radar plots is based on the spatial pattern-recognition facilities this plot offers [38], which has been shown to be useful particularly in multivariate medical studies for hypothesis generation and verification [39]. In exploring the statistics for each dimension, users may recognize common shapes formed by the radar plot. This may indicate relatedness of dimensions by multiple measures.

In dimension space, hue provides a distinction between the data types; blue, yellow and red indicate quantitative, date, and qualitative dimensions, respectively. We draw a visual differentiation between these elements to give the user feedback about how many quantitative, qualitative, and date dimensions are included in the dataset. This allows for rapid observation of outlying dimension composition and pattern observation. Although dates are read as quantitative data, we declare a different hue for this data type, since dates may be particularly useful for identifying subsets, e.g., a patient cohort with a series of MR scans over a period of years, where a specific year is of interest. Our visual distinction allows for easy interaction facilities to isolate date(s) of interest through subset selection mechanisms.

Item Exploration. While the dimension overview offers one data perspective, it is also useful to obtain an overview of item correlation between dimensions. Numerous graphical methods have been proposed to visualize high dimensional data, such as Parallel Coordinates Plots (PCP). This is a standard graphical tool for multidimensional visualization and correlation that make

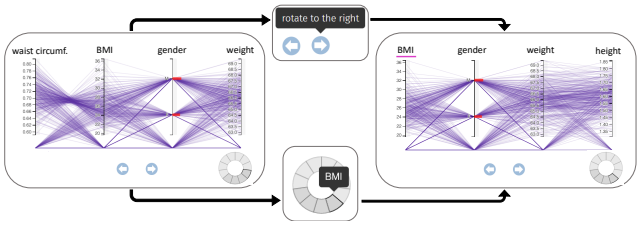

Fig. 5. Interaction facilities allow for navigating all dimensions in the items plot. When clicking on a rotation button or on a dimension in the context donut-chart-based glyph, the parallel coordinates plot automatically rotates to the selected dimension in an animated fashion.

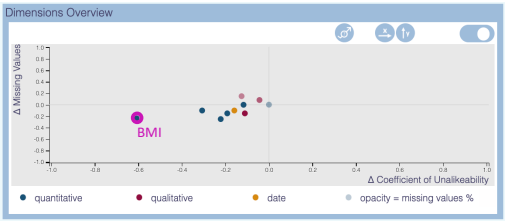

Fig. 6. Deviation plot of all dimensions with the statistical measures *coefficient of unalikeability* and *relative frequency of missing values* resulting from subset selection of people taller than 1.80m. In this view, we are interested in looking at the dimensions that show the largest deviations (in this case, e.g., BMI).

efficient use of space [40]. A number of variations on the traditional parallel coordinates plot exist for the visualization of data containing, e.g., 20 or more dimensions. Bifocal parallel coordinates, presented by Kaur and Karki [41] constitutes one such approach by splitting the parallel coordinates plot into focus and context regions, where dimensions of interest are mapped having sufficient space and the remaining dimensions are presented in a compact manner. However, with dimensions numbering in the hundreds, the simultaneous presentation of all dimensions fails to display a coherent focus and context view, as the context area would be nearly opaque with the necessary degree of condensation. We instead propose a carousel-inspired plot, similar to the perspective walls technique [42], to visualize many dimensions without overcrowding (Fig. 4 (B1)). The parallel coordinates plot automatically rotates to the dimension which has been brushed in the dimension view, or rotates incrementally when pushing a rotation button (Fig. 5). Our approach, paired with sorting by the aforementioned statistical measures (Sec. 4.1 and Sec. 4.2), preserves an appropriate data-to-ink ratio without increasing cognitive load or introducing visual distortion in axis height that has been problematic in other methods [41], [42].

Since our data include both quantitative and qualitative dimensions, we utilize a parallel bubbles [43] technique, as opposed to parallel sets which are more conducive to qualitative-only data [44]. We use red bars to better localize the categories along qualitative dimensional axes, sized by relative frequencies (Fig. 4 (B1)). This color choice helps to distinguish qualitative from quantitative dimensions, since the interface is predominately in cool tones.

The axis background color in a shade of grey encodes the corresponding value of the selected descriptive statistical measure for sorting (Fig. 9). Interaction techniques such as axis reordering, rotation and dimension-search further assist in item exploration. Furthermore, orientation within the carousel-inspired view is provided through a context donut-chart-based glyph at the bottom right, in which the current set of visualized dimensions is shown in dark grey (Fig. 5). Dimensions with subset selections change to a purple color for quick identification.

Subset Selection and Analysis. Subset selections are made initially in the parallel coordinates view. These selections may be refined in the same view or in the subset summary panel (Fig. 4 (B2)) using range sliders for each selected dimension. Subset selection with the aid of parallel coordinates provides additional information that is impossible to obtain with range sliders alone. Parallel coordinates enable the user to quickly discover dimensional correlations between the variables represented by neighboring axes. Secondly, subset selection in one dimension causes simultaneous selection in all other dimensions. This "unintentional co-selection" is visualized in the PCP but cannot be seen solely based on range sliders. After subset selection the changes among descriptive statistical measures of the dimensions are emphasized by light grey lines. These lines point from the original descriptive statistical measures, based on the whole dataset, to the measures for the new subset (Fig. 4 (B1)). A scatterplot-based deviations plot is also available after clicking on the toggle button (Fig. 4 (A1)). In this view, all dimensions are plotted by their deviation in descriptive statistics, as resulting from the subset (Fig. 6). This allows the user to easily visualize the degree and direction of change in statistics, which is helpful for hypothesis formation. Additionally, the *possibility of correlations plot* (Fig. 4 (C)) presents the dimensions' overall deviation. This is defined as the normalized aggregation of the deviations among all descriptive statistical measures resulting from subset selection sorted by data type. This serves as a concise presentation of the most likely correlated dimensions with the subset selector.

Missingness. As previously discussed, the degree of missingness in the data is crucial to understand in the formation of hypotheses. Our choice to include, rather than exclude, missing data is predicated on the idea that missing data are data elements themselves. These deserve visual representation, as they can provide invaluable information about the reliability of an item or dimension. Dimensions with predominately missing data may be less reliable to use for analysis, but reviewing and filtering out the missing items may produce an interesting and reliable subcohort for exploratory analysis with that dimension. We utilize different approaches for conveying missingness in dimension and items space. In dimension space we use transparency to encode the relative frequency of missingness for each dimension. More transparent items correspond to more incomplete dimensions. As we expect most real-world data to exhibit a degree of missingness, our use of transparency allows us to emphasize clustering of dimensions around certain descriptive measure values, since these regions will be more opaque [45]. In item space we treat missing data as real values, displayed at the bottom of the parallel coordinates plot. Our inclusion of missing values as visual design elements provides users with this crucial information so that they may make informed choices based on the certainty of their data. For the statistical analysis, the user can toggle between integrating and excluding missing values since two missing values cannot always be interpreted as the same value. This especially has an impact on the *coefficient of unalikeability* since this measure either treats missing values as a category or not.

5 Visual Hypothesis Generation

We demonstrate the general process of hypothesis generation using our approach in Fig. 2. This process includes the simultaneous

exploration of items and dimensions to obtain an overview of the data, subset selection and analysis, and investigation of sub-space deviations.

Simultaneous Dimension and Item Exploration. Our approach to hypothesis generation enables an overview at two different intrinsically linked levels: (1) dimension space, and (2) item space (Fig. 2 (A). Beginning in dimension space, one can freely explore the entire dataset at a bird's eye level. With this, the user can quickly develop an overview of common characteristics of the data, including general spread and data type characteristics. In this context, dimension space is useful for identifying dimensions with special features (structure of missing values, special distributional properties, etc.). Dimension space also allows for observation of the structure between the dimensions, answering questions like: *"Are there groups of dimensions that show similarities? Are there 'outlying dimensions'?"*

In practice, high-dimensional datasets often come with a high portion of missing data. There is a need for visual facilities to reveal the degree and possible patterns of missing data. This lack of information has an understandably high impact on hypothesis generation, although recent efforts have begun to cope with this information gap. Alemzadeh et al. [46], for example, developed new approaches to visually encode missing data in item space. However, there is no analogous approach to emphasize missing data in dimension space. Integrated overviews of all missing data, for all data types, can give a sense of the completeness and quality of the dataset.

Statistical measures for dimension characterization and comparison allow for hypotheses generation. For example, by investigating outlying or grouped dimensions, information about their similarity can be gathered: *"Do they share similar properties?"* These statistical measures may be observed individually in the dimension plot, or simultaneously through the radar plot revealed when hovering over a dimension.

The item level provides facilities for further investigation of the dimensions within item space. This allows for in-depth exploration of distributions and inter-dimensional correlations, as well as outlier and missingness identification. Axis reordering based on the same statistical measures and missingness frequency facilitates inter-dimension comparison. Users may also begin directly with this view and generate hypotheses within this area by creating subsets of the data. These subsets can then be used for analysis in the dimensions overview. The core idea with both overviews is to smoothly move between both dimension and item space in this exploratory analysis phase, where quantitative and qualitative dimensions are studied together.

Subset Analysis. Zooming and subsetting provide a means of refining hypotheses formed in the overview process (Fig. 2 (B)). At this stage, users can selectively view dimensions of interest deemed valid for the current hypothesis. Subset seletcion in item space allows for redefining the desired subcohort of data. While brushing, the related changes in descriptive statistics emphasize possible correlations. Here, we want to follow the change in any outlying dimension or outlying group: *"Does a group move together?"* and *"Which impact do the applied selections have on statistical measures of all dimensions (both quantitative and qualitative dimensions)?"*

Analysis of Sub-Space Deviations. After selecting a data subset, dimensions with related large deviations are highlighted regardless of their data type, since they convey possible correlations (Fig. 2 (B)). Inspection of the possibility of correlations plot (Fig. 4 (C)) provides a rapid summary of dimensions with the largest deviations. The radar chart on hover reveals a detailed breakdown of descriptive statistics for each dimension of interest. This information can be used to analyze possible correlations by individual descriptive statistics in the deviation plot. A review of the missingness, through the radar plot or in the deviation view, for each dimension provides a mechanism to understand the reliability of the indicated hypothesis. Identification of possible correlations through large deviations and outliers can be a major step in hypotheses generation.

6 Case Study

Clinical cohort data represent an ideal use case to demonstrate our approach since they are typically highly complex and mixed (i.e., contain a combination of quantitative and qualitative attributes) with many missing elements [47]. For this case study, we employ a clinical routine dataset of 307 patients suffering from Cerebral Small Vessel Disease (CSVD) recorded at the university hospital of Magdeburg, Germany. CSVD is an umbrella term for abnormalities related to small brain vessels, such as white matter hyperintensities, microbleeds, lacunes, subcortical infarcts, and enlarged perivascular spaces (EPVS). It has been reported as an important precursor of stroke, dementia, cognitive decline, and psychiatric disorders [48], [49]. The research around biomarkers in CSVD is complex and highly debated – biomarker expression is highly variable, even among patients with similar risk profiles, for the disease. This dataset comprises 193 mixed dimensions related to demographic information, laboratory results, genetic data, education, and lifestyle, as well as 24 dimensions derived from medical images of the patients. In a volumetry analysis of T1-weighted Magnetic Resonance Imaging data using FreeSurfer [50], the volume of 24 brain structures, e.g., hippocampus, putamen, and caudate, was determined. Missing values comprise 76% of the data among dimensions of all data types.

In joint paired analysis sessions with two medical experts we investigated this clinical routine data using our Integrated Dual Analysis protoype implemented as an interactive client-server web environment utilizing JavaScript, D3.js [51], Python, and Flask framework [52]. We chose a joint paired analysis to understand the cognitive processes of the domain experts and allow for a fluent data exploration and hypotheses generation [53]. One expert participant is a senior neurologist with more than ten years of experience in this area of research and mainly assisted in designing the tool. The other participant is a medical student specializing in the study of CSVD with one year of experience. Both domain specialists are co-authors of this paper. Although CSVD is well-studied, numerous open questions remain. Our collaborating medical experts are especially interested in finding reasons why CSVD patients with similar disease risk profiles may express different biomarker patterns. Their standard analysis approach would be to utilize advanced and mature statistical analysis tools, such as SPSS [6], to identify such biomarker relations.

Prior to exploration with our approach, we observed one expert analyze the data using their standard approach with SPSS. The purpose of this analysis was to establish a baseline for the analysis effort in an exploratory approach, relative to our method; we describe the key takeaways and results with this goal in mind. The paired session lasted two hours in total, and followed a "think-aloud" protocol. Since the expert was unfamiliar with this dataset, they were chiefly interested in establishing a general picture of

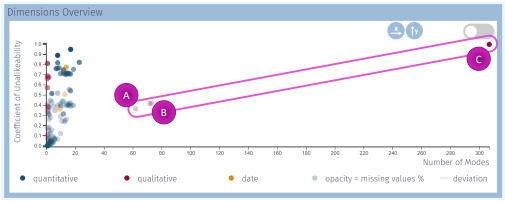

Fig. 7. Four dimensions (*pathology finding* (A), *liquor examination identifiers* and *liquor sample identifiers* (B), and *patient id* (C)) within the dataset contain a significantly outlying number of modes (circled), which is linked to a larger unalikeability within these dimensions.

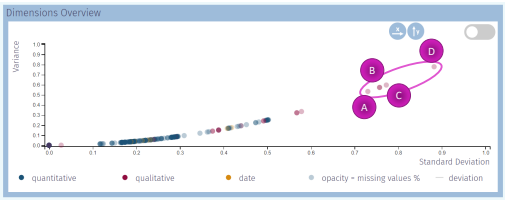

Fig. 8. Comparison of *standard deviation/stDev* and *variance/VA* among all dimensions. We can see two primary groupings of dimensions, divided at about 0.6 along the x-axis. The group of dimensions with higher values for both measures include: *diagnosis group* (A), *fluid-attenuated inversion recovery (FLAIR) imaging sequence* (B), *recorded diagnosis* (C), and *susceptibility weighted imaging modality (SWI)* (D) (circled region). The notably larger intradimensional spread for this grouping may be interesting to follow up on in later analysis.

patterns in the data, centered around the dimension "Group", which specifies patient diagnostic subgroups exhibiting different pathologies of CSVD. The expert's main method of doing this was a series of comparisons of diagnostic test and lifestyle variables for each diagnostic group. The building of each of these factor tables, with subsequent analysis of these tables, was extensive, and required 1.5 hours. Through this process, the expert began to build a general sense of the typical patient characteristics for each diagnostic subgroup. Building of these factor tables is a lengthy process that is often conducted over several sessions. The expert explained that their next series of steps would be pairwise or factor analysis on each of the variables that appeared interesting from these lifestyle and diagnostic variable tables. This analysis sequence would aim to further identify interesting correlated variables in several iterations with different subsets. The expert stated that this is a lengthy and cumbersome procedure that may be spread out over the course of several days.

Although SPSS produced statistical results for a priori hypotheses in a compact form, with basic visualization options, we observed that selecting and iterating analyses over various possibly interesting subcohorts was inefficient and cumbersome. Establishing correlation between dimensions of different data types requires different measures, each of which takes time to identify and run. When interested in comparing a subset of the data to the whole, there also is a time cost in identifying dimensions of interest for analyzing the subset relative to the entire dataset. The expert noted that they found SPSS most powerful if they already have a hypothesis established with variables of interest; using SPSS to simply explore a dataset for interesting relationships is not part of their ideal workflow. By investigating only a priori hypotheses, however, interesting new relationships can be missed in the process and therefore, an exploratory analysis for hypotheses generation is preferred.

We now demonstrate our approach for exploring and reasoning about interesting correlated subcohorts and outliers prior, or in addition, to a more targeted analysis in SPSS or similar tools with two medical experts.

Data overview. In our exploratory analysis, we focus on exploring patient demographic and test score variables alongside imaging-related dimensions. In the course of this exploration we take note of the degree of missing data for each variable we analyze, since this has an impact on the reliability and reasoning for hypotheses we can generate. The first task is dimension exploration, where we investigate the data using descriptive statistical measures to obtain an overview of all dimensions.

Interestingly, we see that the number of modes within each dimension acts as an indicator of central tendency for both quantitative and qualitative data types. By investigating the modal distribution of the data, we observe that there are four dimensions with notably more modes (*pathology finding*, *liquor examination identifiers*, *liquor sample identifiers*, and *patient id*) as shown in Fig. 7. The likely reason for such a high number of modes can be confirmed by high unalikeability within these dimensions. Since the *pathology finding* is recorded as free text and identifiers are naturally unique, these dimensions contain almost only unique values. We expect these dimensions to be outliers then when plotted with these types of descriptive statistics, and our expectation is verified visually.

By investigating *standard deviation* and *variance* in comparison, clusters of dimensions become visible (Fig. 8). Dimensions in the encircled cluster with higher *standard deviations* and *variances* include *susceptibility weighted imaging modality (SWI)*, *recorded diagnosis*, *fluid-attenuated inversion recovery (FLAIR) imaging sequence*, and *diagnosis group*. This shows us that these dimensions are more spread out and could be interesting for further investigation into relatedness. It could be that the higher values for these dimensions arise from the entry type, i.e., free form text entry, or are indicative generally of a broad range of conditions to select from, as in the *diagnosis* dimension. Through this type of graphical exploration we can easily visualize outlier dimensions and make inferences about the range and consistency of the data input method in clinical routine.

APOE. Following our data overview we want to identify possible correlations that could merit further exploration. We first investigate the *apolipoprotein E (APOE) genotype*. This genotype is a combination of *APOE-A1* and *APOE-A2*, and is a known risk factor for *cerebral amyloid angiopathy (CAA)*, a combination of pathologies forming a subtype of CSVD. We select the four patients with an *APOE genotype* of $[4; 4]$ in the item (parallel coordinates) plot (Fig. 9-1,2). Observing the changes in descriptive statistics in the dimension scatterplot (Fig. 9-3), we note that the largest changes arise from the *coefficient of unalikeability* and *percentage of missing values*. The contextual radar plot (Fig. 9-4) that is available by hovering over each dimension assists in identifying these large changes, alongside the trail lines extending from the dimension plot points. Setting *unalikeability* and *percent missingness* as the x- and y-axes, respectively, we switch to the deviation plot view to better observe the deviations resulting from the *APOE* subset selection (Fig. 9-5). Within the deviation plot we note two clusters, one cluster consisting of more complete data, and containing the *APOE*variables that defined our subset.

We then use the *possibility of correlations* plot to clearly identify dimensions with the greatest overall deviation, and thus most likely correlated with the *APOE* subset (Fig. 9-6). The top

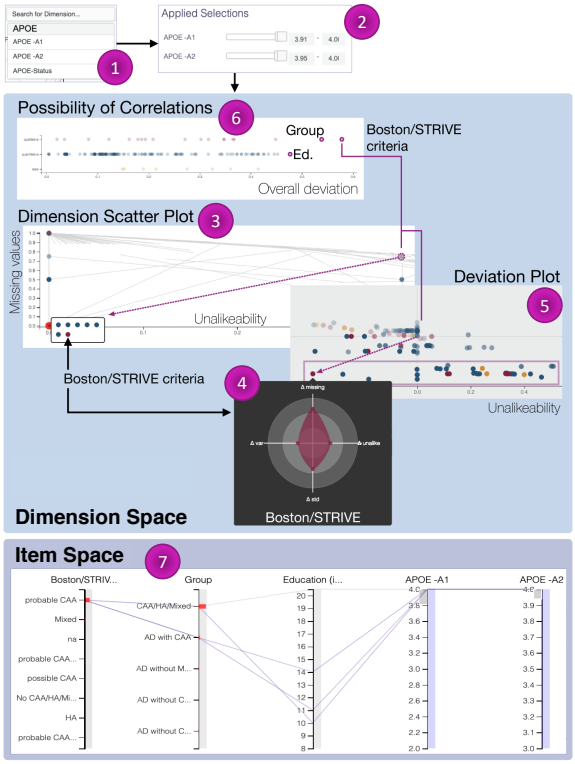

Fig. 9. When selecting only patients having a combination of the known risk factor *apolipoprotein E (APOE) genotype-A1* and *APOE-2* of $[4; 4]$, large value changes in descriptive statistics for *Boston/STRIVE criteria*, *Group (CAA/HA/Mixed)*, and *education in years* become visible. Inspection in the parallel coordinates with axis ordering by overall deviation allows for inspection of correlation direction.

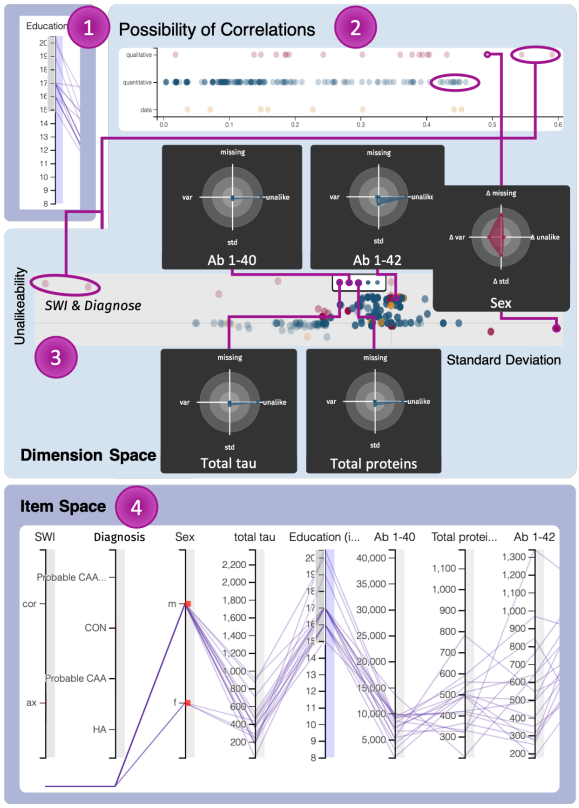

Fig. 10. Selecting patients having more than 15 *years of education* highlights possible correlations with *sex*, *total tau*, *amyloid-beta 40 (Ab 1-40)*, *total proteins*, and *amyloid-beta 42 (Ab 1-42)* in the cerebrospinal fluid. *Susceptibility-weighted images (SWI)* and *diagnosis* are neglected since they include missing values only after subset selection.

dimensions with the greatest changes include: *Boston/STRIVE criteria*, *Group (CAA/HA/Mixed)*, and *education in years* (Fig. 9-6). Although *education in years* is quantitative, we note that *Boston/STRIVE* and *Group (CAA/HA/Mixed)* are qualitative, and their relations would not have been observed with common statistical tools (at least not without searching for this relation explicitly in these tools).

It is established clinical knowledge that *APOE 4* is a genetic risk factor for CAA and correlates with a positive diagnosis of CAA applying the *Boston criteria* [54], [55]. Since the *STRIVE criteria* are imaging features for all CSVD subtypes and not specific for CAA, they are not applicable in this case. After subset selection, the subgroups are not equally distributed and are rather small, a clear limitation, but recognizable through our framework. To validate the identified possible correlations, one should consider a more equally distributed and complete dataset or conduct a clinical study. In the future, testing for *APOE genetic variants* could be used as an early biomarker determining the risk and course of disease.

The relationship between *APOE* expression and the remaining dimension (*education in years*) in this group is well studied but very complex, and requires deeper investigation and validation [56]. Having narrowed our dimensions of interest, we can revisit and reorder the parallel coordinates axes by overall deviation for more information on the correlations between these dimensions (Fig. 9-7). Observing a possible negative correlation between *APOE 4* and *education*, we hypothesize that CAA, since this relates to *APOE 4*, may be connected to fewer *years of education*. However, this hypothesis requires detailed investigation with a larger expression cohort and additional cognitive data.

Education level. The possible influence of *education* on CSVD is a ripe area of inquiry in CSVD research; for some high-risk patients there is a possibility that increased *education levels* might mitigate the cognitive effects of CSVD pathology [57]. In the parallel coordinates plot we create a subset of patients with 15 *years or more of education* (comparable to a Bachelor's degree) (Fig. 10-1). To quickly identify the dimensions most affected by our *education* subset, we examine overall deviations in the bottom panel. The highest-deviating dimensions, and therefore the most likely to be correlated with high education level, are three qualitative (*susceptibility-weighted images (SWI)*, *Diagnose based on Boston/Strive criteria*, and *sex*) and four quantitative dimensions (*total tau*, *amyloid-beta 40 (Ab 1-40)*, *total proteins*, and *amyloid-beta 42 (Ab 1-42)*).

Hovering with the radar plot provides detailed information on how the descriptive statistics have changed for these dimensions. We set the x- and y-axes in the deviation plot to *standard deviation* and *coefficient of unalikeability*, respectively, to visually search for any possible clusters in our high *education* subset. Although no distinct clusters form, we see that *sex* has changed significantly by the two measures shown in the plot; we also note that the cerebrospinal fluid tests for *Ab 1-40*, *Ab 1-42*, *total proteins*, and *total tau* have increased in *unalikeability* with little change in *standard deviation* (Fig. 10-3). Although large deviations for *SWI* and *Diagnose* have been observed, these dimensions are neglected since they do not contain any values after subset selection (Fig. 10-4).

Revisiting the parallel coordinates panel and reordering axes by overall deviation allows us to discover more details on the

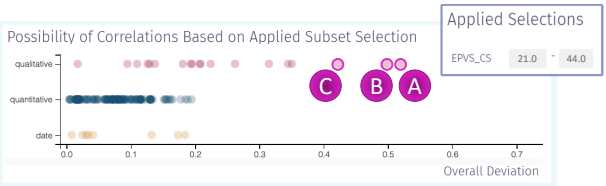

Fig. 11. Exploring the relationship between many (>20) *enlarged perivascular spaces* in the centrum semiovale and other dimensions, possible correlations with *diagnosis* (A), *susceptibility-weighted images* (B), and *Boston/STRIVE criteria* are indicated (C).

relationship of *sex*, *total tau*, *Ab 1-40*, *total proteins*, *Ab 1-42* with *education* (Fig. 10-4). We note that *sex* mostly is restricted to the male item for this subcohort; this is likely a product of the social norms of previous decades, where men were more likely to complete a university education.

Amyloid beta (Ab 1-40 and Ab 1-42) is a hallmark of the Alzheimer's disease pathology. We observe that by applying the *education level* selection ($\geq$ 15 *years of education*), *amyloid-beta* deposition is decreased in cerebrospinal fluid and thus, increased and having a higher pathology within the brain. This finding may support clinical literature that higher *education levels* provide resilience against disturbed amyloid metabolism (deposition or clearance mechanisms) and Alzheimer's disease [57]. Follow up with a detailed analysis of patient cognitive status and demographic information is a necessary next step to investigate the validity of this very interesting hypothesis.

Furthermore, we observe that the cerebrospinal fluid test biomarkers *total tau*, indicating neurodegeneration, as well as *total protein*, which is linked to disturbed blood-brain-barrier function, also decrease with higher *levels of education* [58]. This is an indicator for less pathology and thus, more resistance, within the brain. To the best of our knowledge, this relation is not well studied, and thus forms a new hypothesis valuable for further investigation by our medical experts.

Enlarged perivascular spaces. We lastly explore the relationship between *enlarged perivascular spaces (EPVS)*, an ordinal qualitative variable, evaluated in different brain regions. *EPVS ratings* for the *hippocampus (hc)*, *basal ganglia (bg)*, *centrum semiovale (cs)*, and *midbrain (mb)* are often reported for elderly patients and thought to be early indicators for CSVD, particularly in *bg* and *cs*. However, recent studies have found that this relationship is more ambiguous than previously thought [59], [60]. In the parallel coordinates plot we create a subcohort of patients with an *EPVS rating in cs* rating of 3 or higher (more than 20 *EPVS in cs*) and again check the deviation in descriptive statistics in the dimension deviation plot for possible correlated dimensions.

We note that there is some change in descriptive statistics for *diagnosis (probable CAA, CON, HA)*, *SWI.*, and *Boston/STRIVE criteria* with application of this subset selection (Fig. 11). With our findings that a higher *EPVS rating* is more likely related to CSVD *diagnosis (CAA or HA)* and a higher *Boston/STRIVE criteria score*, our dataset corroborates the established association to CSVD pathology.

SWI is a MRI sequence especially sensitive to venous blood and iron and useful for microbleeds detection [61]. *EPVS*, however, are usually counted in T1- or T2-weighted images. The medical domain experts rated this possible correlation as artificial and the *presence of SWI images* is related to logistic, patient-specific or medical reasons.

Our clinical collaborators were surprised that *EPVS* showed no correlation with, e.g., *APOE genotypes*,—this merits deeper investigation. [49]. Furthermore, they suspected correlations with the *EPVS ratings* in other brain regions. This observation, however, was not present within our dataset and forms a new and interesting hypothesis in CSVD research.

Practical impact. The domain experts emphasized the fast, straightforward, and iterative analysis process using our approach. Although it indicates *possible* correlations only, the domain experts state that it allows for identification of interesting patterns in "wide and shallow" datasets, such as clinical routine data, which would be hard or very time-consuming to analyze using common statistical tools, such as SPSS. These indicators can then form a basis for future clinical studies.

7 Discussion

This paper presents the successful use and integration of qualitative statistical measures into the joint analysis of quantitative and qualitative data. Through our case study of real clinical data we found a number of interesting patterns and relationships for the underlying patient cohort using our integrated approach. We were able to identify well-known relations and support previously known knowledge. Furthermore, we identified a hypothesis that is of interest for the clinician. Our approach allows users to get an overview about the dimensions within their (high-dimensional) data and to identify interesting variables for further investigation or exclusion from statistical analysis. By integrating a video tutorial before entering the framework, we addressed the challenges medical experts might have when moving from standard statistical tools, such as SPSS, to using our proposed framework. The case study participants stated that this movie fulfills their demands on the framework explanation but stated that users probably need to watch it multiple times to understand all functionalities. One concept that was hard to assess for medical experts in the beginning was the differentiation between correlation indicators and p-values. However, since our Integrated Dual Analysis is conducted after data gathering and before statistical analysis for hypothesis verification, the participants appreciated its value and especially emphasized the fast hypothesis generation using the tool.

The primary limitation to our approach lies in the restrictions we face in its application to all possible data types in a dataset. In our search we found a small handful descriptive statistics suiting our needs. Although it is substantiated that *stDev* and *VA* for qualitative data are analogous measures to *standard deviation* and *variance* for quantitative data, respectively, the depiction of the analogous measures on one axis may yield to confusion. To overcome this issue, we emphasize the dimension's data type via color and thus, also the related statistical measure. Since all of these mentioned measures result in a range of $[0, 1]$, we argue for showing them on one axis. Furthermore, we neglect well-known and widely applied descriptive statistics, such as *mean*, *skewness*, and *kurtosis*, because of the data type restrictions. However, in applying the proposed methods and demonstrating the identification of valuable hypotheses, we have shown the suitability of the measures applicable for all data types.

By adding extra parameters for the computation of the modality and variation as a measure of diversity, we introduce some degree of uncertainty within our results. We decrease this uncertainty by providing pre-defined parameter settings and by allowing users

to modify the thresholds for modality and unalikeability computation. Thus, the user is provided with direct visual feedback about the related deviations resulting from parameter adaptation. As future work we see a number of opportunities to explore these parameters in more detail for fine-tuned hypothesis generation. Additionally, a number of nominal qualitative variation measures exist that do not have direct analogs in quantitative statistics. An interesting avenue of further exploration may be to test how each of these nominal measures vary from or relate to the quantitative measures of *variance* and *standard deviation*, as a method to determine the degree of coherence.

8 Conclusion

Exploratory methods for iterative hypothesis generation are becoming essential with increasing data complexity and mixed data types across all domains. In this paper we introduced an Integrated Dual Analysis approach to hypothesis generation via the joint analysis of quantitative and qualitative data. This method extends the Dual Analysis framework by Turkay et al. [4], offering the opportunity for iterative, free form exploration of linked dimension and item spaces within complex and mixed data. To enable such a combined analysis of mixed data we introduced a set of descriptive measures, including a relatively new statistical measure, the *coefficient of unalikeability*, as well as measures of the modal distribution, *variance* and its qualitative analog, and *standard deviation* and its qualitative analog. Unlike the Dual Analysis approach in its original form, we do not perform data imputation, instead preserving the missing data items and explicitly presenting them for comparative analysis. This serves as a highly informative visualization for the uncertainty of the given data, and was exemplified by our clinical routine case data that contained over 70% of missing entries.

Our chosen visual encodings and interactions for each aspect of the Integrated Dual Analysis method are designed to support an iterative and exploratory workflow. We described this workflow in a generalized analysis of the hypothesis generation process, and provided a concrete example of this workflow in a paired clinical case study with a senior neurologist. Using this workflow to analyze a cerebral small vessel disease dataset, our clinical collaborator was able to iteratively explore dimensions and item subgroups to form new hypotheses and corroborate findings from medical literature. These hypotheses were formed from the simultaneous analysis of both qualitative and quantitative data dimensions and items, which would not have been possible to perform as efficiently or easily with existing statistical toolsets.

While our presented approach introduces new opportunities in the visual analysis of mixed data, we also see areas for further development. Looking to the future, expansion into statistical paired correlation analysis and dimensionality reduction techniques are natural areas for investigation to increase the power and efficiency of the Integrated Dual Analysis method. Furthermore, our approach can be applied beyond the medical domain to similarly heterogeneous datasets. Future work includes investigating our method within such domains as ecology, climate change and finance.

Acknowledgments

The research leading to this work was supported by the Federal State of Saxony-Anhalt, Germany (FKZ: I 88), the University of Bergen and the Trond Mohn Foundation in Bergen (#813558, Visualizing Data Science for Large Scale Hypothesis Management in Imaging Biomarker Discovery (VIDI)). Furthermore, we thank Frank Schreiber for his valuable contribution. Parts of this work have been done in the context of CEDAS, i.e., UiB's Center for Data Science.

References

[1] J. Kehrer, H. Piringer, W. Berger, and M. E. Gröller, "A model for structure-based comparison of many categories in small-multiple displays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2287–2296, 2013.

[2] C. Stolte, D. Tang, and P. Hanrahan, "Polaris: A system for query, analysis, and visualization of multidimensional relational databases," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 52–65, 2002.

[3] C. Turkay, J. Parulek, and H. Hauser, "Dual analysis of DNA microarrays," in *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '12*, 2012, p. 1.

[4] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing Dimensions - A Dual Visual Analysis Model for High-Dimensional Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2591–2599, 2011.

[5] S. L. Taylor, L. R. Ruhaak, K. Kelly, R. H. Weiss, and K. Kim, "Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices," *Briefings in Bioinformatics*, vol. 18, no. 2, pp. 312–320, 02 2016. [Online]. Available: https://doi.org/10.1093/bib/bbw010

[6] S. B. Green and N. J. Salkind, *Using SPSS for Windows and Macintosh, books a la carte*. Pearson, 2016.

[7] G. PSPP, "PSPP statistical analysis software: user guide," 2005.

[8] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser, "Hypothesis Generation by Interactive Visual Exploration of Heterogeneous Medical Data," in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Springer Berlin Heidelberg, 2013, vol. 7947, pp. 1–12.

[9] J. Kehrer and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 3, pp. 495–513, 2012.

[10] C. Chabot, C. Stolte, and P. Hanrahan, "Tableau software," *Tableau Software*, 2003.

[11] S. Oeltze, H. Doleisch, H. Hauser, P. Muigg, and B. Preim, "Interactive visual analysis of perfusion data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1392–1399, 2007.

[12] J. Kehrer, P. Filzmoser, and H. Hauser, "Brushing moments in interactive visual analysis," in *Computer Graphics Forum*, vol. 29, no. 3. Wiley Online Library, 2010, pp. 813–822.

[13] P. Angelelli, S. Oeltze, J. Haász, C. Turkay, E. Hodneland, A. Lundervold, A. J. Lundervold, B. Preim, and H. Hauser, "Interactive Visual Analysis of Heterogeneous Cohort-Study Data," *IEEE Computer Graphics and Applications*, vol. 34, no. 5, pp. 70–82, Sep. 2014.

[14] J. Krause, A. Dasgupta, J.-D. Fekete, and E. Bertini, "Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces," in *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, 2016, pp. 11–19.

[15] J. Wang and K. Mueller, "The visual causality analyst: An interactive interface for causal reasoning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 230–239, 2015.

[16] J. Wang and K. Mueller, "Visual causality analysis made practical," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2017, pp. 151–161.

[17] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller, "Visual correlation analysis of numerical and categorical data on the correlation map," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 2, pp. 289–303, 2014.

[18] T. O. Kvålseth, "Variation for categorical variables," *International Encyclopedia of Statistical Science*, pp. 1642–1645, 2011.

[19] L. C. Freeman, *Elementary applied statistics: for students in behavioral science*. John Wiley & Sons, 1965.

[20] H. F. Weisberg, *Central tendency and variability*. SAGE Publications, 1992, no. 83.

[21] T. O. Kvålseth, "Coefficients of variation for nominal and ordinal categorical data," *Perceptual and Motor Skills*, vol. 80, no. 3, pp. 843–847, 1995.

[22] G. D. Kader and M. Perry, "Variability for categorical variables," *Journal of Statistics Education*, vol. 15, no. 2, 2007.
[23] A. R. Wilcox, "Indices of qualitative variation." Oak Ridge National Lab., Tenn., Tech. Rep., 1967.
[24] E. H. Simpson, "Measurement of diversity," *Nature*, vol. 163, no. 4148, p. 688, 1949.
[25] J. Pearlman, P. Rheingans, and M. des Jardins, "Visualizing diversity and depth over a set of objects," *IEEE Computer Graphics and Applications*, vol. 27, no. 5, pp. 35–45, 2007.
[26] E. Mackin and S. Patterson, "Maximizing diversity of opinion in social networks," in *2019 American Control Conference (ACC)*, 2019, pp. 2728–2734.
[27] T. Pham, R. Hess, C. Ju, E. Zhang, and R. Metoyer, "Visualization of diversity in large multivariate data sets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1053–1062, 2010.
[28] M. C. Wee, "An improved diversity visualization system for multivariate data," *Journal of Visualization*, vol. 20, no. 1, pp. 163–179, Feb 2017.
[29] H. Kobayashi, H. Suzuki, and K. Misue, "A visualization technique to support searching and comparing features of multivariate datasets," in *2015 19th International Conference on Information Visualisation*. IEEE, 2015, pp. 310–315.
[30] A. Evren and E. Ustaoglu, "Measures of qualitative variation in the case of maximum entropy," *Entropy*, vol. 19, p. 204, 05 2017.
[31] M. G. Kendall and A. Stuart, *The advanced theory of statistics*. Hafner New York, 1958, vol. 1.
[32] S.-T. Chiu *et al.*, "Bandwidth selection for kernel density estimation," *The Annals of Statistics*, vol. 19, no. 4, pp. 1883–1905, 1991.
[33] M. Florek and H. Hauser, "Quantitative data visualization with interactive kde surfaces," in *Proceedings of the 26th Spring Conference on Computer Graphics*, 2010, pp. 33–42.
[34] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L 2 theory," *Gebiete*, vol. 57, pp. 453–476, 1981.
[35] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. CRC Press, 1986, vol. 26.
[36] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American statistical association*, vol. 79, no. 387, pp. 531–554, 1984.
[37] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, and H. Schumann, "A survey on interactive lenses in visualization." in *EuroVis (STARs)*, 2014.
[38] M. J. Saary, "Radar plots: a useful way for presenting multivariate health care data," *Journal of clinical epidemiology*, vol. 61, no. 4, pp. 311–317, 2008.
[39] K. Jafari, A. Tierens, A. Rajab, R. Musani, A. Schuh, and A. Porwit, "Visualization of cell composition and maturation in the bone marrow using 10-color flow cytometry and radar plots," *Cytometry Part B: Clinical Cytometry*, vol. 94, no. 2, pp. 219–229, 2018.
[40] J. Heinrich and D. Weiskopf, "State of the Art of Parallel Coordinates," in *Eurographics 2013 - State of the Art Reports*, M. Sbert and L. Szirmay-Kalos, Eds. The Eurographics Association, 2013.
[41] G. Kaur and B. B. Karki, "Bifocal parallel coordinates plot for multivariate data visualization," in *VISIGRAPP (3: IVAPP)*, 2018, pp. 176–183.
[42] G. G. Robertson, J. D. Mackinlay, and S. Card, "The perspective wall: Detail and context smoothly integrated," in *Proceedings of ACM CHI*, vol. 91, 1991, pp. 173–179.
[43] R. Tuor, F. Evéquoz, and D. Lalanne, "Parallel bubbles-evaluation of three techniques for representing mixed categorical and continuous data in parallel coordinates." in *VISIGRAPP (3: IVAPP)*, 2018, pp. 252–263.
[44] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 558–568, 2006.
[45] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger, "Looks good to me: Visualizations as sanity checks," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 830–839, 2018.
[46] S. Alemzadeh, U. Niemann, T. Ittermann, H. Völzke, D. Schneider, M. Spiliopoulou, K. Bühler, and B. Preim, "Visual Analysis of Missing Values in Longitudinal Cohort Study Data," *Computer Graphics Forum*, p. cgf.13662, 2019.
[47] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Tenth International Conference on Information Visualisation (IV'06)*, July 2006, pp. 9–16.
[48] Q. Li, Y. Yang, C. Reis, T. Tao, W. Li, X. Li, and J. H. Zhang, "Cerebral Small Vessel Disease," *Cell Transplantation*, vol. 27, no. 12, pp. 1711–1722, 2018.
[49] M. Pasi and C. Cordonnier, "Clinical relevance of cerebral small vessel diseases," *Stroke*, vol. 51, no. 1, pp. 47–53, 2020.
[50] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.
[51] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.
[52] M. Grinberg, *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.
[53] N. Elmqvist and J. S. Yi, "Patterns for visualization evaluation," *Information Visualization*, vol. 14, no. 3, pp. 250–269, 2015.
[54] S. Andrews, D. Das, K. J. Anstey, and S. Easteal, "Interactive effect of apoe genotype and blood pressure on cognitive decline: the path through life study," *Journal of Alzheimer's Disease*, vol. 44, no. 4, pp. 1087–1098, 2015.
[55] D. M. Michaelson, "Apoe ε4: The most prevalent yet understudied risk factor for alzheimer's disease," *Alzheimer's & Dementia*, vol. 10, no. 6, pp. 861 – 868, 2014.
[56] C. Pettigrew, A. Soldan, S. Li, Y. Lu, M.-C. Wang, O. A. Selnes, A. Moghekar, R. O'Brien, M. Albert, and R. T. the BIOCARD, "Relationship of cognitive reserve and apoe status to the emergence of clinical symptoms in preclinical alzheimer's disease," *Cognitive neuroscience*, vol. 4, no. 3-4, pp. 136–142, 2013.
[57] N.-Y. Jung, H. Cho, Y. J. Kim, H. J. Kim, J. M. Lee, S. Park, S. T. Kim, E.-J. Kim, J. S. Kim, S. H. Moon *et al.*, "The impact of education on cortical thickness in amyloid-negative subcortical vascular dementia: cognitive reserve hypothesis," *Alzheimer's research & therapy*, vol. 10, no. 1, p. 103, 2018.
[58] M. Milà-Alomà, G. Salvadó, J. D. Gispert, N. Vilor-Tejedor, O. Grau-Rivera, A. Sala-Vila, G. Sánchez-Benavides, E. M. Arenaza-Urquijo, M. Crous-Bou, J. M. González-de Echávarri, C. Minguillon, K. Fauria, M. Simon, G. Kollmorgen, H. Zetterberg, K. Blennow, M. Suárez-Calvet, J. L. Molinuevo, and for the ALFA study, "Amyloid beta, tau, synaptic, neurodegeneration, and glial biomarkers in the preclinical stage of the alzheimer's continuum," *Alzheimer's & Dementia*, vol. 16, no. 10, pp. 1358–1371, 2020.
[59] D. Smeijer, M. K. Ikram, and S. Hilal, "Enlarged perivascular spaces and dementia: A systematic review," *Journal of Alzheimer's Disease*, no. Preprint, pp. 1–10, 2019.
[60] B. Gyanwali, H. Vrooman, N. Venketasubramanian, T. Y. Wong, C.-Y. Cheng, C. Chen, and S. Hilal, "Cerebral small vessel disease and enlarged perivascular spaces-data from memory clinic and population-based settings," *Frontiers in Neurology*, vol. 10, p. 669, 2019.
[61] S. Mittal, Z. Wu, J. Neelavalli, and E. M. Haacke, "Susceptibility-weighted imaging: technical aspects and clinical applications, part 2," *American Journal of neuroradiology*, vol. 30, no. 2, pp. 232–252, 2009.

Juliane Müller joined the *Medicine and Digitalization* (MedDigit) group at the Department of Neurology, Univ. of Magdeburg, Germany as a doctoral researcher in 2018. She received her M.Sc. in Computer Science in 2016 from TU Braunschweig. Her research interests include information visualization, the visual analysis of medical data, and visual explainability of model-based clinical decision support.

Laura Garrison joined the Visualization Research Group in the Department of Informatics at the Univ. of Bergen, Norway as a doctoral researcher in 2018. She received her M.Sc. in biomedical visualization in 2012 from the Univ. of Illinois. Her research focuses medical visualization and visual analytics, drawing from her background as a medical illustrator.

Philipp Ulbrich is a medical student at the Otto von Guericke University Magdeburg, Germany. He joined the research group of Prof. Stefanie Schreiber at the Department of Neurology in 2019 as a doctoral researcher. His research interests include cerebral small vessel disease and central nervous system extracellular matrix.

Stefanie Schreiber is a professor at Department of Neurology, Univ. of Magdeburg, Germany since 2018. In 2014, she received a habilitation (venia legendi) in Neurology and in 2007 a Ph.D. in Medicine from Univ. of Magdeburg. Her research interests include Cerebral Small Vessel Disease (CSVD) and Neuromuscular Disorders.

Stefan Bruckner is a full professor in Visualization at the Department of Informatics of the Univ. of Bergen, Norway since 2013. He received his master's degree (2004) and Ph.D. (2008), both in Computer Science, from the TU Wien, Austria, and was awarded the habilitation (venia docendi) in Practical Computer Science in 2012. His research interests include all aspects of data visualization, with a particular focus on interactive techniques for the exploration and analysis of spatial data.

Helwig Hauser received his MSc (1995), Ph.D. (1998), and habilitation (2004) in Computer Science from TU Wien, Austria. Since 2007, he is a professor in Visualization at the Department of Informatics of the Univ. of Bergen, Norway. His interests include interactive visual analysis, illustrative visualization, and the combination of scientific and information visualization, with particular focus in the application of visualization to the fields of medicine, geoscience, climatology, biology, engineering, and others.

Steffen Oeltze-Jafra heads the working group *Medicine and Digitalization* at the Department of Neurology, Univ. of Magdeburg, Germany. In 2016, he received a habilitation in Computer Science, in 2010 a Ph.D. in Computer Science, and in 2004, a diploma in Computational Visualistics from Univ. of Magdeburg. His research interests are in the quantitative analysis of clinical routine data, the visual analysis of biomedical data and in model-based clinical decision support.