Insight Into Lumbar Back Pain
What the Lumbar Spine Tells About Your Life

Paul Klemm[1], Sylvia Glaer[1], Kai Lawonn[1] Henry Vlzke[2] and Bernhard Preim[1]
[1]*Department of Simulation and Graphics, University of Magdeburg*
[2]*Greifswald*
{*paul, sylvia*}*@isg.cs.uni-magdeburg.de, voelzke@anderAddy*

Keywords: My Keywords in Title Case

Abstract: The abstract should summarize the contents of the paper and should contain at least 70 and at most 200 words. The text must be set to 9-point font size. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1 INTRODUCTION

Epidemiology is the study of causation of diseases. Large population studies, such as the Study of Health in Pomerania (SHIP) **??** gather as much information as possible about participants to be assessed towards different diseases. These information are used to determine risk factors for diseases, helping people to make their lifestyle healthier or helping in diagnosing a disease. Epidemiological research is strongly hypothesis driven. Observations made by clinicians are translated into hypothesis, which are then statistically evaluated using data variables from epidemiological studies.

Modern cohort studies often comprise medical image data. Short summary

Back pain is one of the most frequent diseases in the western civilization.

Our goal is to combine data mining algorithm with data visualization to provide insight into the quality of image derived data to analyze if it acts as a risk factor for a disease.

More than 5 variables are rarely analyzed simultaneously. Our contributions are:

- Analyzing back pain using image-derived variables of 2,240 subjects.
- Assessing the suitability of lumbar spine shape for diagnosing back pain
- Analyzing correlations between between image-based and socio-demographic as well as medical parameters.
- Identification of most important variables using data mining methods noch schoen schreiben
- semiquantitative Auswertmglichkeiten werden dem User oft zur verfgung gestellt
- Techniken sind als Teil des IVA-Frameworkes zu verstehen

2 EPIDEMIOLOGICAL BACKGROUND

Epidemiology is the study of dissemination, causes and results of health-related states and events. Epidemiological reasoning relies on a strict statistically driven workflow (**?**):

- Physicians formulate hypotheses based on observations made in their clinical practice.
- To assess a hypothesis, epidemiologists compile a list of variables depicting it.
- Statistical methods, such as regression analysis, assess the association of selected variables with the investigated disease.

Mutually dependent variables make this analysis challenging. Many diseases, such as different cancer types, are more likely with increasing age. When for example analyzing influences of nutrition to prostate cancer, the results need to be age-normalized. Age

acts as an *confounder* for prostate cancer. These *confounding* variables are often hard to find. Statistical correlation does not imply causation–epidemiologists need to assess the medical soundness of the statistical results.

2.1 Epidemiological Data

Epidemiological data can stem from a wide range of studies. The study type depends on the condition of interest. Most common are case-control studies, analyzing one specific disease and its influences. We focus on data from large scale cohort studies. These studies aim to collect as much data as possible for each subject. As a result, these data can be analyzed regarding many diseases and conditions.

Epidemiological data is heterogenous and incomplete. Women-specific question can only be acquired for female subjects, data about a disease treatment only for subjects suffering from this condition. Therefore, statistical analysis has to take missing data into account.

Epidemiologists acquire data using a wide range of techniques, such as medical examinations, self reported questionnaires or genetic examinations. This yields a heterogenous information space. To compare these data, information reduction techniques are applied. For example, continuous data, such as age is often discretized into age-bins. Every information reduction can introduce a bias to the data, since it reflects an assumption about the data. Using age to divide subjects into *young* and *old* categories can distort statistical results. This distortion is reduced with increasing number of discretization steps.

Modern cohort studies often comprise medical image data. These data are hard to analyze, since segmentation algorithms are not generally available and need to be custom-made for each body structure. Segmentation data is usually analyzed by abstracting it into key figures, such as diameters or distances. These numeric values can be compared with non-image variables to retrieve correlations.

2.2 Lower Back Pain

The lower (*lumbar*) spine is the most stressed part of the spine. Lower back pain is one of the most frequent diseases in the western civilization. Epidemiologists assume associations between lumbar back pain and lifestyle factors. These include nutrition, sporting activities and body posture at work. The exact causes as well as particularly vulnerable risk groups are not known. Potential *confounding* effects are also subject of current research.

Epidemiologists want to characterize the healthy aging process of the spine. To achieve this, they have to analyze the lumbar spine shape as well as the mentioned lifestyle factors.

3 RELATED WORK

Our own work (VIS, VMV, BVM). Sylvias paper with reference to the methods. **More information necessary here!**

4 The Lumbar Spine Data Set

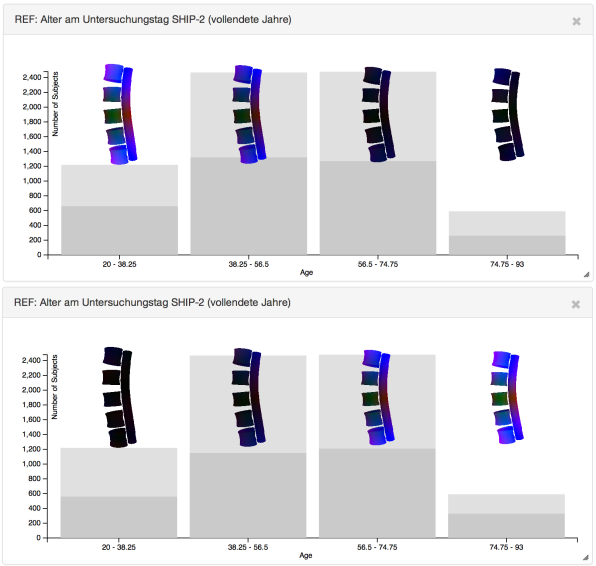

Figure 1: Age-Gender

Our approach allows to analyze many variables simultaneously. Therefore, epidemiologists compiled the data set with a with a wide range of variables possibly correlating with lumbar back pain. The data set comprises of 6,753 subjects from two cohorts (4,420 from `SHIP-Trend-0` and 2,333 from `SHIP-2`).

4.1 Non-Image Data

The variables range from somatometric variables describing body measures to medical examinations, such as laboratory tests as well as lifestyle factors as sport activity or nutrition.

4.2 Image Data

From VIS'14 Paper The lumbar spine was detected in the image data using a hierarchical finite element

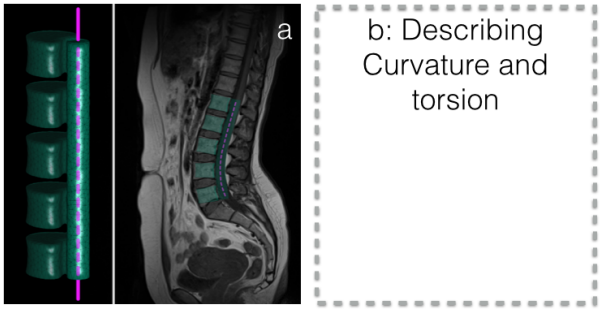

Figure 2: Centerline

method by Rak et al. [32]. This semiautomatic method requires the user to initialize the tetrahedron-based finite element models (FEM) 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 T1and 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 [21]. Due to incorrect initialization, strongly deformed spines, contrast differences and artifacts, the model was not able to detect lumbar spines for all subjects. We obtained and worked with 2,540 tetrahedron models of the lumbar spine. For clustering, we extracted the centerline of the lumbar spine canal, which captures information about lordosis and scoliosis (the medical terms for spine curvature) [21].

We have to assess the model accuracy to extract key figures from it. The detection model depicts the vertebrae positions, but lacks detailed information about their volume. It captures reliable information about spine canal curvature. In a previous work (**?**; **?**), we extracted a centerline representation of the lumbar spine canal from the detection model (Fig. 2 (a)). Using the Frenet formulas (**?**), we extracted the following metrics from the model (Fig. 2 (b)):

- *Curvature* is calculated as weighted sum of curvature between all adjacent points describing the centerline: $\sum \frac{curvature_i}{curvature_{all}}$.
- *Torsion* (deviation of a curve from the course) is calculated as weighted sum of torsion between all adjacent points describing the centerline: $\sum \frac{torsion_i}{torsion_{all}}$.
- *Curvature angle* is defined by the middle point of the spine canal centerline as *vertex* and the line between middle point and top/bottom point as *sides*.

These figures are also extracted in the sagittal, coronal and transversal projection of the model. We assess the information gain of each dimension using these projections. This gives us a total of 9 image-derived parameters.

5 EXPERIMENTS

Spine shape is confounded by several somatometric variables. Larger people have also a longer spine and its shape is more straight compared to smaller subjects. Since men are on average taller than women, gender is another confounder. Large body weight increases the spine load, resulting in a bend shape. All these variables need to be taken into account, when correlating spine curvature and torsion with non-image parameter. Since the gender confounder mainly encodes body-height, we decided to divide subjects into body-size groups. To avoid small outlier groups, epidemiologists recommend using quartiles to discretize metric variables (**?**).

5.1 Preliminary Results

As first experiment we correlated the shape parameter with the binary back pain indicator using a pairs plot (Schloerke et al., 2011). Pairs plots visualize continuous and categorical variables pairwise in a matrix view. Scatterplots display pairs of continuous variables, bar charts for continuous variable as a function of a categorical variable and box plots pairs of categorical variables. Since our image-derived variables are metric, their pairwise combinations are visualized using scatterplots. The combination of the image variables with back pain is visualized as histogram at the left side of the matrix and as box plot on the right side. Figure 3 (right) shows the range of each variable as box plot. The projections to the transversal planes attract attention as they have many outliers. We conclude that curvature is not as reliable on the transversal plane as it is on the other planes.

5.1.1 Correlation Matrix

We calculated an association matrix to to assess correlations between the image parameters. The pearson correlation coefficients between the numeric variables are depicted right of the main diagonal of Figure 3. Curvature, curvature angle and torsion correlate strongly with their planar projections. Also the mean curvature and the curvature angle correlate by a factor of -0.89. Torsion does not correlate with any other image-derived variable.

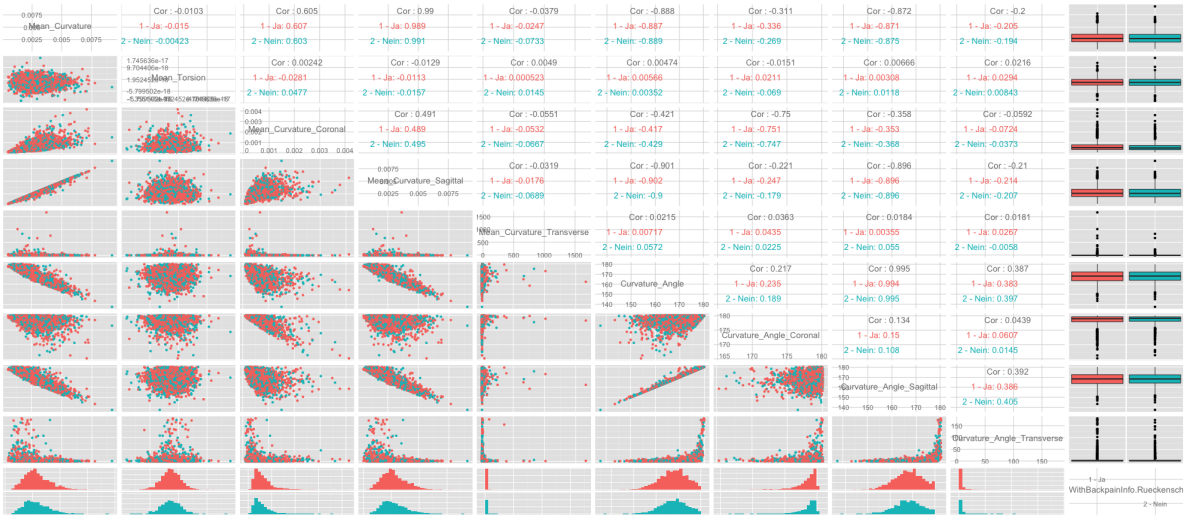

Figure 3: image-parameter-range

5.1.2 Correlation of Image Parameters With Back Pain

Figure 3 shows the distribution for all subjects. We created a pairs plot for subject groups derived by dividing them into quantile groups of body-size. No statistical significant correlation could be observed through all subject groups. The box plots show no difference between subjects with and without back pain.

5.1.3 Assessing the Information Gain Using the PCA

To determine the information gain per image-derived variables, we calculated a PCA and compared the loadings per dimension. The first three principal components explain 75% of variance in the image-derived variables. The first principal component explains 47% of the variance and weights primarily *mean curvature*, *curvature sagittal*, *curvature angle* and *curvature angle sagittal*. The second component, adding 16% of variance, weigths *curvature coronal* and *curvature angle coronal*. The third component explains 12% of variance and weights *torsion* and *curvature transverse*. This supports our prior conclusion about the low information gain of the transverse planes. *Torsion* also adds little variance to the information space.

5.1.4 Heterogenous Correlations

We then expanded our focus on correlations of image-derived parameters with all other non-image variables. We applied a heterogenous correlation technique to derive correlations between all variables is the data set. The method uses the following correlation metrics for the different type combination:

- *pearson product-moment* for two continuous variables,
- *polyserial* correlation for one continuous and one categorical variable, and
- *polychoric* correlations for two categorical variables.

All correlation values are scaled between 0 - no correlation and 1 - identical. Some variables are too sparse for calculating correlations, for example TODO: some variables from calculation here; TODO: Image?, eventuell interactiven Tile-Plot. We display the resulting *contingency matrix* using a heat map, mapping correlation values to color brightness with white for 0 and dark blue for 1 (**?**). We calculated the contingency matrix for all size groups and looked for correlations between image- and non-image derived variables. The resulting contingency matrix shows no strong correlation with any of the parameter. 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 with *torsion*, which correlated with almost no variables (p values between 0 and 0.05) and *parkinson* (0.24). This observation brought us to the conclusion to incorporate more sophisticated data mining techniques to assess the influence of the image-derived parameters.

5.2 Experimental Settings and Results

Decision Tree erklaeren, ebenfalls Umrechnen in die Dummy-Variablen. Was sind die Zielvariablen

- Altersgruppierung
- Male/Female
- Gewicht (BMI - dnn, normal, dick)

6 EVALUATION OF DECISION TREES

As described before, correlation coefficients fail to infer back pain status based on lumbar spine canal curvature and torsion. We rely on predictive classification trained to obtain a complex rule set on how combination of the image-parameter explain non-image variables. Decision trees are a popular classification method in data mining for creating predictive models. Leaves of a decision tree represent class labels, branches represent feature conjunctions leading to the class labels. Decision trees are easy to understand and to read. They work with numerical as well as categorical data. This allows epidemiological domain experts to interpret the results without having deep knowledge about the algorithm creating the tree. Readability is only granted for small trees, complex structures with many branches are not desirable. Too many branches also impose to the data. Quelle Decision Tree

6.1 The C4.5/C5.0 Algorithm

Using a trained data set, the C4.5 algorithm builds decision trees based on information entropy. Such a calculation requires a numeric or categorical target variable. The algorithm then tries to find a decision tree, which divides the samples using the input variables just like the target variables. This means, that every node in the tree represents the attribute which splits the data most efficient into the target subset. The pseudocode for the algorithm is defined in Algorithm 1 (Kotsiantis et al., 2007). C5.0 is developed to produce smaller decision trees than C4.5 and improve the execution time. We use the `R` implementation of C5.0 (**?**). Categorical attributes with more levels are biased with more information gain in a decision tree (Deng et al., 2011). Creating dummy variables bypasses this problem.

The actual use for the resulting tree is the classification of new observations (subjects). Yet, we are interested in the decision rules and the classification accuracy.

Check for base cases;
for *each attribute a* **do**
  Find the normalized information gain ratio from splitting on a;
end
Let a_{best} be the attribute with the highest normalized information gain;
Create a decision node that splits on a_{best};
Recurse on the sublists obtained by splitting on a_{best}, and add those nodes as children of node;

Algorithm 1: Building a decision tree using the C4.5 Algorithm

6.2 Interactive Display of Decision Trees

We have to create a decision tree for every non-image variable to see which one can be explained by image-derived variables. Since we have 134 non-image variables, the calculation yields a corresponding number of trees. Dividing the subjects into four groups using body-size, the number increases to 402 trees. To make the results cognitive feasible, we have to abstract the results of the classification.

6.2.1 Visualization of Classification Results

We follow the visual information-seeking mantra and want to provide an overview first, then details-on-demand **??**. The optimal classification uses a few rules to precisely characterize the target variable. Therefore, we are interested in *small trees* with a *low classification error rate*. The two measures form the axis for a *scatter plot* of the classification results.

The Error Term. Normally the error rate for a classification is calculated with $error = \frac{n_{classifiedCorrectly}}{n}$ where n is the number of subjects. The metric usually works well for variables with uniform distribution. It distorts the result for other distribution types. If for example a variable indicating a disease is negative for 90% of the subject and the classifier simply assigns all subjects to *not ill* the error metric would yield an error of 10%, even though it is very bad. Our error term therefore incorporates the discriminative power of each manifestation and is denoted as follows:

$$error = 1 - \frac{\sum_{m=0}^{M} \frac{m_{classifiedCorrectly}}{m_{all}}}{number_{dimensions}} \quad (1)$$

M represents the set of manifestation of each variable. The error is scaled to denote perfect classification with 0 and 1 is equal to random selection. It allows for comparability of error rates between variables with different manifestation count.

Attribute Mapping The scatter plot axis are defined by tree size and the previously described error metric. This allows us to visualize a multitude of classification results in one plot. Lets say we want to classify and compare the same variables for different subject groups, for example male and female subjects. We can support this by color coding the data points according to group affiliation.

Many follow-up 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. This makes these results less reliable. We map the number of subjects associated with a variable to point diameter in the scatter plot. This allows to instantly assess the reliability of the result.

We apply a square root scale for the tree size axis to highlight data points with few decision rules. Outlier results with very large decision trees distort otherwise the resulting plot.

Needs to be moved We removed variables with less than 100 subjects, since they have to few subjects to be statistically significant. ToDo: Manche Variablen rausgeschmissen, weil sie zu wenig Probanden enthielten, oder Auspraegungen nur duenn besetzt waren!

6.2.2 Dummy Variables

Dummy variables abstract a categorical variable with multiple manifestations into several binary variables. Each binary variables 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 binary variables (*No pain - Yes/No* to *Large Pain - Yes/no*). One subject can only have one of these variables set to true. This is useful for our classification, because it allows to determine which manifestation of a variable can be described best using the image data parameter.

6.2.3 Interaction With the Visualization

The described visualization provides a good overview over the classification results. We however still want to be able to look display *details-on-demand* **??** and examine a decision tree in detail. This is realized by clicking on a entry on the visualization, which then displays the corresponding decision tree in detail (Fig. **??** b). This allows for sequentially analyzing the classifications.

6.2.4 Implementation

All analysis are carried out using `R`, a popular 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 parameters using UI controls, such as sliders. We implemented our application as web based prototype using `R Shiny`[2]. The package combines the power of `R` with advantages of web-based applications. It allows us to quickly exchange results with our collaborating epidemiological partners. They can try out the technique without installing anything. Exchanging software becomes as easy as exchanging a hyperlink.

6.3 Results

We ran the analysis using different subject groups:

1. All subjects,
2. subdivision into *male* and *female*,
3. subdivision into *age quantiles*, yielding the groups (21, 43.8] (43.8, 54] (54, 64] (64, 90],
4. subdivision into *size quantiles*, yielding the groups (139, 164], (164, 171], (171, 177], (177, 202].

We plotted each group twice. The first plot shows all original variables. The second plot shows all categorical variables transformed into dichotomous dummy variables.

6.3.1 All Variables

The vast majority of parameters can not be described well using the classifier. This is reflected in the large amount of variables classified with a error rate above $0,6$.

None of the pain indicators can be described reliably using the image-based parameter. The only variable reliably classified in this group is gender.It can be discriminated with an error rate of $0,31$ using 7 rules and incorporates only curvature and curvature angle variables. Surprisingly high is the classification of medication for high blood pressure. 1,058 subjects are classified with an error rate of $0,47$ solely based on coronal mean curvature. Almost all subjects who are medicated $(796/1,058)$ were correctly classified, the vast majority of non-medicated subjects $(262/1,058)$ are false-positive classified. Therefore the classifier is not as useful as the error-rate would

[1] GGVIS Footnote
[2] R Shiny footnote

indicate. The four body size groups could be characterized with an error rate $0,48$, but the decision tree comprises of 71 rules and imposes overfitting.

The analysis of the dummy variable yields similar results like the blood pressure medication. Variables, such subjects sized 139-164 cm, between 64 and 90 years of age or nutrition related parameters are dominantly populated by one manifestation. The classifier neglects the other group and yields a error rate below $0,5$.

6.3.2 Gender Groups

Classification using groups divided by gender do not produce satisfying results. Only hypothyroidism could be described for male subjects with an error rate of $0,24$ for 110 subjects using the *mean curvature* and *curvature angle*. Since there are only 30 male subjects diagnosed with hypothyroidism, reducing the statistical power of the result. The dummy variable analysis showed that female subjects of 139-164 cm body height could be discriminated using the mean curvature and curvature angle, with an error of $0,38$.

6.3.3 Age Groups

Gender could be described for each age group using *mean curvature* and *curvature angle*. The accuracy varies between $0,35$ ($43,8-54$ years of age, 3 decision rules) to $0,27$ ($21-43,8$ years of age, 6 decision rules). Many variables, such as body size can be described with an error rate of $0,35$ to $0,45$ but only using large decision trees with over 20 rules. Notable is also the increasing accuracy with increasing decision tree size. For subjects between $43,8-54$ years of age, the classifier discriminated subjects with thyroid nodules with an error rate of $0,32$.

The dummy variable analysis shows many results using a decision tree with one rule based on *mean curvature* with accuracy between $0,4$ and $0,5$. It shows that mean curvature can be used to predict dependencies to variables, such as *high blood pressure*, *hypothyroidism* or *nutrition*.

6.3.4 Size Groups

Many previously described results can be associated to be confounded with subject size. Differences in the gender analysis are mostly due to the average height difference between males and females. For example, large subjects are already characterized by their rather straight spine. The question is, whether the intergroup spine-variability parameter is enough for predicting other parameter or not. Dividing subjects into height groups potentially highlights classifications not confounded by body-height.

Large Decision Trees. Back-pain associated variables can be explained for various size-groups, but we could not extract universal rules. Radiating back pain could be described with error rate of $0,39$ using 23 rules for subjects between $171-177cm$ body size using torsion and mean-curvature. For subjects sized $177-202cm$ the accuracy drops to $0,47$ using 20 decision rules. There are several decision trees for laboratory values, such as alanine aminotransferase value in the blood can be described with a high precision of $0,4$ ($139-164cm$) to $0,36$ ($164-171cm$). 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 described using only one decision rule with accuracy between $0,42$ to $0,47$. Most of these variables have a dominant manifestation and the classifier shows a low detection precision for the second manifestation. These variables include nutrition parameter, thyroid disorder and social problems induced by back pain.

6.3.5 Concluding Remarks

Except for gender in the age-plot, no variable could be explained with reasonable accuracy for age or body size.

Manchmal kann es passieren, dass die Quantile fuer eine Variable gleich sind - z.B. bei Dauer Jahre Schwerarbeit, weil viele Probanden hier eben eine 0 haben. Deswegen kann das Ding nicht in vier Qunatile unterteilt werden, die ersten 3 sind 0. Deswegen wurden nur die einzigartigen Unterscheidungen genommen (statt *quantile(myVar)* wurde *unique(quantile(myVar))* verwendet.)

6.4 Themen hier unterbringen

We can discriminate back pain using non-image data, but not with image-derived parameter. Welche parameter koennen wir gut unterscheiden

Allgmeine Analyse der Daten mit einbringen. Hierfuer statistisch Auswertung machen, welche Parameter am meisten verwendet wurden bei der Kategorisierung. Nur Parameter verwenden, die genauer sind als 0.5%, der Entscheidungsbaum kleiner als 15 ist und das Feature mindestens 100 Probanden enthlt

Weiterhin interessant die Beschreibung rueckenschmerzassoziitierter Parameter, bei einigen zeigt sich, dass Torsion sehr dominant vertreten ist.

Algorithmus schreiben, der eng zusammenliegende Parameter fuer alle Probanden sucht.

7 CONCLUSION

ACKNOWLEDGEMENTS

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grant no. 03ZIK012), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania. Whole-body MR imaging was supported by a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg-Vorpommern. The University of Greifswald is a member of the Centre of Knowledge Interchange program of the Siemens AG. This work was supported by the DFG Priority Program 1335: Scalable Visual Analytics.

REFERENCES

Deng, H., Runger, G., and Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 293–300. Springer.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., and Marbach, M. (2011). Ggally: Extension to ggplot2.

APPENDIX

If any, the appendix should appear directly after the references without numbering, and not on a new page. To do so please use the following command: *\section*{APPENDIX}*