Optimal Classification of Long Echo Time In Vivo Magnetic Resonance Spectra in the Detection of Recurrent Brain Tumors

B. H. Menze[1]; M. P. Lichy[2,3]; P. Bachert[4]; B. M. Kelm[1]; H.-P. Schlemmer[2,3]; F. A. Hamprecht[1],

1 Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany
2 German Cancer Research Center (dkfz), Division of Radiology, Heidelberg, Germany
3 currently: University of Tuebingen, Department of Diagnostic Radiology, Tuebingen, Germany
4 German Cancer Research Center (dkfz), Division of Medical Physics in Radiology, Heidelberg, Germany

Abstract

We present an optimal high-level postprocessing of single-voxel ^{1}H magnetic resonance spectra, and assess the benefits and limitations of automated methods as diagnostic aids in the detection of recurrent brain tumor.

In a previous clinical study, 90 long-echo-time single voxel spectra were obtained from 52 patients and classified during follow-up (30/28/32 normal/non-progressive tumor/tumor). On this data a large number of evaluation strategies, including both standard resonance line quantification and algorithms from pattern recognition and machine learning were compared in a quantitative evaluation. Results from linear and nonlinear feature extraction, including ICA, PCA, and wavelet transformations, but also the data from resonance line quantification were combined systematically with different classifiers such as LDA, chemometric methods (PLS, PCR), support vector machines, and ensemble methods. Classification accuracy was assessed using a leave-one-out cross-validation scheme and the area under the curve (AUC) of the receiver-operator-characteristic (ROC).

A regularized linear regression on spectra with binned channels reached 91% classification accuracy compared to 83% from quantification. Interpreting the loadings of these regressions, we find that lipid and lactate signals are too unreliable to be used in a simple machine rule. Choline and NAA are the main source of relevant information.

Overall, we find that fully automated pattern recognition algorithms perform as well as, or slightly better than, a manually controlled and optimized resonance line quantification.

Keywords: Statistical Learning, Chemometrics, Preprocessing, Postprocessing, Benchmark, Magnetic Resonance Spectroscopy, Human Brain Tumor

1 Introduction

In vivo nuclear magnetic resonance spectroscopy (MRS) opens a window into certain metabolic processes in living tissue. However, the biochemical function of the metabolites attributed to the observed resonances is presently not fully understood. So the detailed analysis of empirical data and the generation of optimal rules is crucial for the success of MRS in medical diagnostics.

Two different approaches are discussed for the extraction of spectral information: inverse modeling of the resonance lines can be used to infer the absolute or relative concentrations of the biochemical agents ("quantification") [21, 35, 14], or the full spectrum is input to statistical decision methods, an approach often termed "pattern recognition" in MRS literature [10, 8].

An early review on MRS data analysis stated in 1997 that "benchmarking[1] is probably the single most important step towards implementing practical and clinical information [of MRS analysis into computer algorithms]" [8]. In the meantime, MRS-based diagnostic methods have been developed for a number of clinical applications (e.g., in the diagnosis of brain tumors and lesions [13, 22]). These methods are almost exclusively based on the quantification of resonance lines, whereas pattern recognition methods [8, 10] do not seem to have found routine use since the time of that review [31, 17, 32, 12].

The objective of this study is to provide such a benchmark in terms of a systematic and quantitative comparison of both data analysis approaches for the detection of recurrent tumor after radiotherapy.

In this diagnostic question, both radiation-induced tissue changes in a state of stable disease and recurrent tumor may be observed at the same location and both show contrast enhancement in standard CT and MRI [4, 25]. MRS and,

[1] Benchmarking is the process of assessing the performance of competing products, services or practices against one another with the aim of setting standards and guidance choice. [8, p.117]

more recently, MR spectroscopic imaging (SI) have demonstrated their value in this ambiguous situation [25, 16].

While the pre-therapeutic diagnostics of brain tumor have been studied extensively [9, 5], less attention has been devoted so far to the optimization of decision rules for the analysis of MRS data in the post-therapeutic diagnostics. New strategies of multimodal MR diagnostic [28] depend crucially on a reliable and robust machine guided analysis of the spectral information, so the present work also aims towards the development of similar techniques in the postoperative diagnostic setting.

Pattern recognition is a long-term standard in the classification of spectra of all kinds and its application to in vivo MR spectra has been discussed for a while [1, 17, 32]. Conceptually it consists of a preprocessing step, including feature extraction and feature selection, mainly for the purpose of dimensionality reduction (discussed in [8, 11]), and a subsequent classification step (see fig. 1). Although the importance of preprocessing is widely recognized, only few studies focus on that aspect [15, 5, 27].

While previous studies have used statistical decision methods such as linear discriminant analysis [32], logistic regression [4], decision trees [18] and nonlinear support vector machines [5], primarily in conjunction with quantification, a comprehensive and quantitative benchmark of optimal combinations of preprocessing strategies *and* classification methods is not available to the best of our knowledge. We will discuss appropriate methods in order to find the best such combination on the given data set.

In the following, we will shortly discuss differences between pattern recognition methods and a conventional quantification (section 2), compare the two approaches quantitatively on our data set (section 3) and present and discuss our results (sections 4 and 5).

2 Concepts of Spectral Analysis

2.1 Quantification

A spectral pattern can be understood and modeled as superposition of individual resonance lines. This description is very general and useful both in an explanatory, purely descriptive setting, and in a supervised learning context (calibration, classification). Especially for the latter, methods of inverse modeling are needed to infer the parameters of the single resonance lines [35, 21, 24].

These methods often rely on prior knowledge about the individual resonance lines, such as the expected position, width, or shape. Basis functions are either determined empirically [14] (e.g., measured or "learned" for single metabolites, as in "LCmodel" [24]), or are parametric models deduced from the physics of MR (e.g., assuming Lorentzian, Gaussian, Voigt line shape functions, as realized in "AMARES" [36]).

From a data analyst's point of view, an important advantage of this approach is the low dimensionality of the resulting data representation – corresponding to the number of metabolites or resonance lines rather than the number of spectral channels – which allows the use of simple decision rules and a straightforward interpretation of the extracted features. This biochemical interpretability and the possibility to check the results for plausibility halfway between the spectral pattern and the diagnosis explains the high confidence in this approach.

Unfortunately, in the presence of noise or artifacts, the quality of such a quantification suffers from its high flexibility and the fact that the optimization in the inverse modeling may converge to a wrong solution [1]. As a consequence, a fully unsupervised fitting of resonance lines may result in misleading output in practice. In the presence of noise, results can only be trusted after visual inspection and reassurance that the fitting procedure has worked successfully [37, 30].

2.2 Pattern recognition

From a methodological point of view [6], any resonance line quantification followed by a decision rule operating on the resulting parameters (fig. 1) can be seen as “pattern recognition”.

However, literature on MRS data analysis generally subsumes only nonparametric approaches under this term [10], a convention that we shall adopt for the remainder of the paper. A large variety of algorithms originally developed in the context of signal processing, chemometric or machine learning (e.g., [5, 28]) fall into this class. Their common feature is that they assess the physiological state of the tissue directly from the spectral pattern and that they bypass the intermediate step of resonance line quantification.

Availability of a representative training sample comprising the states of disease of diagnostic interest (e.g. tumor discriminated vs. normal tissue, neoplastic lesion vs. non-neoplastic lesion) is a prerequisite for pattern recognition. Also, as the statistical evaluation now has to deal with hundreds of channels instead of a small number of resonance line parameters (see [8, 11] for a discussion), dimensionality reduction prior to classification becomes a major concern (fig. 1). This conceptual separation between primary feature extraction and subsequent classification is not possible for all pattern recognition algorithms [14], which hampers insight into the decision process of these methods and evokes the main criticism concerning this approach.

Therefore, pattern recognition often appears as a “black-box” method with no biochemical interpretability. In section 4, we will illustrate how the diagnostic rule at least of a linear classifiers can be compared against biochemical knowledge.

In contrast to resonance quantification, pattern recognition methods do not require an optimization step in their application. As shown in section 4, this might result in procedures that are less susceptible to noise and artifacts. Pattern recognition methods therefore have a high potential to operate reliably and fully automatically even on low quality spectra [32].

3 Methods

3.1 Data

3.1.1 Patients/Study design

Our data set comprised single-voxel ^{1}H MR spectra of 58 lesions from 52 patients after initial treatment by radiotherapy, examined at the German Cancer Research Center (dkfz), Heidelberg [25]. The lesions were classified during follow-up on the basis of magnetic resonance imaging (MRI) or positron emission tomography (PET) examinations. Recurrent tumor was diagnosed in the case of an increase of the hyperintense area on T_2-weighted MR images by more than 25% in size, or appearance of a new contrast-enhancing area with subsequent enlargement on MRI and (in patients with PET examinations) significant tracer (^{18}Fluoro-2-deoxy-D-glucose) uptake. MR follow-up examinations started 6 weeks after radiotherapy and were repeated every 3 to 6 months. The mean length of follow-up for patients with stable disease was 15.3 months.

For 30 lesions, a recurrent tumor was confirmed, whereas 28 were diagnosed with non-progressive tumor, comprising both radiation injury and stable disease (fig. 2). For the reference group of normal physiological state (i.e. normal tissue), 32 spectra were acquired from the contralateral unaffected regions of these patients. Primary tumors were in the majority astrocytoma of grade 1-4, but the study also included patients with meningioma and metastases (also see [25] for details on the study).

– – – Figure 1 about here – – –

– – – Figure 2 about here – – –

3.1.2 Data acquisition

MRI and single-voxel ^{1}H MRS examinations were performed using a 1.5-T whole-body scanner (Magnetom Vision; Siemens, Erlangen, Germany) with commercially available pulse sequences and the standard head coil. Voxel sizes varied from $1.5 \times 1.5 \times 1.5$ cm^3 to $2 \times 2 \times 3$ cm^3, with a majority of the voxels sized $2 \times 2 \times 2$ cm^3. All lesions were larger than the selected voxels. Normal brain tissue, cerebrospinal fluid and edema were avoided, but (central) areas of the lesions could comprise necrotic tissue.

The MR spectra were obtained with a double spin-echo sequence with one-pulse water-signal suppression and long echo time (1500/135/200-300 [T$_R$ / T$_E$ / number of excitations], spectral width 1 kHz, 1024 data points).

3.2 Statistical analysis

The search for the best decision rule was performed on the basis of several different representations of the spectral information (see fig. 1).

3.2.1 Feature extraction by resonance line quantification

In the original study, a commercial program (Luise, Siemens, available at the tomograph) was used for the quantification of the three major resonance lines in the spectral region under study (cholines (Cho, chemical shift $\delta = 3.22$ ppm), creatines (Cr, $\delta = 3.01$ ppm), N-acetyl-aspartate (NAA, $\delta = 2.01$ ppm)). It included apodization, Fourier transformation, phase correction, and baseline flattening by fitting and subtracting a spline function. Results from this software were readjusted by the operator in the case of dissatisfactory line fits. To assess the performance of an automated quantification without a final operator control, we used jMRUI 2.0 [23] with soft constraints allowing $+/-0.03$ ppm frequency shift and a line width of 6.25 (range: 0 - 31.25) Hz.

As input for the following classification, we studied different representations of this data: values of the metabolite peaks as obtained from the integration (neither normalized with respect to a water reference nor to a metabolite value, table 1: "NAA Cr Cho" and "NAA Cr Cho auto" as obtained from MRUI), normalized by creatine ("NAA/Cr Cho/Cr") and a normalization proposed in [25] ("Cho/Cr Cho/NAA"). These data sets were optionally augmented with a categorical indication of lipid/lactate occurrence (lipid: true/false, lactate: true/false).

3.2.2 Feature extraction by nonparametric transformations

Prior to the application of any pattern recognition method to the spectrum, the residual water signal was removed using the HLSVD methods implemented in MRUI. Input for further processing was either the absolute Fourier transformed signal (magnitude spectrum), or the real phased part of the Fourier transform after a manual adjustment of the phase. In either case, the spectrum was normalized to the integrated amplitudes of the spectral region between and including the Cho and NAA peaks (3.4 to 2.0 ppm). Lipid and lactate resonances were excluded from the normalization to keep the pattern of the Cho, Cr and NAA region as constant as possible within the data set.

Two different kinds of transformations were applied to the spectra:

- transformations that are independent of the specific set of spectra, such as different wavelet transformations, the integration of the spectrum over isolated spectral regions around peaks, or the summation over a predefined number of neighboring channels (table 1, rows 9-16)
- transformations that were deduced from the pooled data set, for instance the principal and the independent component analysis (PCA, ICA, table 4).

Wavelets decompose a signal into more or less localized short and long range components and are frequently used in the denoising and compression of spectral data (for applications on MRS, e.g., see [33]). As spectra generally represent highly correlated signals, wavelets have the potential to express the relevant variation of a spectrum in a low number of wavelet coefficients. Besides the standard dyadic wavelets (of Daubechies-4 type), we used wavelet packages and continuous wavelets, which contain all possible combinations of low- and high-pass filtering or possible shifts, respectively. Binning, the integration over neighboring channels, amounts to a smoothing and subsampling of the spectral vector. It can be advantageous both in the presence of noise and small shifts of the resonance lines.

PCA extracts features from the data set that optimally represent the correlated variation in the data. In situations where this variation is due to strong interclass differences rather than noise or independent artifacts, PCA leads to a set of meaningful variables. ICA assumes that a signal is the superposition of several independent processes. These latent processes, or sources, are recovered in search of a transformation that, in our chosen algorithm, maximizes nonlinear correlation of the sources within the data set. For example: on a data set of spectra that show independently varying intensities of three resonance lines, ICA is supposed to extract these three peaks as latent sources of variation within the data set.

Some of these transformations increase the number of features (P) drastically. Thus, when P exceeded the number of spectra (N) by far (e.g. wavelet packages: $P = 4096 >> N = 58$), a mild univariate selection of the N most informative features was performed, namely according to the correlation between feature value and class label (table 1, rows 10,11). For a comparison of the feature representations, we included spectral vectors without any transformation into the evaluation of the classifiers. They either comprised solely the Cho, Cr and NAA spectral regions (table 1, rows 5-6, $P = 141$), or were extended to lower frequencies including lipids and lactate (table 1, row 7, $P = 256$; basis for all feature transformations in rows 9-16) or to both lower and higher frequencies (table 1, row 8, $P = 356$). All feature transformations described above were generally performed on the medium range, comprising resonance lines from Cho to Lip. For a quantitative comparison of a pattern recognition on magnitude (table 1, rows 6-8) and real phased spectra, we also included the latter in our benchmark (table 1, row 5)

3.2.3 Classification

We evaluated twelve different classification or regression methods both on the features from resonance line quantification and the other preprocessing methods (table 1). For the regression methods, a threshold on the response variable was learned to obtain a binary result from the predicted values. Due to the lack of reasonable assumptions on the clinical frequency of the two classes, balanced weights and cost functions were chosen for training and evaluation.

When applied to data sets such as the presented, standard methods (LDA, k-nearest-neighbors (k-NN), regression) tend to result in overtrained and unstable classifiers. Therefore, regularized multivariate linear regression methods were also evaluated: While ridge, lars and lasso regression are often motivated from statistical learning theory [11], partial least squares (PLS) and principal component regression (PCR) are traditional chemometrical methods [11]. Support vector machines (SVM) with linear and nonlinear (RBF) kernel and randomForest [3] decision tree ensembles are based on machine learning, the latter two being nonlinear learning algorithms.

All classifiers were evaluated on a reasonable range of parameters [20]. The classification error was assessed by leave-one-out cross-validation (table 1, figure 3). For the regression methods, the area under curve of the receiver operator characteristic (ROC-AuC) was additionally used to assess the quality of class separation (table 2). Lower and upper quartiles of the classification results were determined under the assumption of a binomial distribution (table 1, figure 3). Robust summaries of these distributions [19] were used to test for significant differences between the various processing strategies (fig. 3).

4 Results

In the following, benchmark results (table 1) on optimal feature extraction and classification will be summarized, followed by an analysis of the biological implications of the optimal decision rule.

– – – tables about here – – –

4.1 Feature extraction

Following the scheme of fig.1, feature extraction comprises two major approaches.

Resonance line quantification

Visual checking and manual readjustment leads to an improved classification performance (compare rows 1 and 4 in table 1). Normalization as proposed in

[25] (table 1, row 3) performs better than normalization with creatine intensity (table 1, row 2). Overall, the performance of a classification based on unnormalized quantification values is equivalent to results after the best normalization strategy (table 1: compare row 1 vs. 2 and 3). Additional use of categorical information concerning the occurrence of lipid or lactate resonances in the spectrum did not result in a notable change of the classifier performance (results not shown).

Pattern recognition

A pattern recognition based on magnitude spectra was more successful than one based on the real part of manually rephased spectra (table 1: compare rows 5 and 6). Compared to the pattern of magnitude spectra (figure 1), spectra phased to the real part show a higher variance of the spectral pattern. We cannot observe significant performance differences when using spectral regions of different widths (table 1, row 6-8; figure 3, number 1-3). While a classification based on coefficients from dyadic wavelets or wavelet packages performed as well as one on raw data (table 1, rows 9-11 vs. rows 6-8; fig. 3), a continuous wavelet significantly decreased the classifier performance.

Allowing for a multivariate selection of features, principal component analysis (PCA) was applied as part of principal component regression (PCR) with a constraint on the allowed dimensionality (similar to PLS). Even without this constraint and in conjunction with other classifiers, PCA proved to be beneficial when applied to a subset of the feature representations (table 4).

Independent component analysis (ICA) was applied in all principal component subspaces of up to twelve dimensions. When classifying on the basis of the scores of any of these ICA models, we could not observe an advantage compared with a direct classification based on the respective PCA scores (table 4).

Integrating the spectrum over consecutive bins ("binning") improved the average classification result of subsequent algorithms maximally (table 1, rows 13-16).

A restriction of the binning to the peak regions of Cho, Cr and NAA significantly decreased the performance on magnitude spectra (table 1, row 12).

Classification results of approximately 85–90% (table 1, last three rows, fig. 3) for most of the classifiers on any of the bin widths from five to fifteen channels easily surpassed performances from raw spectral vectors (≈ 80%–85%, table 1, rows 6–8) or the best results based on quantified resonance lines (≈ 75%–80%, table 1, rows 1–3).

4.2 Classification

All methods perform nearly equally well on the low-dimensional representation obtained from resonance line quantification. Here, PCR performs best (up to 83% classification accuracy, table 1, rows 1-3), although this difference is not significant.

The performance of standard methods such as k-NN and linear discriminant analysis (table 1, columns 7-8) on the differently preprocessed data sets is similar to the performance of all other classifiers, but generally tends to belong to the weaker part of this group. Tree ensembles (table 1, column 11) and RBF-kernel SVMs (table 1, columns 9-10) show good performance on binned parameter sets, although they never performed better than linear methods (table 1, columns 9-10), e.g. the linear kernel SVM.

Comparing standard multivariate linear regression (table 1, last column) and the related LDA with their constrained variants (table 1, columns 1-6) emphasizes the importance of regularization when N is small compared to the number of features.

In a comparison between algorithms deducing non-sparse features from the full spectrum (such as ridge, PLS, PCR – table 1, columns 1-3) on the one hand, and selective feature extractors (e.g. forward selection, lasso, lars – columns 4–6) on the other hand, the dimensionality of the regression model is of importance (see also table 4). A comparison in terms of the classification accuracy – the optimization criterion for all methods – of all preprocessing algorithms suggest the conclusion that forward selection, lars, and lasso offer a slight advantage in classification performance (table 1, row 4–6 vs. 1–3). However, the more informative measure of the ROC-AuC contradicts this observation (table 2). In summary, the exploitation of non-sparse, spectrum-wide features as performed by ridge regression, PLS and PCR appears more favorable than the "channel picking" by the selective algorithms. On the basis of the available data it is not possible to single out one of PCR, PLS or ridge regression as the best performing technique. These closely related [11] techniques all perform very well in the presence of high noise levels.

Feature extraction and classification

Overall, the classification problem is adequately addressed by linear methods. A smoothing and downsampling approach as realized in binning is the optimal feature representation of the given data. Classifiers tend to perform best under a (non-sparse) regularization.

In the comparison of pattern recognition and quantification, a conservative summary of the results is that pattern recognition methods perform at least as well as a manually supervised quantification (fig. 3, table 3). A less conservative assessment of our results ascribes a 5-10% higher accuracy to the pattern recognition methods.

An overall upper limit in the attainable classification accuracy for the distinction of tumor vs. non-progressive tumor (i.e., radiation injury and stable disease) after radiotherapy based on single-voxel ^{1}H MR spectra at 1.5 T is approximately 90%.

– – – table 3 about here – – –

– – – figure 4 about here – – –

– – – figure 5 about here – – –

4.3 Evaluation of the decision rules

In chemometrics, a careful inspection of the coefficients – the regression weights or "loadings" – of the applied linear regression methods is an established standard. On the given data set (regularized) linear regression methods performs best. Differently from non-linear classifiers, the decision rule of these models can be understood and an analysis concerning the importance of the five metabolites visible in the spectrum is possible (figs. 4, 5).

A comparison of the coefficient vector (fig. 4, box 3) with the average spectrum of the tumor group and the non-progressive tumor group (both in fig. 4, box 2) offers insight in the relation between learned decision rule and the biological signature if the tissue in the MR spectrum: The loadings are equivalent to a template pattern of the group means. A strong positive correlation with this pattern is given in the case of an (average) tumor spectrum (fig. 5, box 2), a strong negative correlation for an (average) non-progressive tumor spectrum. This observation corresponds to a well-known rule from signal processing: The optimal filter that can be used to detect a known signal corrupted by uncorrelated noise is the known signal itself - the "matched filter". Similarly, the optimal classifier detects spectra of recurrent tumor by a comparison or correlation of their pattern with the template pattern of a "typical" tumor spectrum.

Although very common, a direct biological interpretation of the loadings should be made with care. Loadings only represent channel "weights" for a decision about inter-group differences. More informative is the study of the channel-wise product of the *mean* tumor signal (as in fig. 5, box 2) with the loadings (loadings in fig. 5, box 3; product in box 4). It reveals the significance of the respective channels, as the outcome of the regression on an average tumor signal – the score – is the sum over this vector. Therefore, large positive entries along this graph indicate a high diagnostic relevance of the corresponding spectral regions in tumor detection. Similarly, a robust summary of the typical tumor pattern may be used: the *median* tumor signal of our data set is equivalent to the *mean* tumor pattern, but does not show high entries in the Lip/Lac region, as an increase of these resonance lines can only be observed for a subset of the tumor spectra.

In a channel-wise multiplication of the median signal with the regression weights, the sign of the Lip/Lac region is negative. This reduces the score of the tumor pattern from the Cho, Cr, and NAA spectral region alone, weakening the diagnosis of tumor for an otherwise unambigious tumor spectrum. It explains why the inclusion of lipid/lactate information is not useful for the classifiers discussed here.

In general, the lipid/lactate information is only marginally useful in the differentiation between recurrent tumor and non-tumorous lesions. Sufficient information for the diagnostic problem is, according to our data, already contained in the Cho, Cr, and NAA resonances (see fig. 4): the loadings of ridge regression/PLS/forward selection do not show high values in the spectral region of lipid/lactate[2].

[2] While this observation holds for a classification using solely MR spectra, we cannot exclude that the lipid/lactate region does hold information for a medical expert with access to supplementary information.

The amplitude of the creatine resonance is of comparatively minor interest. As with NAA, its relative decrease indicates the occurrence of tumor (figs. 4, 5). Studying the loadings on the full vector, a decrease of creatine and an increase in choline intensity marks the difference between the tumor and non-progressive tumor group. This finding is supported by a number of publications, mainly on tumor grading [25, 22, 34, 9, 2]. Nevertheless, the best classifiers on the basis of the "binning" type of preprocessing do not rely on this creatine intensity change (figs. 4, 5). The bad performance of the creatine normalization after resonance line quantification (table 1, row 2; fig. 3, column 13) underlines the low importance of this metabolite for the given classification purpose. When using intensity values normalized by creatine, its variance obviously decreases the reliability of the classifier (table 1, row 1).[3]

Effectively, only the variation of the choline and NAA resonances is the substantial information the optimal classifiers rely on.

4.4 Other binary decisions

Table 4 summarizes classification performances on binary decisions other than between recurrent tumor and non-progressive tumor. Values for pattern recognition are the averages of PLS, PCR and ridge regression after binning; for quantification, the best result within the four feature representations (as in table 1, rows 1-4) was chosen. So, values indicated in this table – in particular for quantification – are at the risk to be overly optimistic, but they might serve as an indicator for an upper limit in the given classification tasks.

Both pattern recognition and quantification perform equally well – with a slight advantage for pattern recognition methods. Normal and tumor spectra are separable with virtually no error. The subgroups of non-progressive tumor,

[3] The low information content does not counterbalance the associated dimensionality increase in the state-of-disease modeling and introduces a new source of noise in the analysis of resonance line integrals.

namely stable disease and radiation injury, are indistinguishable by all methods.

5 Discussion

As an overall result of the present study, we find that methods from pattern recognition and statistical learning can be applied successfully in the discrimination between recurrent brain tumor and a non-progressive state of disease after radiation therapy. These fully automated methods perform as well as semi-automatic procedures involving a manually supervised quantification step; and on our data they surpass results of a quantification without human control and interaction.

Focusing on technical details, we find a superiority of magnitude spectra when classifying the full spectral vector. The real part of phased complex spectra has a smaller linewidth, but also a greater variability of the spectral pattern which reduces the interpretability and decreases the performance of a pattern recognition [32]. A normalization of the spectrum to unit area in the spectral region of Cho, Cr, and NAA is sufficient to reach and possibly surpass classification performances from quantification. An external standard for normalization (i.e. a water reference) was not required.

The strong correlation of the spectral channels provides the basis for the best preprocessing of our noisy in vivo data: smoothing and subsampling and thus regularizing by the application of binning. Binning both reduces the dimensionality and allows for small shifts within the spectral pattern. It trades spectral resolution for increased signal-to-noise.

Other preprocessing methods, such as ICA transformations and wavelet representations could not prove their usefulness, which is in line with previous observations [27]. Transforms using smoother wavelets have not been investigated here, and may be better adapted to the given data. In spite of the argumentation put forward in [5], we observe an increase in performance when applying a dimension reduction prior to a classification by SVMs.

We found non-sparse regularized linear regression methods, such as ridge regression, PLS or PCR, to be optimal. Nonlinear SVMs or ensemble methods were not required. We recommend these linear methods when a discrimination between linearly separable groups is sought in binary settings. As a consequence of the simplicity of these linear regression models, their decision rules can be visualized directly and we have demonstrated how to compare these loadings with established clinical knowledge.

Overall, at the noise level that is typical for in vivo MRS at 1.5 Tesla, regularized linear regression methods taking the entire spectrum as input perform at least as well as, and possibly even better than, classification algorithms relying on manually corrected resonance line quantifications.

As a separation between normal state and tumor spectra is possible without error (table 2), the extension of this binary classification problem to a three-class-classifier, i.e. normal – non-progressive – tumor, for example in the automated classification of whole spectroscopic images is straightforward [11, 6] as long as there are no mixed voxels. Since both the classifier for healthy vs. tumorous tissue and non-progressive vs. progressive tumor rely primarily on the Cho/NAA intensity ratio, a mixture of the former[4] is difficult to distinguish from a non-progressive tumor alone.

As a consequence, a severe limitation of current fully automated approaches is that they must be limited to a given anatomical region by a medical expert. Future algorithms may take into account prior spatial information from pre-therapeutic diagnostics, the radiation therapy plans, or may exploit information from other imaging modalities.

At present, an approximate classification accuracy of 90% between progres-

[4] MR spectra are superpositions [29, 26] of all tissues types in the voxel.

sive and non-progressive tumor after radiotherapy can be reached by means of single-voxel MRS, and the exploitation of complementary information beyond MR spectra will be required to approach the desired 100%. As the joint interpretation of different sources of information is the strength of automated methods, we see the potential of a highly automated multi-modal imaging – combining MRI and MRSI – in the detection and localization of recurrent brain tumor [27]. High-resolution SI [7] can help to increase the sensitivity for small tumors and to reduce problems with mixed voxels. Here, (semi-) manual analysis of an entire data set will become increasingly more time-consuming and we expect automatic classification tools to become an indispensable high-level "preprocessing" for medical experts.

Acknowledgements

The authors gratefully acknowledge the suggestions of the three anonymous reviewers who helped to improve the content and presentation. The authors also acknowledge financial support by the Deutsche Forschungsgemeinschaft under grant DFG-HA-4364.

References

[1] T. F. Bathen, J. Krane, T. Engan, K. S. Bjerve, and D. Axelson, *Quantification of plasma lipids and apolipoproteins by use of proton NMR spectroscopy, multivariate and neural network analysis*, NMR Biomed. (2000), 271–288.

[2] M. Bendszus, M. Warmuth-Metz, R. Klein, R. Burger, C. Schichor, J. C. Tonn, and L. Solymosi, *MR spectroscopy in gliomatosis cerebri*, Am. J. Neuroradiol. **21** (2000), no. 2, 375–80.

[3] L. Breiman, *Random forests*, Machine Learning Journal (2001), 5–32.

[4] J. Butzen, R. Probst, D. Chetty, K. Donahue, R. Neppl, W. Bowen, S. Li, V. Haughton, M. Leighton, T. Kim, W. Mueller, G. Meyer, H. Krouwer, and S. Rand, *Discrimination between neoplastic and nonneoplastic brain lesions by use of proton MR spectroscopy: The limits of accuracy with a logistic regression model*, Am. J. Neuroradiol (2000), 1213–1219.

[5] A. Devos, L. Lukas, J. A. Suykens, L. Vanhamme, A. R. Tate, F. A. Howe, C. Majos, A Moreno-Torres, M. van der Graaf, C. Arus, and S. van Huffel, *Classification of brain tumours using short echo time* ^{1}H *MR spectra*, J. Magn. Reson. **170** (2004), no. 1, 164–75.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Wiley, New York, 2000.

[7] U. Dydak, K. P. Prüssmann, M. Weiger, J. Tsao, D. Meier, and P. Bösiger, *Parallel spectroscopic imaging with spin-echo trains*, Magn. Res. Med. **50** (2003), 196–200.

[8] W. El-Deredy, *Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: a review*, NMR Biomed. **10** (1997), 99–124.

[9] D. Galanaud, O. Chinot, F. Nicoli, S. Confort-Gouny, Y. Le Fur, M. Barrie-Attarian, J. P. Ranjeva, S. Fuentes, P. Viout, D. Figarella-Branger, and P. J. Cozzone, *Use of proton magnetic resonance spectroscopy of the brain to differentiate gliomatosis cerebri from low-grade glioma*, J. Neurosurg. **98** (2003), 269–76.

[10] G. Hagberg, *From magnetic resonance spectroscopy to classification of tumors. A review of pattern recognition methods*, NMR Biomed. (1998), 148–156.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer Series in Statistics, Springer, New York, 2001.

[12] S. Herminghaus, T. Dierks, U. Pilatus, W. Moller-Hartmann, J. Wittsack, G. Marquard, C. Labisch, H. Lanfermann, W. Schlote, and E. Zanella, *Determination of histopathological tumor grade in neuroepithal brain tumors by using spectral pattern analysis of in vivo spectroscopic data*, J. Neurosurg. (2003), 74–81.

[13] A. F. Howe and K.S. Opstad, *^1H MR spectroscopy of brain tumour and masses*, NMR Biomed. (2003), 123–131.

[14] H. J. A. in't Zandt, M. van der Graf, and A. Heerschap, *Common processing of in vivo MR spectra*, NMR Biomed. (2001), 224–232.

[15] C. Ladroue, F. A. Howe, J. R. Griffiths, and A. R. Tate, *Independent component analysis for automated decomposition of in vivo magnetic resonance spectra*, Mag. Res. Med. **50** (2003), 697–703.

[16] M. Lichy, C. Plathow, D. Schulz-Ertner, H. Kauzcor, and H.-P. Schlemmer, *Follow-up of gliomas after radiotherapy: ^1H MR spectroscopic imaging for increasing diagnostic accuracy*, Neuroradiology (2005), 1–12.

[17] P. J. G. Lisboa, S. P. J. Kirby, A. Vellido, Y. Y. B. Lee, and W. El-Deredy, *Assessment of statistical and neural networks methods in NMR spectral classification*, NMR Biomed. **11** (1998), 225–234.

[18] C. Majos, J. Alonso, C. Aguilera, M. Serrallonga, J. Perez-Martin, J.J. Acebes, C. Arus, and J. Gill, *Proton magnetic spectroscopy (^1H MRS) of*

human brain tumours: assessment of differences between tumour types and its applicability in brain tumour categorization, Eur. Radiol. (2003), 582–591.

[19] R. McGill, J. W. Tukey, and W. A. Larsen, *Variations of box plots*, The American Statistician **32** (1978), 12–16.

[20] B. H. Menze, M. Wormit, P. Bachert, M. P. Lichy, H.-P. Schlemmer, and F. A. Hamprecht, *Classification of in vivo magnetic resonance spectra*, Classification, the ubiquitous challenge: Proceedings of GfKl 2004, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 2005.

[21] S. Mierisova and M. Ala-Korpels, *MR spectroscopic quantitation: a review of frequency domain methods*, NMR Biomed. (2001), 247–259.

[22] W. Moller-Hartmann, S. Herminghaus, T. Krings, G. Marquart, H. Lanfermann, U. Pilatus, and F. E. Zanella, *Clinical application of proton magnetic resonance spectroscopy in the diagnosis of intracranial mass lesions*, Neuroradiology (2002), 371–381.

[23] A. Naressi, C. Couturier, J. M. Devos, M. Janssen, C. Mangeat, R. de Beer, and D. Graveron-Demilly, *jMRUI, MRUI for Java*, MAGMA **12** (2001), 141–152.

[24] S. W. Provencher, *Estimation of metabolite concentrations from localized in vivo proton NMR spectra*, Magn. Reson. Med. **30** (1993), 672–9.

[25] H.-P. Schlemmer, P. Bachert, K. K. Herfarth, I. Zuna, J. Debus, and G. van Kaick, *Proton MR spectroscopic evaluation of suspicious brain lesions after stereotactic radiotherapy*, Am. J. Neurorad. **22** (2001), 1316–1324.

[26] U. Seeger, T. Naegele, I. Mader, M. Erb, and U. Klose, *Calculation of pure tissue spectra in proton MR spectroscopy of brain diseases*, Proc. ESMRMB, 2003, p. 331.

[27] A. W. Simonetti, W. J. Melssen, F. Szabo de Edelenyi, J. J. van Asten, A. Heerschap, and L. M. Buydens, *Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification*, NMR Biomed. (2005), 34–43.

[28] A. W. Simonetti, W. J. Melssen, M. Van der Graaf, G. J. Postma, A. Heerschap, and L. M. C. Buydens, *A chemometric approach for brain tumor classification using magnetic resonance imaging and spectroscopy*, Anal. Chem. **75** (2003), 5352–5361.

[29] A. Stadlbauer, O. Ganslandt, S. Gruber, Nimsky C., R. Busler, R. Fahlbusch, and E. Moser, *Improved preoperative diagnostics of brain tumors by quantification of ^1H MRSI metabolites*, Proc. ISMRM, 2004, p. 2053.

[30] R. Stoyanova and T. R. Brown, *NMR spectral quantitation by principal component analysis. III. A generalized procedure for determination of lineshape variations*, J. Magn. Reson. **154** (2002), 163–75.

[31] A. R. Tate, *Pattern recognition of in vivo magnetic resonance spectra*, Ph.D. thesis, Sussex University, 1996.

[32] A. R. Tate, C. Majos, A. Moreno, F. A. Howe, J. R. Griffith, and C. Arus, *Automated classification of short echo time in vivo ^1H brain tumor spectra: a multicenter study*, Magn. Res. Med. (2003), 29–36.

[33] A. R. Tate, D. Watson, S. Eglen, T. N. Arvanitis, E. L. Thomas, and J. D. Bell, *Automated feature extraction for the classification of human in vivo*

^{13}C *NMR spectra using statistical pattern recognition and wavelets*, Magn. Res. Med. (1996), 834–40.

[34] A. A. Tzika, D. Zurakowski, T. Young Poussaint, L. Goumnerova, L. G. Astrakas, P. D. Barnes, D. C. Anthony, A. L. Billet, N. J. Tarbell, R. M. Scott, and P. M. Black, *Proton magnetic spectroscopic imaging of the child's brain. The response of tumors to treatment*, Neuroradiology **43** (2001), 169–177.

[35] L. Vanhamme, T. Sundin, P. van Hecke, and S. van Huffel, *MR spectroscopic quantititation: a review of time-domain methods*, NMR Biomed. (2001), 233–246.

[36] L. Vanhamme, A. van den Boogaart, and S. van Huffel, *Improved method for accurate and efficient quantification of MRS data with use of prior knowledge*, J. Magn. Reson. **129** (1997), 35–43.

[37] H. Witjes, W. J. Melssen, H. J. in't Zandt, M. van der Graaf, A. Heerschap, and L. M. C. Buydens, *Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets*, J. Magn. Reson. **144** (2000), 35–44.

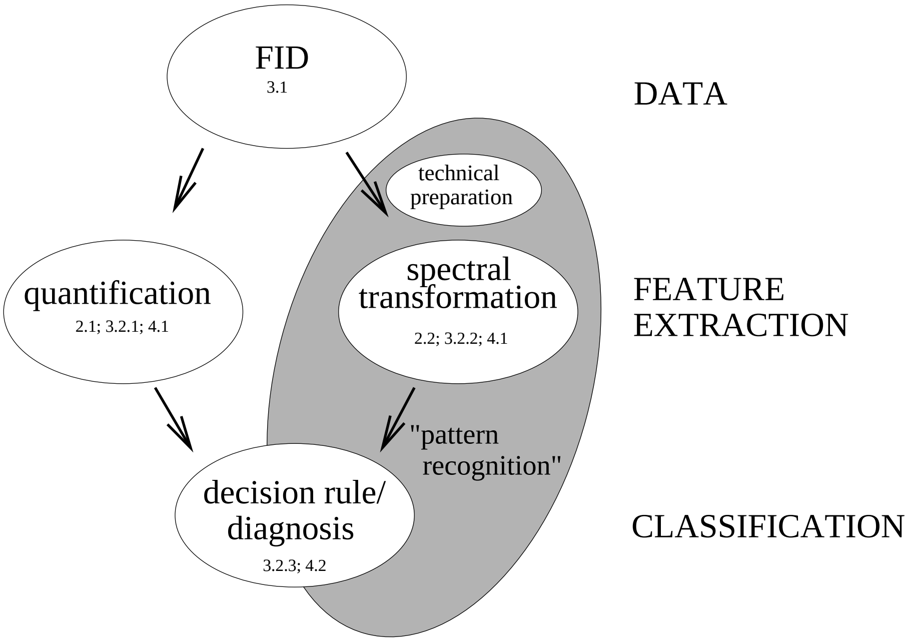

Figure 1: Flowchart of the different approaches in the extraction of diagnostic information from magnetic resonance spectra. Both quantification and model-free processing of the raw data can be seen as feature extraction prior to a classification. The grey ellipsoid indicates the approach of spectral analysis, which is termed “pattern recognition”. The “technical preparation” comprises operations such as water peak removal and Fourier transformation of the FID, which are implicit parts of the quantification (in the time-domain). Numbers refer to sections of this paper.

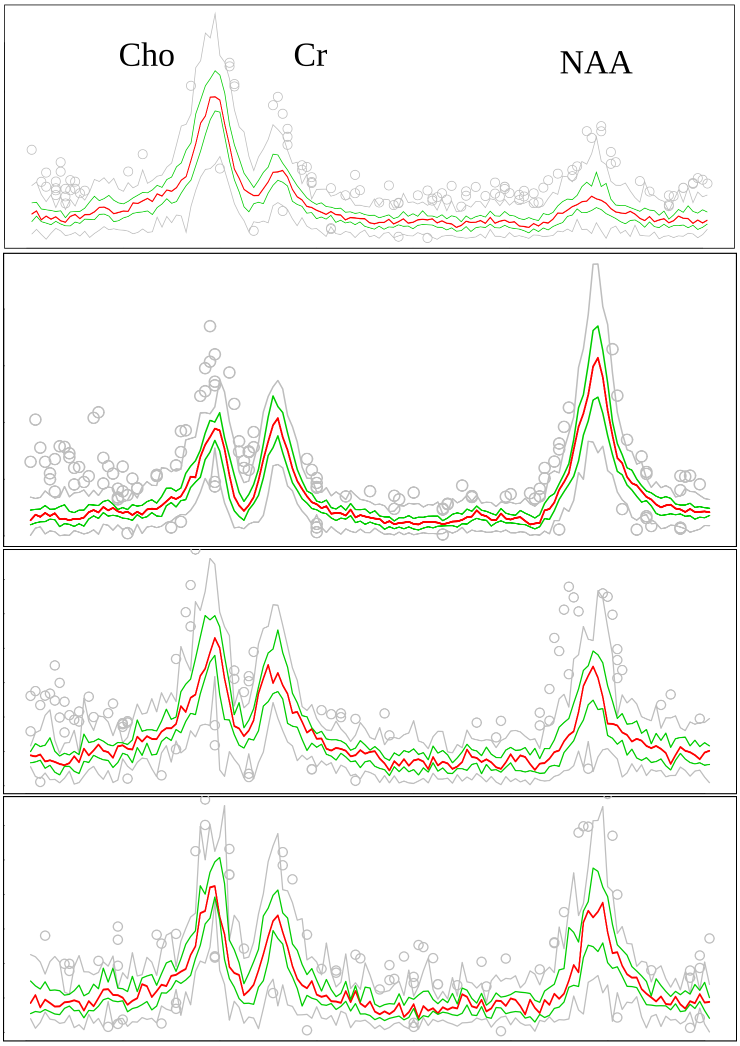

Figure 2: Characteristic patterns of magnitude spectra (central black lines) and their variation (inner and outer grey lines) in the region of the Cho, Cr and NAA resonances. From top to bottom: progressive tumor, healthy tissue, radiation injury, stable disease. Lines indicate the 25% and 75% quartiles (inner grey lines) and median (black) of the marginal empirical distribution in each spectral channel, outermost grey lines and points indicate outliers.

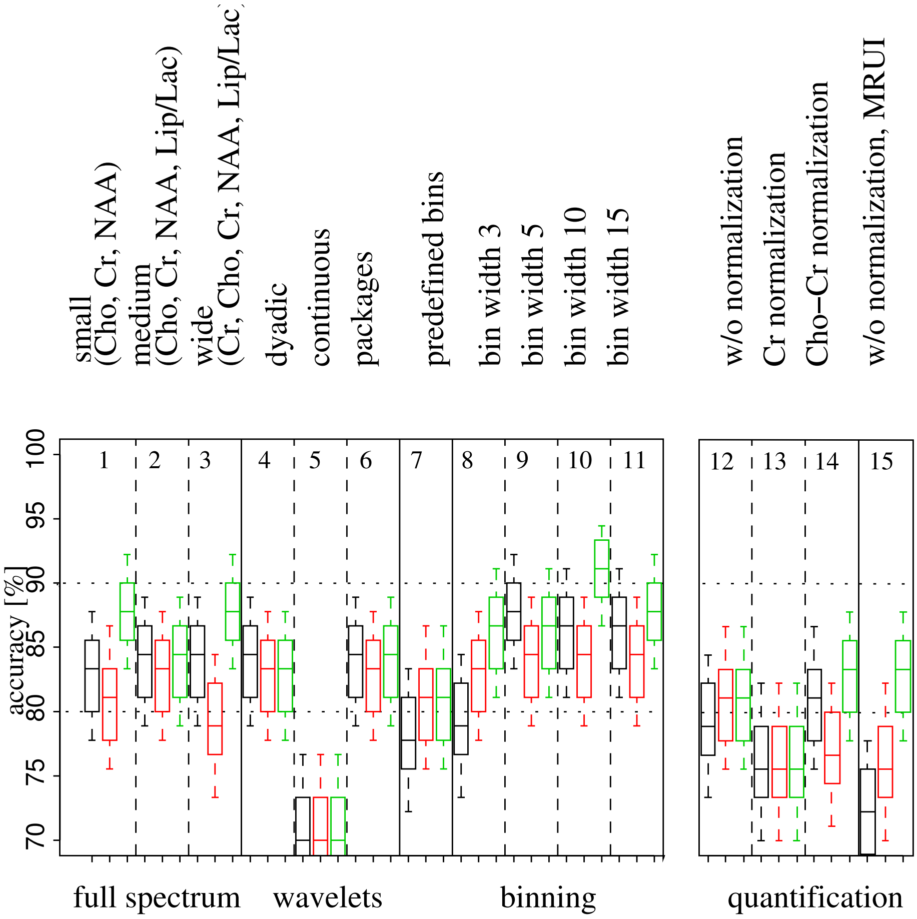

Figure 3: Cross-validated classification accuracy for progressive vs. non-progressive tumor as a result of varying spectral information extraction. The accuracy of the three top performing algorithms is indicated: ridge regression, PLS and PCR (left/middle/right bar for each method, box indicating median and quartiles, whiskers marking 10%/90% quantiles of the assumed binomial distribution). For the presented data, an estimation based on [19] comes to the conclusion that a significant difference of the median is given at $p < 0.05$ for differences $> 2.5\%$ accuracy, at $p < 0.005$ for differences $> 4\%$.

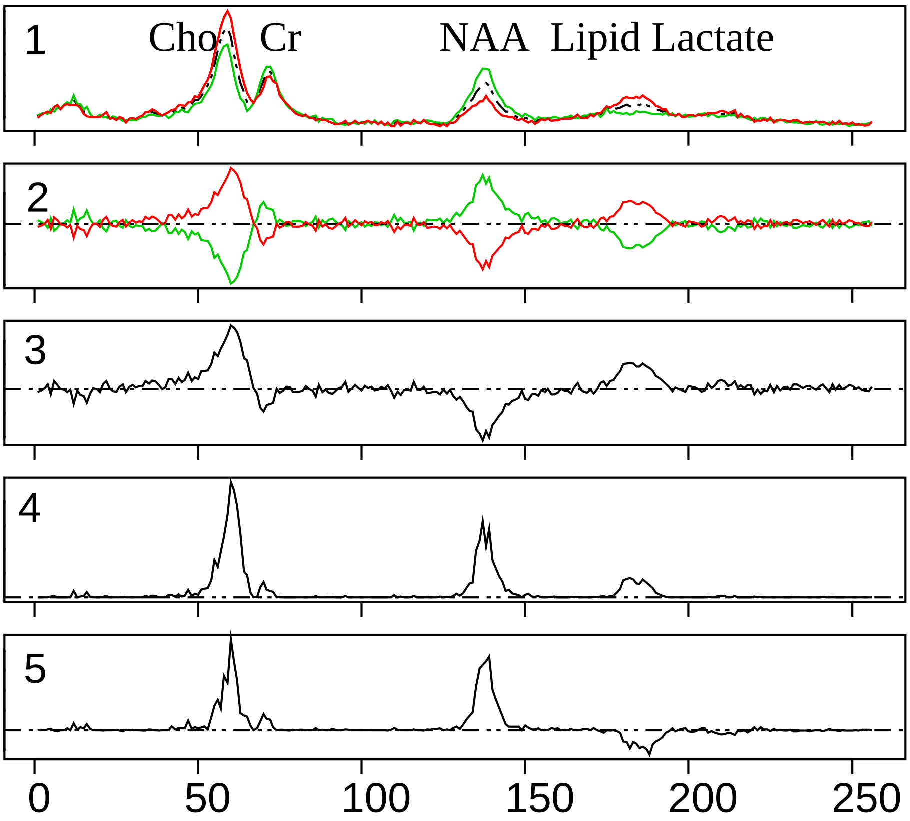

Figure 4: Interpretation of the linear classifiers. Box 1: Mean of the tumor (grey line) and non-progressive tumor group (dark line), and the overall mean of the data set (dashed line). Box 2: Similar to box 1, with the overall mean removed. Box 3: Coefficients of the ridge regression (84% accuracy) without binning. Box 4: Channelwise product of the *mean* tumor signal (as in box 2) with the coefficients from box 3. Box 5: Channel-wise product of the *median* tumor signal with the coefficients from box 3. See section 4.3 for details.

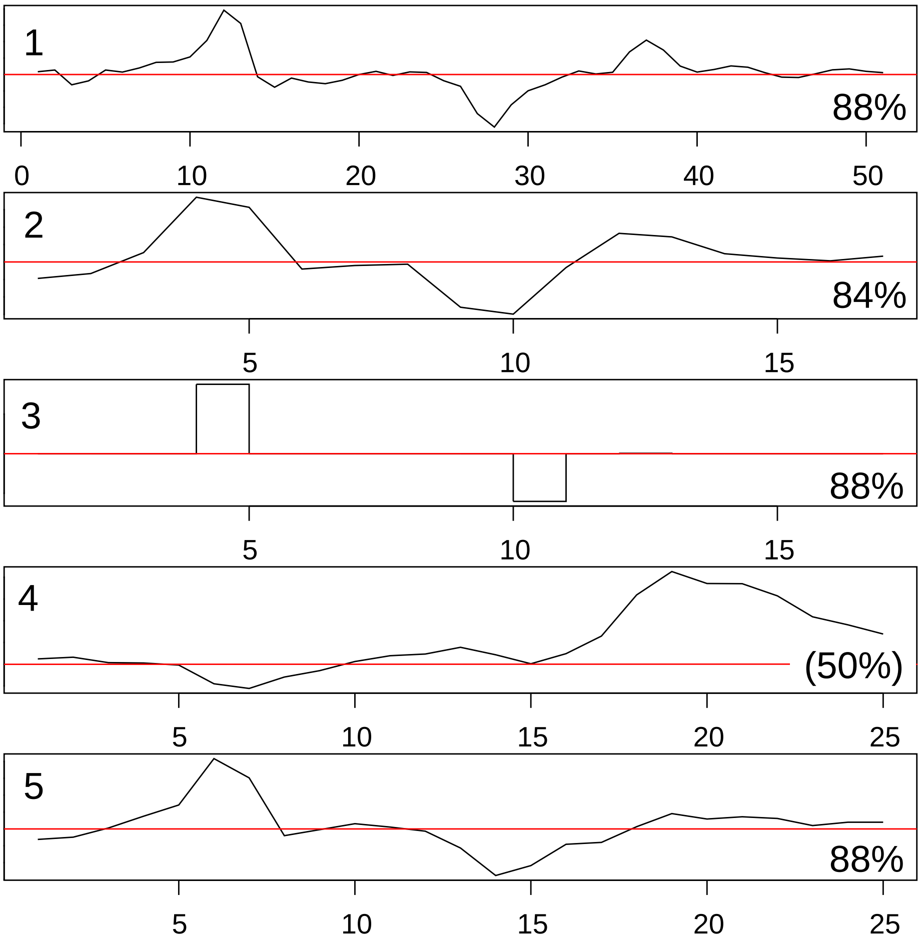

Figure 5: Regression coefficients after different preprocessing. Box 1: ridge coefficients (bin 5). Box 2: PLS coefficients (bin 15). Box 3: coefficients from forward selection (bin 15). Boxes 4 and 5: PCR first and second component.– The accuracy of the respective classification is given in percent. While the sole use of the first PCA component (accounting for lipid/lactate variation) does not allow for a classification (50% accuracy = error of random classification), the additional use of the Cho/NAA information improves the result significantly.

		ridge	PLS	PCR	forw	lasso	lars
	Preprocessing	1	2	3	4	5	6
1	NAA Cr Cho	79 (77, 82)	81 (78, 83)	81 (78, 83)	78 (76, 81)	81 (78, 83)	81 (78, 83)
2	NAA/Cr Cho/Cr	76 (73, 79)	76 (73, 79)	76 (73, 79)	<70	<70	<70
3	Cho/Cr Cr/NAA	81 (78, 83)	77 (74, 80)	83 (80, 86)	<70	78 (76, 81)	78 (76, 81)
4	NAA Cr Cho auto	72 (69, 76)	76 (73, 79)	83 (80, 86)	<70	<70	<70
5	small real	78 (76, 81)	71 (68, 75)	76 (73, 79)	<70	<70	<70
6	small abs	83 (80, 86)	81 (78, 83)	88 (86, 90)	83 (80, 86)	84 (81, 87)	84 (81, 87)
7	medium abs	84 (81, 87)	83 (80, 86)	84 (81, 87)	83 (80, 86)	84 (81, 87)	84 (81, 87)
8	wide abs	84 (81, 87)	79 (77, 82)	88 (86, 90)	83 (80, 86)	84 (81, 87)	84 (81, 87)
9	wavelet	84 (81, 87)	83 (80, 86)	83 (80, 86)	74 (71, 77)	<70	<70
10	contw	<70	<70	<70	<70	83 (80, 86)	79 (77, 82)
11	wpack	84 (81, 87)	83 (80, 86)	84 (81, 87)	83 (80, 86)	83 (80, 86)	81 (78, 83)
12	pbins	78 (76, 81)	81 (78, 83)	81 (78, 83)	81 (78, 83)	81 (78, 83)	81 (78, 83)
13	bin3	79 (77, 82)	83 (80, 86)	86 (83, 89)	81 (78, 83)	81 (78, 83)	81 (78, 83)
14	bin5	88 (86, 90)	84 (81, 87)	86 (83, 89)	88 (86, 90)	84 (81, 87)	84 (81, 87)
15	bin10	86 (83, 89)	84 (81, 87)	91 (89, 93)	88 (86, 90)	88 (86, 90)	88 (86, 90)
16	bin15	86 (83, 89)	84 (81, 87)	88 (86, 90)	90 (88, 92)	88 (86, 90)	88 (86, 90)

		k-NN	LDA	SVM RBF	SVM linear	RForest	regr
	Preprocessing	7	8	9	10	11	12
1	NAA Cr Cho	<70	78 (76, 81)	79 (77, 82)	83 (80, 86)	76 (73, 79)	79 (77, 82)
2	NAA/Cr Cho/Cr	<70	78 (76, 81)	79 (77, 82)	79 (77, 82)	72 (69, 76)	76 (73, 79)
3	Cho/Cr Cho/NAA	<70	74 (71, 77)	74 (71, 77)	74 (71, 77)	71 (68, 74)	81 (78, 83)
4	NAA Cr Cho auto	<70	<70	<70	<70	<70	<70
5	small real	78 (76, 81)	<70	74 (71, 77)	74 (71, 77)	< 70	<70
6	small abs	84 (81, 87)	72 (69, 76)	81 (78, 83)	76 (73, 79)	90 (88, 92)	< 70
7	medium abs	79 (77, 82)	78 (76, 81)	78 (76, 81)	74 (71, 77)	88 (86, 90)	< 70
8	wide abs	83 (80, 86)	<70	76 (73, 79)	<70	88 (86, 90)	< 70
9	wavelet	81 (78, 83)	83 (80, 86)	76 (73, 79)	76 (73, 79)	78 (76, 81)	< 70
10	contw	81 (78, 83)	79 (77, 82)	79 (77, 82)	80 (78 82)	79 (77, 82)	81 (78, 83)
11	wpack	79 (77, 82)	79 (77, 82)	79 (77, 82)	81 (70 83)	80 (78 82)	< 70
12	pbins	72 (69, 76)	78 (76, 81)	79 (77, 82)	79 (77, 82)	79 (77, 82)	78 (76, 81)
13	bin3	76 (73, 79)	<70	84 (81, 87)	76 (73, 79)	83 (80, 86)	< 70
14	bin5	76 (73, 79)	<70	88 (86, 90)	86 (83, 89)	86 (83, 89)	< 70
15	bin10	81 (78, 83)	76 (73, 79)	88 (86, 90)	88 (86, 88)	87 (84, 90)	74 (71, 77)
16	bin15	84 (81, 87)	79 (77, 82)	88 (86, 90)	90 (88, 92)	87 (84, 90)	78 (76, 81)

Table 1: Cross-validated classification accuracy (in %) for the discrimination between progressive and non-progressive tumor (comprising radiation injury and stable disease, excluding normal tissue). Numbers in parentheses are the quartiles obtained from leave-one-out cross-validation under the assumption of a binomial distribution. Rows and columns correspond to different preprocessing methods and classifiers discussed in the text.

	ridge	PLS	PCR	forw	lasso	lars	reg
small	89	87	89	87	87	87	62
medium	90	90	90	87	87	87	67
wide	90	91	90	87	87	87	56
wavelet	89	89	89	88	87	85	67
contw	87	87	87	85	86	87	81
wpack	89	89	89	88	87	85	62
pbins	80	82	82	81	81	81	79
bin3	89	90	89	88	88	88	54
bin5	93	91	91	89	89	89	67
bin10	93	93	92	92	92	92	84
bin15	92	91	91	92	92	92	86

Table 2: Area under curve of the receiver operator characteristic (ROC - AuC), as evaluated for the regression-based classifiers under study in the discrimination of progressive tumor vs. non-progressive tumor.

group	tumor	SD	RI	nPT	group size (# patients)
normal	100 (100, 100) 98 (97, 100)	93 (91, 96) 89 (85, 91)	87 (84, 91) 90 (87, 93)	88 (85, 90) 84 (81, 87)	32
tumor		83 (80, 87) 88 (84, 91)	87 (83, 90) 90 (86, 93)	83 (80, 87) 88 (84, 91)	30
SD			50 (42, 57) 55 (50, 60)	–	15
RI				–	13

Table 3: Accuracy for binary classification tasks from the data set of this study. Accuracy is assessed from the average performance of the three best classifiers (PLS, PCR, ridge after binning – fig. 3, compare with "tumor vs. non-progressive tumor" here). Values are given on the basis of both data analysis methodologies: upper rows – resonance line quantification, lower rows – pattern recognition. SD – stable disease, RI – radiation injury, nPT - non-progressive tumor. All numbers result from cross-validation, values in parentheses represent quartiles (table 1).

	medium w/o	medium PCA	medium ICA	bin 15 w/o	bin 15 PCA	bin 15 ICA
ridge	84%	84%	81%	86%	86%	86%
forw	82% (6)	88% (3)	79% (7)	88% (6)	88% (2)	84% (2)
lasso	84% (6)	88% (3)	81% (7)	84% (5)	86% (2)	84% (9)
PCR	84% (3)	84% (3)	81% (7)	88% (2)	88% (2)	86% (9)
lars	84% (6)	88% (3)	81% (7)	88% (5)	86% (2)	84% (9)
PLS	83% (1)	NA	81% (1)	84% (1)	NA	86% (1)

Table 4: Merit of principal component / independent component analysis (PCA/ICA) and binning. Values in percent indicate the cross-validated classification accuracy, the numbers of nonzero regression coefficients for the optimal model are given in brackets. From left to right: medium range spectrum: without transformation, with PCA, with ICA; the same spectral region binned with bin width 15: without transformation, with PCA, with ICA.