Automated Estimation of Tumor Probability in Prostate MRSI: Pattern Recognition vs. Quantification

B. Michael Kelm[1], Bjoern H. Menze[1], Christian M. Zechmann[2], Klaus T. Baudendistel[2], and Fred A. Hamprecht[1]

[1]Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg
[2]German Cancer Research Center (dkfz), Heidelberg

July 5, 2006

Automated Estimation of Tumor Probability in Prostate MRSI

Word Count: 5250

Corresponding Author:
Fred Hamprecht
Interdisciplinary Center for Scientific Computing
Im Neuenheimer Feld 368, 69120 Heidelberg
phone: +49 (0) 6221 548800
fax: +49 (0) 6221 548850
fred.hamprecht@iwr.uni-heidelberg.de

Abstract

Despite its diagnostic value and technological availability, ^{1}H NMR spectroscopic imaging (MRSI) has not found its way into clinical routine yet. Prerequisite for the clinical application is an automated and reliable method for the diagnostic evaluation of MRS images. In the present paper, different approaches to the estimation of tumor probability from MRSI in the prostate are assessed. Two approaches to feature extraction are compared: quantification (VARPRO, AMARES, QUEST) and subspace methods on spectral patterns (principal components, independent components, nonnegative matrix factorization, partial least squares). Linear as well as non-linear classifiers (support vector machines, Gaussian processes, random forests) are applied and discussed. It is found that quantification-based approaches are much more sensitive to the choice and parameterization of the quantification algorithm than to the choice of the classifier. Furthermore, linear methods based on magnitude spectra easily achieve equal performance and also allow for biochemical interpretation in combination with subspace methods. Nonlinear methods operating directly on magnitude spectra achieve the best results but are less transparent than the linear methods.

Keywords: magnetic resonance spectroscopic imaging; pattern recognition; classification; quantification;

1 Introduction

Clinical studies have shown significant diagnostic value of ^{1}H magnetic resonance spectroscopic imaging (MRSI) for the detection of tumorous tissue in the prostate (3, 19, 24, 25, 28, 33). Despite the promising results of these and other studies, the integration of MRSI in the clinical routine remains difficult, mainly because of the complexity and effort associated with the evaluation of the acquired data.

Two basic approaches to the evaluation of MRSI can be distinguished: the *quantification-based* approach and the *pattern recognition* (PR) approach. Quantification aims at estimating relative metabolite concentrations as accurately as possible. For that purpose, the most likely parameter estimate for a given signal model is usually determined with a nonlinear least squares (NLS) approach. However, quantification may fail for various reasons. In particular, in the presence of artifacts and severe noise the NLS objective can have many local optima and the

result becomes very sensitive to the choice of initial values. Prior knowledge about the expected signal shape can help to alleviate these problems (30), but, it also leads to an estimation bias and can be harmful in unanticipated cases where the employed prior knowledge is inadequate. A subsequent statistical analysis which gains diagnostic information from the spectral data relies on these parameter estimates and therefore inherits all problems associated with the quantification.

Pattern recognition approaches do not require an explicit quantification step. Although the same methods and classifiers (e.g. logistic regression (9), artificial neural networks (9), support vector machines (SVM) (26), etc.) can be used for both, quantified signals and spectral patterns, only methods applied to the latter will be referred to as "pattern recognition" (PR) approaches in accordance with, for example, (8). The PR approach is characterized by minimal preprocessing, thus avoiding errors introduced by feature calculation steps. It is left to the classifier to construct features and extract the relevant information to distinguish random effects from significant changes in the spectral pattern. Since it is not exact quantification that is the main goal in clinical applications but accurate diagnostic information, we suggest to address the diagnostic problem directly without prior quantification.

In the following, we briefly review related work in order to emphasize common ideas and highlight differences with our approach. Recently, encouraging results have been reported (5, 6, 13, 27, 29) from studies on the automated classification of brain tumor spectra in the context of the INTERPRET project (1). Tate et al. (29) show that the influence of acquisition parameters (manufacturer, sequence, TE, TR) on the spectral pattern is small enough to allow for stable classification results across multiple centers. In (5) and (13), Devos and Lukas et al. examine and compare different preprocessing strategies and classifiers for long and short echo time brain spectra respectively. They show that the best results are obtained with L_2-normalized magnitude spectra, omitting for example baseline and phase corrections. Although a nonlinear classifier has been employed, no improvement over linear classifiers could be observed in both studies, which the authors attribute to the limited amount of available data.

Laudadio et al. (12) propose a PR approach using magnitude spectra that incorporates spatial context. It is applied to simulated as well as *in vivo* prostate MRSI data and focuses on evaluating the benefit of incorporating spatial information.

In (17), Menze et al. examine classifiers for the discrimination of recurrent tumor and brain

lesions after radiotherapy based on single voxel MRS. An exhaustive combination of feature extraction methods and classifiers is benchmarked according to several error measures. Regularized linear classifiers with preceding dimensionality reduction (binning) are found to perform best on the given data set.

To the best of our knowledge, a similar comparative study has not been performed on prostate MRSI data yet. In the present work an extensive collection of linear subspace methods and a representative set of state-of-the-art nonlinear classifiers are evaluated on prostate data. For the first time also the influence of different quantification algorithms on the classification results is examined. Furthermore, we conduct experiments comparing the use of magnitude and real spectra. Since it is common practice in prostate MRSI to analyze the acquired data based on quantification (3, 19, 24, 25, 28, 33), we emphasize the comparison of quantification-based approaches with PR approaches. R3.5

2 Methods

Only approaches that can be used for a fully automated analysis of MRSI data are considered in this study because extensive user interaction is not acceptable in clinical routine use. We also concentrate on methods that can provide tumor *probability* estimates, a much richer description of classification results than hard class labels. Finally, all selected methods have either been proposed for NMR spectroscopic data before or are closely related to such methods.

The section starts with a short description of the employed data set. Subsequently, the used feature extraction and classification methods are concisely summarized with ample references to the literature. The last subsection is devoted to the error measure used to compare the different approaches.

2.1 Data

^{1}H-NMR spectroscopic image volumes from an ongoing prostate MRSI study have been collected at the German Cancer Research Center (dkfz, Heidelberg). The data was acquired on a clinical 1.5T scanner (Magnetom Symphony; Siemens Medical Solutions, Erlangen, Germany) with a disposable endorectal coil (MRInnervu; Medrad Inc., Indianola, PA, USA) and the pro-

tocol described in (23, 24). 512 datapoints with a bandwidth of 1000-1250 Hz were acquired (TE/TR=120/650 ms). The field of view (FOV) and the volume of interest (VOI; selected with PRESS pulses) were adapted to the size of the individual prostates. Typical FOVs were around 60-66 $\times$ 78-84 $\times$ 66-78 mm. An elliptical k-space acquisition scheme and apodization with a Hanning filter was employed (23). The total acquisition time was limited to 10 minutes and the spectral data was interpolated to yield a volume of 16^3 voxels. Along with the MRSI data, T_2-weighted axial MR images (turbo-spin echo, TE=129, TR=4000-4800 ms, FOV= 140 $\times$ 140 mm, matrix size 512 $\times$ 512, 20-25 slices, slice thickness = 4 mm) were acquired. Two exemplary spectra are shown in Fig. 1.

- - - - - - - - - - Fig. 1 about here - - - - - - - - - -

For 12 of the 36 recorded patients, poor shimming, ineffective fat suppression or problems with the endorectal coil resulted in corrupted MRSI data. These patients have been excluded from the data set. For several patients, results from a histologic step-section examination were available. These could be used as "gold standard" for a qualitative evaluation. The training set was created using a semimanual analysis of the spectra according to standard decision rules based on the metabolite resonances of Cho, Cr and Ci (19, 28, 32) . Altogether, 76 slices with 256 voxels each have been labeled with respect to their spectral pattern class (healthy, undecided, tumor) and the signal quality (not evaluable, poor, good). In judging the signal quality both, low signal-to-noise ratios and artifacts (nuisance resonances, heavy baselines) have been considered. An overview of the collected data is given in Tab. 1. Only spectra that are evaluable (signal quality "poor" and "good") have been used in this study. The large number of "not evaluable" voxels is due to outer volume suppression and the limiting coil sensitivity profile in prostate MRSI. Only about one fourth of the voxels in the FOV actually lie within the prostate.

- - - - - - - - - - Tab. 1 about here - - - - - - - - - -

2.2 Preprocessing and Feature Extraction

Both quantification and PR profit from the prior removal of nuisance peaks and baselines in the spectra. Therefore, prior to further processing, the residual water and lipid resonances were removed by time-domain selective HSVD filtering (20), i.e. by removing all signal components with poles outside the interesting frequency range of 2.4 to 3.6 ppm.

Quantification. Three different methods have been used for quantification: QUEST (21) and AMARES (30) from the jMRUI tool (18) and a custom implementation of a constrained VARPRO approach which used an interior trust region algorithm for optimization (4). Quantification was performed with four Lorentzian components (cf. Tab. 2). Besides small frequency shifts of $\pm.03$ ppm for the individual components, a common shift of up to $\pm.625$ ppm was allowed for in the VARPRO approach. Furthermore, the zero-order phases of components have been tied. Similar constraints have been used for AMARES. Since AMARES does not support constraints on the overall frequency shift, the individual components have been constrained to $\pm.625$ ppm. In addition, the amplitudes of the two citrate peaks have been tied. For QUEST, three metabolite templates have been constructed by simulating noise-free Lorentzian lines according to Tab. 2.

- - - - - - - - - - Tab. 2 about here - - - - - - - - - -

Spectral Patterns. For the PR approach, zerofilling yielded an interpolated spectrum at 1024 frequencies. Automatic zero-order phase correction was performed based on the first recorded data point. From both magnitude and real spectra, 40 values at equidistant frequencies between 3.34 ppm and 2.36 ppm have been calculated by linear interpolation to account for differences in the imaging frequency and the bandwidth. Finally the spectral patterns have been L_1-normalized, i.e. each channel was divided by the sum of the absolute values over all channels. Fig. 2 shows robust statistics of the extracted spectral magnitude patterns as obtained from the evaluable spectra in the prostate data set. The general tumor pattern of elevated Cho + Cr peak (channels 8/14) versus a reduced Ci peak (channel 31) is clearly recognizable.

R3.5
R1.3

- - - - - - - - - - Fig. 2 about here - - - - - - - - - -

Different subspace methods have been used for dimensionality reduction of the spectral patterns. They are particularly appropriate for prostate MRSI since, ideally, only three metabolites contribute to the spectral shape. In particular, four subspace methods have been considered: principal components analysis (PCA), partial least squares (PLS), independent component analysis (ICA) and nonnegative matrix factorization (NMF) which we briefly describe.

- PCA seeks K uncorrelated latent variables $z_k(x) = \alpha_k^T x$ (factors, score variables) that capture all relevant information of the original predictors x. The loadings α_k are obtained as the directions of maximum variance. PCA is described, for example, in (9) and has successfully been used in (5, 13, 17, 27).

- PLS also seeks uncorrelated factors but additionally considers the given classification task. R1.2
The loadings α_k are determined by maximizing both the variance and the correlation with the class label (9). The determined subspace can thus be expected to better capture the information relevant for classification. PLS has originally been proposed in chemometrics (31) and is therefore designed for spectral data. Its good performance in clinical MRSI has been demonstrated in (17).

- ICA is a subspace method that has been used for MR spectra for example in (27). As opposed to PLS and PCA, ICA not only requires uncorrelated but statistically independent components. After centering, prewhitening and dimensionality reduction, ICA reduces to a search over rotations that minimize the mutual information between the components or equivalently maximize the negentropy (9, p.498). In this study, the FastICA algorithm (10) has been used with the logcosh approximation to negentropy.

- NMF has also recently been proposed for the extraction of spectral components (22). It enforces nonnegative loadings and scores which is a reasonable constraint for magnitude spectra. Here we have used a robust version of the alternating nonnegative least squares algorithm.

An important advantage of linear subspace methods is their amenability to interpretation. The weighting of the spectral channels expressed in the constructed components or loadings can be visualized and helps to understand the decision process of the trained classifier.

2.3 Classification

Linear classifiers model the decision boundary as a hyperplane in the space of the explanatory variables. Several studies on MRS classification have applied linear discriminant analysis (LDA) (5, 13, 29) which models the feature distributions as Gaussians with common covariance. Instead, we opted for *logistic regression* (LR) which can be derived from the same probabilistic model by using conditional likelihood (9, pp.103ff). LR is designed for discriminating classes instead of modeling feature distributions which is appropriate for classification tasks (9, p.105).

Furthermore, two linear classifiers which have explicitly been designed for spectral data were considered. *Generalized PLS* (GPLS) can be used to perform LR and PLS in one step (14). *P-spline signal regression* (PSR) exploits the prior knowledge that neighboring spectral channels are correlated by modeling the coefficient profile as a cubic spline function (15).

Nonlinear classifiers are more powerful than linear classifiers in that nonlinear decision boundaries can be constructed. However, this also leads to "black-box" methods which, in general, are hardly interpretable. Here we consider three nonlinear classifiers: *random forests* (RF) (2), an ensemble method, and *support vector machines* (SVM) and *Gaussian processes* (GP), two kernel methods (26).

In short, the RF classifier learns a collection of a few hundred slightly different decision trees (2). The diversity of the trees is encouraged by using bootstraps of the given sample and by randomly selecting a subset of feature variables considered in each node when growing the decision trees. A new example is classified according to the majority vote of the trees in the forest. Thus, the RF classifier employs ideas common with bagging and boosting (9).

Kernel methods perform an implicit mapping to a high-dimensional feature space. The constructed linear decision boundary (a hyperplane) in this high-dimensional feature space corresponds to a nonlinear decision boundary in the original feature space. By using a positive definite kernel instead of the usual dot-product, most linear classifiers can be "kernelized" to

yield nonlinear classifiers. In this study two kernel methods have been used, support vector machines (SVM) and Gaussian processes (GP) (26). The least-squares SVM used for MRS classification for example in (5, 13) can be viewed as a kernelized ridge regression and, except for an additional bias term, is identical to the GP method used in this study (7).

2.4 Error Measure

The area under curve (AUC) of the receiver operator characteristic was used to measure classification performance. It is determined as the area under the graph obtained by plotting $sensitivity$ against $1 - specificity$. Since it does not depend on the chosen threshold that determines the tradeoff between the true positive and true negative rates, it is independent of class priors and misclassification costs. It is therefore an appropriate performance measure for comparing binary classifiers. The AUC attains its maximum value of 1 for perfect separation, whereas it is .5 for random predictions.

Cross-validation (CV) has been used to obtain reliable estimates for the AUCs. In using CV, it should be considered that spectra obtained from the same patient are certainly correlated, violating the i.i.d. assumption in CV. We have therefore employed a "leave-one-patient-out" scheme which determines the performance measure (AUC) for every patient with the classifier trained on all other patients.

3 Results

All reasonable combinations of feature extraction methods and classifiers have been evaluated. The tested combinations are listed in Fig. 3 where the methods have been abbreviated as described in the previous section. The employed box-and-whiskers plots (16) are robust summaries of the 24 AUC values obtained from leave-one-patient-out cross-validation. The thick line within the box marks the median value and the box itself is bounded by the two hinges which are versions of the first and third quartiles. The whiskers extend to the most extreme data points which are no more than 1.5 times the interquartile range from the box.

Linear PR methods vs. quantification. Fig. 3a compares quantification approaches based on VARPRO (v), AMARES (a), QUEST (q) and two PR approaches. In addition to various classifiers, results from the conventional metabolite ratio rule (Cho+Cr)/Ci (and Ci/(Cho+Cr+Ci) in the case of VARPRO) are provided. Since, given a particular quantification algorithm, all classifiers performed similarly, not all results are depicted for AMARES and QUEST. It should be noted that only spectra for which at least one of the peaks was found ($a_k > 0$) have been used in the evaluation of AMARES and QUEST. For AMARES this was about 74% and for QUEST 97% of the data set. The performance of the two PR approaches PCA/LR and PLS/LR (PLS and PCA with logistic regression) based on magnitude (m) spectra is comparable with that of QUEST-based quantification approaches.

Linear vs. nonlinear PR methods. Fig. 3b compares linear and nonlinear PR approaches based on magnitude spectra. For comparison, the first compartment repeats the results for QUEST. The second compartment summarizes linear and the third compartment nonlinear PR approaches.

LR (m) shows results with (unregularized) logistic regression based on all 40 spectral channels. Then, results for the five subspace methods PCA, ICA, NMF, PLS and GPLS are given. In our experiments we have used the four most important loadings which, in the case of PCA and ICA, covered about 80% of the variance and seemed sufficient according to a scree plot (not shown). For PSR, a generalized linear model (GLM) with logistic link function and binomial posterior has been used, the same GLM which yields LR. The SVM with linear kernel (SVM-lin) is listed as a linear method since the decision boundary remains a hyperplane in the original feature space.

Finally, results for the nonlinear PR methods are provided. The random forest (RF) classifier has been trained with 500 trees, nodesize 1 and a subset of 13 considered variables in each split. For the SVM as well as for the GP method, the width of the employed radial basis function (RBF) kernel has been estimated from a fraction of the respective training data set (procedure sigest, cf. (11)).

- - - - - - - - - - Fig. 3 about here - - - - - - - - - -

R3.5

Real vs. magnitude spectra. In Fig. 3c, PR methods using magnitude and real spectra are compared. First, results for the subspace methods PCA, ICA and PLS are provided (NMF does not make sense for real spectra), followed by PSR and the linear SVM. The results for RF (m)

The corresponding results for magnitude spectra are repeated for comparison.

The last two compartments show results obtained with nonlinear classifiers. Based on real spectra, results for the SVM and GP classifiers with RBF kernel and the RF classifier are provided. Representative for the nonlinear classifiers applied to magnitude spectra, RF (m) is repeated.

Detailed comparison. Detailed cross-validation results for six of the tested methods are provided in Tab. 3. For each of the 24 patients one minus the AUC of the respective classifier is given when trained on all patients but this one. In the first four columns, results for three nonlinear classifiers (SVM, GP and RF) and LR with PLS-subspace based on magnitude (m) spectra are listed. Then, two quantification approaches based on QUEST (q) follow. Since no training is required for the ratio rule, (Cho+Cr)/Ci (q) just reflects the AUC results when broken down to individual patients. The last row provides mean values for the respective methods.

- - - - - - - - - - Tab. 3 about here - - - - - - - - - -

R3.2

Although the performance differences between the classifiers in Tab. 3 seem to be small, statistical significance of some differences can be established using a Wilcoxon signed rank test. Concerning the question whether nonlinear classifiers can improve results over linear methods, it is observed that based on magnitude spectra, SVM-rbf (m), GP-rbf (m) and RF (m) significantly outperform PLS/LR (m) (p = .0002/.0012/.0006). Also, based on QUEST the SVM-rbf (q) performs significantly better than the ratio rule (Cho+Cr)/Ci (q) (p = .0034). Comparing quantification-based (SVM-rbf (q)) and PR methods (SVM-rbf (m)/GP-rbf (m)), the performance gain could still be considered significant (p = .0269/.0261). The performances of the linear PR approach PLS/LR (m) compared with SVM-rbf (q) and (Cho+Cr)/Ci (q), however, are statistically indistinguishable (p = .1531/.5966).

Interpretation of PR approaches. PLS loadings obtained from the whole prostate dataset are presented in Fig. 4. The first row shows the L_2-normalized loadings along with a typical spectrum (dashed line). The last two rows show statistics (median, hinges and extreme points) of the upper/lower 5% of the training sample, sorted according to their PLS scores. This reveals spectral patterns which score high/low for the respective PLS loading and facilitates their interpretation.

- - - - - - - - - - Fig. 4 about here - - - - - - - - - -

Fig. 5 contrasts coefficient profiles obtained from three linear classifiers trained on the whole data set. Fig. 5a shows the coefficients obtained with unregularized LR on all 40 channels, Fig. 5b with LR on PLS scores and Fig. 5c shows the result for PSR.

- - - - - - - - - - Fig. 5 about here - - - - - - - - - -

Fig. 6 shows the color-coded probability map obtained from PLS/LR. Next to it, results from a histologic step-section examination are shown. It should be noted that the slice planes obtained from histologic examinations and MRSI are unlikely to coincide exactly. Also, since the histologic samples easily deform after radical prostatectomy, only qualitative comparisons are possible.

- - - - - - - - - - Fig. 6 about here - - - - - - - - - -

4 Discussion

4.1 Spectral Preprocessing

For PR, two spectral representations have been employed in this paper, namely real and magnitude spectra. As opposed to real spectra, magnitude spectra are invariant w.r.t. zero-order phase shifts. The additional variation in the spectral pattern caused by phasing problems R3.5

mainly degrades the performance of PCA/LR and ICA/LR (Fig. 3c). Although the difference to magnitude-based methods is smaller for other linear and nonlinear classifiers, magnitude spectra consistently yield better results. Improvements with real spectra might be obtainable when using more sophisticated automatic phasing algorithms, however, these might also be prone to similar robustness problems as quantification algorithms. We conclude that the advantage obtained from omitting phase correction in magnitude spectra outweighs the disadvantage of increased line widths and peak overlap for prostate MRSI. Other studies present analogous results for brain MRS (5, 13, 17, 29). R1.3

It has also previously been found that some kind of normalization (L_1, L_2, L_∞) of the spectral patterns is important (17). Experiments with L_1- and L_2-normalized prostate spectra did not yield very different results (not shown here). In contrast to (5, 13, 29) we have used L_1-normalized spectra because of the notable relationship to metabolite ratios. The L_1-norm can be regarded as an approximation to the integrated spectrum and corresponds to Cho+Cr+Ci in the prostate. Hence, linear combinations of the derived spectral features are similar to the ratio $r_2 = \mathrm{Ci}/(\mathrm{Cho}+\mathrm{Cr}+\mathrm{Ci})$ which is related to the usual ratio $r_1 = (\mathrm{Cho}+\mathrm{Cr})/\mathrm{Ci}$ by the monotonous transformation $r_2 = (r_1 + 1)^{-1}$. Therefore, r_1 and r_2 must have the same discriminating power which is confirmed by the results in Fig. 3a. Hence, L_1-normalization addresses the problem that absolute line intensities in MR spectra are unreliable and information is only contained in their ratio.

4.2 Quantification-based approaches

Despite only subtle mathematical differences in performing quantification with VARPRO, AMARES or QUEST (with simulated metabolite templates), the classification results differ considerably (Fig. 3a). All quantification methods have been employed with the same number of Lorentzian shaped components but with slightly different constraints. Hence, the employed prior knowledge has considerable influence on the obtainable classification performance.

Implementation details of the employed algorithm also seem to matter. The superior performance of our VARPRO approach in comparison to AMARES might be surprising at first. However, deviating from the original VARPRO approach, which uses a modified Levenberg-Marquardt algorithm (30), we have used an interior trust region algorithm (4) that appears to

cope better with the variable projection functional. The excellent performance of QUEST, on the other hand, might be due to its implicit baseline correction (21).

Most of the differences between quantification-based approaches are due to choosing different quantification methods and not due to using different classifiers (Fig. 3a). However, none of the classifiers employed on quantified data could improve over the results obtained with the conventional (Cho + Cr)/Ci ratio. This indicates that the ratio rule is indeed a good approach for the discrimination of tissue classes in the prostate, provided that the quantification results are reliable. However, the latter is difficult to judge in the absence of ground truth.

4.3 Subspace Methods

The results listed in the second compartment of Fig. 3b show no significant difference between the tested subspace methods. In particular, identical performance is obtained with PCA and ICA. Given that the scores obtained from FastICA are necessarily a linear combination (scaling and rotation) of the PCA scores, this can be explained by noting that LR is invariant w.r.t. such feature transformations. But also NMF cannot improve the AUC. And although PLS and GPLS can increase the lower hinge in the discrimination of healthy and tumor tissue, these effects are not observed in the discrimination against voxels of the "undecided" class.

One reason for the use of subspace methods is that the basis of the constructed subspace is amenable to interpretation. Optimal subspaces along with the most important spectral patterns are automatically determined based on *in vivo* data. Therefore, not only protocol and metabolite dependent features of the signal but also the *in vivo* situation is considered. Furthermore, for PLS also the classification task at hand has an influence on the choice of the subspace. This distinguishes subspace from quantification approaches which either use theoretical models or metabolite templates derived from *in vitro* measurements.

The PLS loadings derived for the prostate data allow for a consistent interpretation (cf. Fig. 4). As published in various clinical studies on prostate spectroscopy (e.g. (19, 24, 25)), the ratio between Cho + Cr and Ci is the most important feature in discriminating cancerous from healthy tissue. Together with the L_1-normalization, the first loading clearly reflects this ratio. The second loading rewards high Cho-to-Cr ratios in the presence of a clear Ci peak. Spectra with small line widths and clear peaks for all metabolites get a high score on this component if Cho is

elevated in comparison to Cr. This criterion is in accordance with medical studies, for example (19), where the Cho/Cr ratio has also been considered. The third loading reflects frequency shifts of the citrate peak and the fourth loading frequency shifts of the choline peak. Both these and higher order loadings are not relevant for tumor classification.

4.4 Linear and Nonlinear Classifiers

Fig. 5 demonstrates that subspace methods act as a regularizer which helps to overcome the problem of collinearities in spectral data. The unregularized model as applied to the highly correlated spectral channels yields a very rough coefficient profile with high offset (Fig. 5a). As opposed to that, the coefficient profile of the PLS model in Fig. 5b takes small values around zero and shows a clear pattern resulting from a linear combination only of the first four loadings. A similar profile is obtained for the logistic PSR model in Fig. 5c for which the coefficient profile has explicitly been modeled as a smooth spline function. Anyhow, the regularizing influence is not reflected in a clear performance gain (Fig. 3).

Increased performance is obtained when using nonlinear PR approaches. A SVM with linear kernel is a linear classifier and performs no better than its cognates. As evidenced by Fig. 3b, an improvement is obtained only when switching to the nonlinear RBF kernel. A significant improvement from using nonlinear classifiers can also be observed in Tab. 3. The performance of the RF and GP methods are very similar, indicating that some nonlinearity is indeed present in the prostate tumor classification task. However, an interpretation of the decision rules of a nonlinear classifier remains difficult. Significant differences between the nonlinear classifiers could not be observed.

Nonlinear classifiers can manifest their superiority only when applied to spectral patterns. In view of Tab. 3 and Fig. 3, if enough data is available and a nonlinear "black-box" method is acceptable, there remains little reason to use quantification for feature extraction. R1.1 R3.4

Finally, diagnostic maps such as shown in Fig. 6 allow for a time-efficient evaluation of NMR spectroscopic images. In this example, the spectral patterns obtained with MRSI and the histopathological ground truth agree very well. Further clinical evaluation is certainly required to confirm this correlation.

5 Conclusions

In this study, we have compared different approaches to the automated estimation of tumor probability in ^{1}H NMR spectroscopic images of the prostate. The emphasis has been put on developing a fully automatic and reliable approach with optimal diagnostic results.

In particular, we have found that quantification-based approaches heavily rely on an optimal choice of prior knowledge and on the algorithm used for quantification. In contrast, PR approaches do not require specific prior knowledge and can infer important spectral patterns from the *in vivo* training data automatically. The PR approach attempts to address the diagnostic question – healthy vs. tumorous tissue – directly and can therefore use the full statistical information contained in the raw spectral data.

Among the quantification-based approaches, best results have been obtained with classifiers based on metabolite concentrations estimated with QUEST. However, the performance was not superior to the conceptually simple linear PR approaches based on magnitude spectra.

Several subspace methods proposed for spectral MR data before have been compared. In particular, we have used PCA, ICA, NMF and PLS. Hardly any difference in performance could be observed between these. Still, methods especially designed for spectral data such as PLS and GPSR seemed to have slight advantages.

If a "black-box" approach is acceptable, superior performance can be obtained by using a suitable nonlinear classifier in conjunction with magnitude spectra.

6 Acknowledgements

The authors would like to extend their thanks to Lutz Trojan for taking care of the patients; to Rainer Grobholz for providing the results of the histologic step-section examinations; to Peter Bachert for kindly sharing his invaluable experience in MRSI; and to three anonymous reviewers for their comments that have helped to improve the presentation. Finally, the authors would like to acknowledge financial support by the Deutsche Forschungsgemeinschaft (grant DFG-HA-4364).

References

[1] INTERPRET project (IST-1999-10310) funded by the European Comission. http://carbon.uab.es/INTERPRET/.

[2] Breiman L. Random forests. Mach Learning 2001;45(1):5–32.

[3] Coackley FV, Teh HS, Qayyum A, Swanson MG, Lu Y, Roach III M, Pickett B, Shinohara K, Vigneron DB, Kurhanewicz J. Endorectal MR imaging and MR spectroscopic imaging for locally recurrent prostate cancer after external beam radiation therapy: Preliminary experience. Radiology 2004;233:441–448.

[4] Coleman T, Li Y. An interior trust region approach for nonlinear minimization subject to bounds. SIAM J Opt 1996;6:418–445.

[5] Devos A, Lukas L, Suykens JAK, Vanhamme L, Tate A, Howe F, Majós C, Moreno-Torres A, van der Graaf M, Arús C, Van Huffel S. Classification of brain tumours using short echo time ^{1}H MR spectra. J Magn Reson 2004;170(1):164–175.

[6] Devos A, Simonetti A, van der Graaf M, Lukas L, Suykens JAK, Vanhamme L, Buydens LMC, Heerschap A, Van Huffel S. The use of multivariate MR imaging intensities versus metabolic data from MR spectroscopic imaging for brain tumour classification. J Magn Reson 2005;173(2):218–228.

[7] Gestel TV, Suykens JAK, Lanckriet G, Lambrechts A, Moor BD, Vandewalle J. Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. Neural Comput 2002;14(5):1115–47.

[8] Hagberg G. From magnetic resonance spectroscopy to classification of tumors. A review of pattern recognition methods. NMR Biomed 1998;11:148–156.

[9] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, 2001.

[10] Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw 2000;13(4-5):411–30.

[11] Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab - An S4 package for kernel methods in R. Technical Report 9, Dept. Stat. Math., Wien, 2004.

[12] Laudadio T, Pels P, Lathauwer LD, Hecke PV, Huffel SV. Tissue segmentation and classification of MRSI data using canonical correlation analysis. Magn Reson Med 2005; 54(6):1519–1529.

[13] Lukas L, Devos A, Suykens JAK, Vanhamme L, Howe F, Majós C, Moreno-Torres A, der Graaf MV, Tate A, Arús C, Van Huffel S. Brain tumor classification based on long echo proton MRS signals. Artif Intell Med 2004;31(1):73–89.

[14] Marx BD. Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics 1996;38(4):374–381.

[15] Marx BD, Eilers PHC. Generalized linear regression for sampled signals or curves: A P-spline approach. Technometrics 1999;41(1):1–13.

[16] McGill R, Tukey JW, Larsen WA. Variations of box plots. Am Stat 1978;32(1):12–16.

[17] Menze BH, Lichy MP, Bachert P, Kelm BM, Schlemmer HP, Hamprecht FA. Optimal classification of long echo time in vivo magnetic resonance spectra in the detection of recurrent brain tumors. NMR Biomed 2006;Published Online:26 Apr 2006.

[18] Naressi A, Couturier C, Castang I, de Beer R, Graveron-Demilly D. Java-based graphical user interface for MRUI, a software package for quantitation of in vivo/medical magnetic resonance spectroscopy signals. Comput Bio Med 2001;31:269–86.

[19] Noworolski SM, Henry RG, Vigneron DB, Kurhanewicz J. Dynamic contrast-enhanced MRI in normal and abnormal prostate tissue as defined by biopsy, MRI, and 3D MRSI. Magn Reson Med 2005;53:249–255.

[20] Pijnappel WWF, van den Boogart A, de Beer R, van Ormondt D. SVD-based quantification of magnetic resonance signals. J Magn Reson 1992;97:122–134.

[21] Ratiney H, Sdika M, Coenradie Y, Cavassila S, van Ormondt D, Graveron-Demilly D. Time-domain semi-parametric estimation based on a metabolite basis set. NMR Biomed 2005; 18(1):1–13.

[22] Sajda P, Du S, Brown TR, Stoyanova R, Shungu DC, Mao X, Parra LC. Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. IEEE Trans Med Imaging 2004;23(12):1453–65.

[23] Scheenen T, Klomp D, Röll S, Fütterer J, Barentsz J, Heerschap A. Fast acquisition-weighted three-dimensional proton MR spectroscopic imaging of the human prostate. Magn Reson Med 2004;52(1):80–8.

[24] Scheenen T, Weiland E, Fütterer J, van Hecke P, Bachert P, Villeirs G, Lu J, Lichy M, Holshouser B, Roell S, Barentsz J, Heerschap A. Preliminary results of IMAPS: An international multi-centre assessment of prostate MR spectroscpoy. In Proc Intl Soc Mag Reson Med, 13. Springer, 2005; p. 260.

[25] Scheidler J, Hricak H, Vigneron DB, Yu KK, Sokolov DL, Huang LR, Zaloudek CJ, Nelson SJ, Carroll PR, Kurhanewicz J. Prostate cancer: Localization with three-dimensional proton MR spectrosopic imaging – clinicopathologic study. Radiology 1999;213:473–480.

[26] Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA, 2001.

[27] Simonetti AW, Melssen WJ, de Edelenyi FS, van Asten JJA, Heerschap A, Buydens LMC. Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification. NMR Biomed 2005;18(1):34–43.

[28] Swindle P, McCredie S, Russell P, Himmelreich U, Khadra M, Lean C, Mountford C. Pathologic characterization of human prostate tissue with proton MR spectroscopy. Radiology 2003;228(1):144–151.

[29] Tate AR, Majós C, Moreno A, Howe FA, Griffiths JR, Arús C. Automated classification of short echo time in in vivo ^{1}H brain tumor spectra: a multicenter study. Magn Reson Med 2003;49(1):29–36.

[30] Vanhamme L, van den Boogaart A, Van Huffel S. Improved method for accurate and efficient quantification of MRS data with use of prior knowledge. J Magn Reson 1997; 129(1):35–43.

[31] Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comput 1984;5(3):735–743.

[32] Zakian KL, Eberhardt S, Hricak H, Shukla-Dave A, Kleinman S, Muruganandham M, Sircar K, Kattan MW, Reuter VE, Scardino PT, Koutcher JA. Transition zone prostate cancer: Metabolic characteristics at ^{1}H MR spectroscopic imaging – initial results. Radiology 2003; 229(1):241–151.

[33] Zakian KL, Sircar K, Hricak H, Chen HN, Shukla-Dave A, Eberhardt S, Muruganandham M, Ebora L, Kattan MW, Reuter VE, Scardino PT, Koutcher JA. Correlation of proton MR spectroscopic imaging with gleason score based on step-section pathologic analysis after radical prostatectomy. Radiology 2005;234:804–814.

Tables

Table 1: Distribution of labels in the prostate data set (76 slices from 24 patients).

quality \ class	healthy	undecided	tumor	all
not evaluable	–	–	–	15268
poor	721	437	284	1442
good	1665	629	452	2746
all	2386	1066	736	19456

Table 2: FID components used for quantifying prostate MRSI. The parameters have been initialized with the given value and constrained to the range given in brackets.

Metabolite	Model	Frequency [ppm]	Line Width [Hz]
Choline	Lorentzian	3.22 [±.03]	6.25 $[0, 31.25]$
Creatine	Lorentzian	3.04 [±.03]	6.25 $[0, 31.25]$
Citrate-1	Lorentzian	2.65 [±.03]	6.25 $[0, 31.25]$
Citrate-2	Lorentzian	2.60 [±.03]	6.25 $[0, 31.25]$

Table 3: Cross-validation results for a selection of the tested methods. The given test error values are one minus the AUC of the receiver operator characteristic when training is performed on all but the tested patient, i.e. better performance is indicated by smaller values.

	Pattern Recognition				**Quantification**	
patient	**SVM-rbf (m)**	**GP-rbf (m)**	**RF (m)**	**PLS/LR (m)**	**SVM-rbf (q)**	**(Cho+Cr)/Ci (q)**
1	5.00e-04	2.00e-04	2.00e-04	6.00e-04	2.70e-03	9.60e-03
2	0	0	0	0	0	0
3	0	0	0	2.49e-02	0	0
4	0	0	0	0	0	1.50e-03
5	0	0	0	0	1.00e-04	1.30e-03
6	0	0	0	0	0	0
7	1.90e-03	1.00e-03	2.00e-03	1.00e-02	6.00e-03	4.90e-03
8	1.70e-03	1.70e-03	0	1.70e-03	7.10e-03	1.78e-02
9	0	0	0	0	0	0
10	0	0	0	2.00e-04	0	2.60e-03
11	0	0	8.00e-04	2.30e-03	3.00e-04	1.40e-03
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	3.20e-03	0	0
15	0	1.90e-03	5.00e-04	2.00e-04	3.31e-02	3.56e-02
16	0	0	1.44e-02	8.85e-02	2.87e-02	4.07e-02
17	3.46e-02	5.63e-02	6.49e-02	6.49e-02	0	0
18	0	0	0	0	0	0
19	0	0	0	3.90e-03	3.90e-03	6.00e-03
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	3.00e-04	4.30e-03	4.27e-02	0	2.00e-03
23	0	0	0	0	2.70e-03	0
24	0	0	0	0	0	8.30e-03
mean	**1.60e-03**	**2.60e-03**	**3.60e-03**	**1.01e-02**	**3.50e-03**	**5.50e-03**

Figure Captions

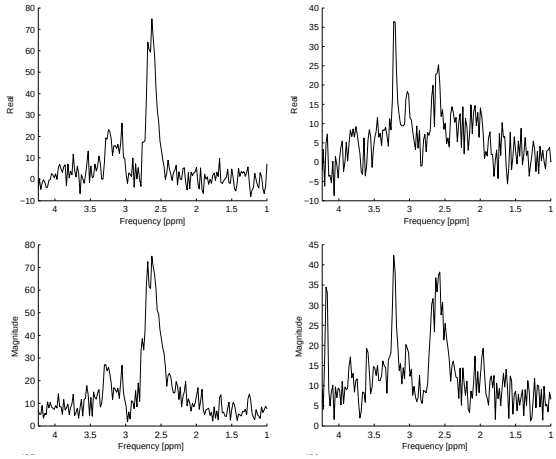


Figure 1: Two example spectra (left: healthy, right: tumor) after HSVD water/lipid removal and zerofilling to 1024 datapoints. The top row shows manually phased real absorption spectra whereas the bottom row shows the corresponding magnitude spectra. A slight increase in line width can be observed when switching from real to magnitude spectra.

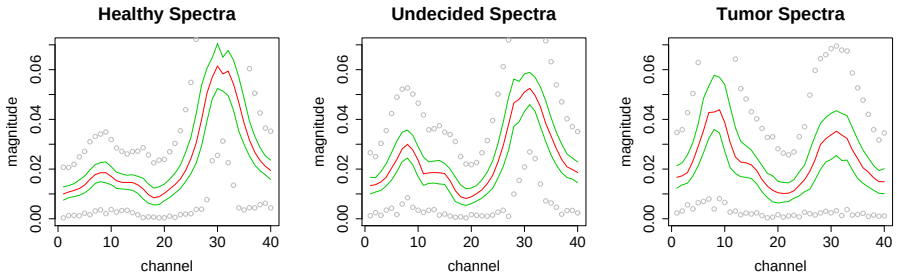
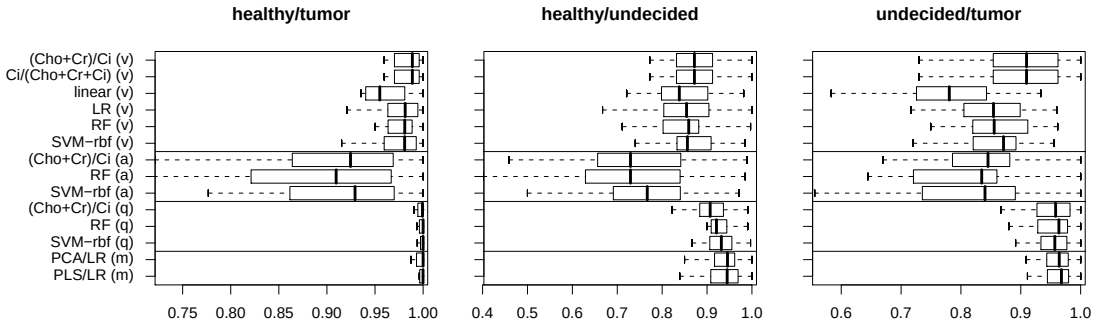
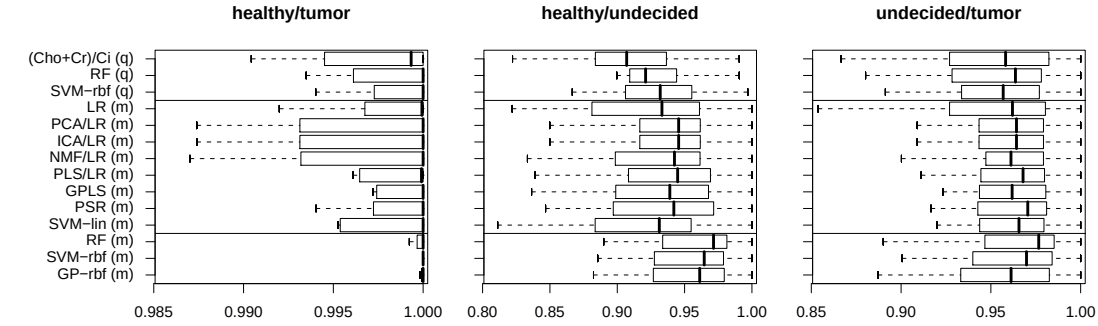


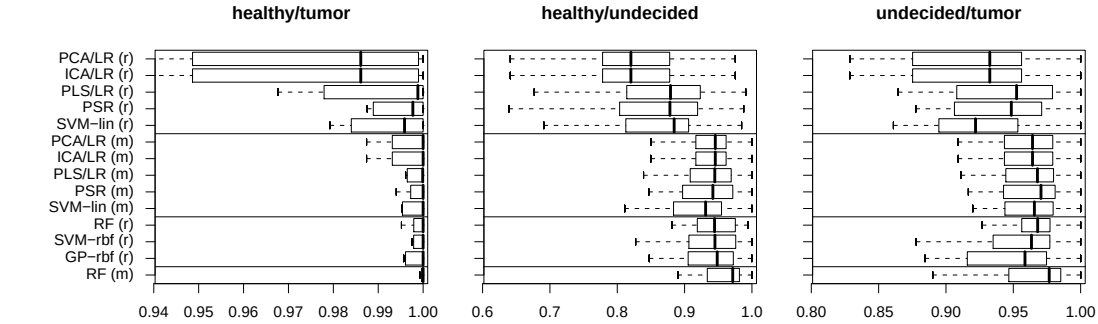
Figure 2: Spectral patterns in the prostate data (3.34-2.36 ppm). From left to right typical patterns of healthy, undecided and tumor tissue can be recognized with their characteristic choline (channel 8), creatine (channel 14) and citrate (channel 31) ratios. In the spirit of a box-and-whiskers plot (16), the median (red), the hinges (green) and extreme points (circles) are shown for each channel.



(a) Linear PR versus quantification-based approaches.



(b) Linear versus nonlinear PR approaches.



(c) Real versus magnitude spectra.

Figure 3: Comparison of various approaches: (v)-VARPRO, (a)-AMARES, (q)-QUEST quantification-based approaches versus PR approaches based on (m)agnitude and (r)eal spectra. The box-and-whiskers plots show the median, the hinges and the extreme points of the area under curve (AUC) values of the receiver operator characteristic obtained from leave-one-patient-out cross-validation. Linear PR approaches combining a subspace method X with logistic regression (X/LR) easily achieve the same performance as the best quantification approaches (i.e. QUEST). Even slightly better results are obtained with nonlinear PR approaches (RF, SVM, GP) applied to raw magnitude spectra (m). Details of the different methods are described in the text. Note that the individual plot scales differ.

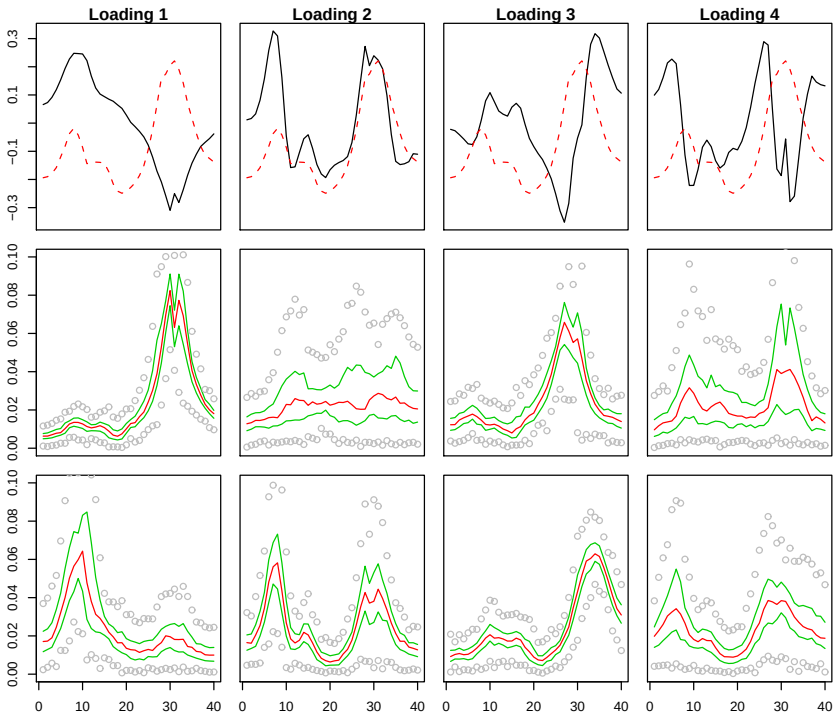
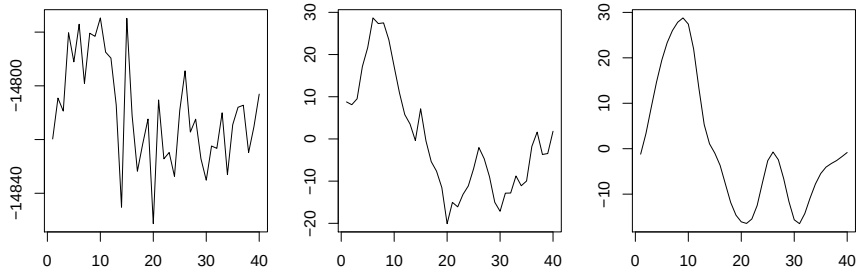


Figure 4: First row: first four PLS loadings (the dashed lines sketch a prototypical spectrum with the three relevant peaks of Cho, Cr and Ci). Last two rows: median (red), hinges (green) and extreme points (circles) of the 5% of the training sample which score highest/lowest for the respective PLS loading.



(a) Logistic Regression (b) LR with PLS (c) Logistic PSR

Figure 5: Comparison of coefficient profiles learned with logistic regression models. The unregularized model (a) does not show a pattern whereas the PSR model (c) seems to oversmooth slightly. In contrast, the PLS model (b) shows a clear pattern and also preserves the details.

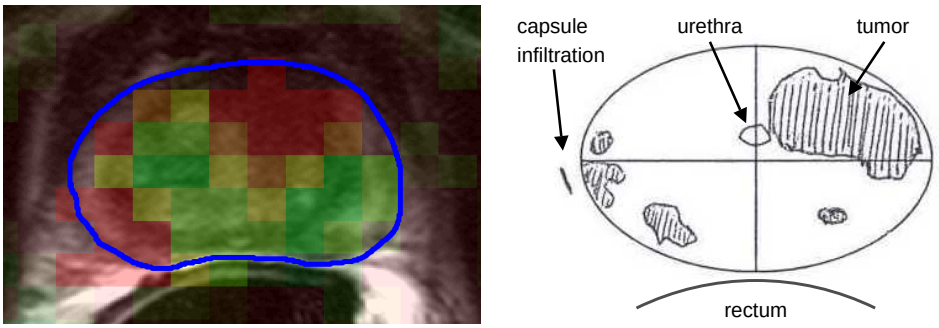


Figure 6: Tumor probability map estimated with logistic regression based on partial least squares scores (PLS/LR) and histologic step-section result for the same slice. Up to minor deformations, the evaluated *in vivo* MRSI agrees very well with the histopathology.