Original article

How much anatomy do we need? Automated *vs*. manual pattern recognition of 3D ^{1}H MRSI data of patients with prostate cancer

Christian M. Zechmann[1*], Bjoern H. Menze[2*], Michael B. Kelm[2], Patrik Zamecnik[1], Uwe Ikinger[3], Rüdiger Waldherr[4], Frederik L. Giesel[1], Christian Thieke[5], Stefan Delorme[1], Fred A. Hamprecht[2], Peter Bachert[6]

[1] German Cancer Research Center (DKFZ), Department of Radiology, Heidelberg, Germany
[2] Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Heidelberg, Germany
[3] Urology Department, Salem Hospital, Heidelberg, Germany
[4] Pathology Institute Prof. Waldherr, Heidelberg, Germany
[5] German Cancer Research Center (DKFZ), Clinical Cooperation Unit for Radiation Therapy, Heidelberg, Germany
[6] German Cancer Research Center (DKFZ), Department of Medical Physics in Radiology, Heidelberg, Germany

*shared first authorship

Corresponding author: Christian M. Zechmann
Department of Radiology (E010)
German Cancer Research Center (DKFZ)
Im Neuenheimer Feld 280
D–69120 Heidelberg
phone:+49 6221 422525
fax: +49 6221 422531
e–mail: c.zechmann@dkfz.de

Key words: prostate cancer; proton MR spectroscopic imaging; postprocessing; pattern recognition

Total word count:

Abstract

Objective

To evaluate quantitatively 3D proton MR spectroscopic imaging (3D ^{1}H MRSI) data of the prostate in patients with known prostate cancer using anatomical knowledge, compared with a single–voxel–spectra evaluation by blinded experts and automated processing methods.

Material and Methods:

MRSI data of 10 patients with histologically proven prostate adenocarcinoma, scheduled either for prostatectomy or intensity–modulated radiation therapy, were evaluated by two MRS experts using information on anatomy and localization of the spectra; and in blinded, randomized lists. Spectra were also classified by automated processing methods based on spectral fitting and pattern recognition. Results were compared using Kendall's *tau* and grouped in a hierarchical segmentation.

Results:

The experts came to more binary decisions – using information of spectra from surrounding tissue in ambiguous cases – when evaluating MR spectroscopic images as a whole. Differences between unblinded and blinded evaluation were larger than differences between blinded and automated processing methods.

Conclusion:

An automated approach can be as good as a blinded reader. However, anatomical and morphological information are routinely used by human experts, supporting their final decisions. Using automated approaches considering anatomical knowledge could therefore improve an automated evaluation, supporting the human reader.

193 words

Key Words: prostate cancer; proton MR spectroscopic imaging; postprocessing; pattern recognition

Introduction

Prostate cancer is still the most frequent malignancy in the western hemisphere with 29% of all newly diagnosed tumors. The recorded incidence increased mainly due to prostate–specific antigen (PSA) screening. Following this a decrease in the death rate from 1990 to 2003 of 31.12% is seen [JeSW07], which is related to many improvements in diagnosis and treatment. Nevertheless prostate cancer still causes the second leading cancer–related mortality in the Western hemisphere (9% in the U.S.).

While today the detected cancers are smaller, at a lower stage, and a lower grade, a wide range of aggressiveness remains. Therefore the identification of patients who will develop a highly infiltrating and metastasizing tumor is crucial [HrCE07].

Adenocarcinoma can be suspected when the level of serum PSA is rising over time. Digital rectal examination performed by experienced urologists is commonly used for non–invasive diagnosis of prostate carcinoma. For conclusive evidence of a tumor lesion the histopathologic grading (Gleason score) determined from a biopsy specimen is still required. But often these techniques fail to localize the suspected tumor.

The role of imaging techniques for diagnosis and treatment planning depends on the modality. Transrectal ultrasound (TRUS) is mainly used for image–guided biopsies, whereas computed X-ray tomography (CT) is performed for the assessment of lymph nodes or distant metastases in bones and other organs [HrCE07]. Magnetic resonance imaging (MRI) permits visualization of the zonal anatomy of the prostate and the tumor itself. The tumor can be differentiated as localized (stage T1–T2), infiltrative beyond the capsule into periprostatic fat, lymph nodes or seminal vesicles (stage T3), or into the surrounding organs (stage T4). High–resolution T_2–weighted MRI performed with a pelvic array coil, provides good specificity (up to 90% are reported), but low sensitivity (27–61%) for tumor detection [ScHV99], [HrWV94]. With use of an endorectal coil the sensitivity for the detection of tumor foci larger than 1 cm in diameter on T_2–weighted MRI is improved up to 85.3%, while for smaller tumors a sensitivity of only 26.2% was reported [NaTI04]. Critical false–positive

results mainly arise from factors that also lead to a signal reduction in T_2–weighted MRI such as biopsy, hemorrhage, prostatitis, and therapeutic effects [PeKJ96].
Therefore functional imaging techniques such as dynamic–contrast–enhanced MRI (DCE–MRI) and proton MR spectroscopic imaging (^{1}H MRSI) were included into MRI studies to increase the sensitivity for detection of prostate tumors. MRSI permits non–invasive detection of small metabolites *in vivo* in the prostate such as citrate (Ci) and free choline and choline–containing compounds (Cho) and thus offers a certain improvement in sensitivity and specificity for the diagnosis of prostate cancer [BaML96], [QuFD94], [DASW98]. Since the concentrations of these compounds change characteristically in pathologic tissue, MRSI is in principle well–suited for the detection and localization of prostatic tumors [RePR07], [ScHV99], [ScYT92].

The application of MRSI, however, comes along with a tremendous amount of data that must be post–processed and evaluated. In addition, the subsequent signal processing demands for a high degree of user interaction, requiring an experienced MR spectroscopist. Both factors aggravate the introduction of MRSI into clinical routine. If this workload can be minimized by a reliable and robust post–processing software of MRSI data this could have a major impact on the diagnosis of prostate cancer.

Different approaches exist to facilitate the analysis of the data and to support the human reader. Most popular is the automated fitting of mathematical models to the spectral data and a diagnostic evaluation using the estimated model parameters. Alternatively – or in addition – the spectral pattern can be evaluated by multivariate classification methods, directly mimicking the visual inspection of a spectrum [KeMZ07]. While both approaches are able to automatically evaluate a large amount of single spectra, they are not specifically adapted to the anatomical information inherent to a spectroscopic image.
Trained spectroscopists, however, have supportive information concerning anatomical margins like the prostate capsule, on suspicious spectra from surrounding tissue or T_2–hypointense areas that indicate tumor. This visual inspection of MRSI data together with supportive anatomical information can be considered as the gold standard in the evaluation of MRSI of the human prostate.

The purpose of our study was therefore to compare such an evaluation using this information about anatomy and localization of a spectrum, with the outcome of an “anatomically blinded” evaluation as performed by the algorithms used in the automated processing. To this end, MRSI data sets of ten patients with histologically proven prostate cancer were analysed by two experienced MR spectroscopists using knowledge about the localization of each single spectrum and localization of the prostate. Then spectra were evaluated separately, by the same experts in a visual inspection and by different automated approaches, i.e. fitting in the time domain, fitting in the frequency domain, and machine–based classification of the spectral pattern. The respective result maps were compared.

Material and Methods

Data

Patients

Ten patients with biopsy–proven prostatic cancer who were scheduled either for prostatectomy ($n = 5$) or intensity–modulated radiation therapy (IMRT; $n = 5$) were included in this study. Before the examination written informed consent was obtained from each patient. None of the patients had prior therapy like antihormonal treatment or radiation therapy.

Spectroscopic imaging

At $B_0 = 1.5$ T (Magnetom Symphony; Siemens Medical Solutions, Erlangen, Germany) T_2–weighted MRI (TR = 4000 ms, TE = 24 ms, FOV = 140 x 140 mm^2, matrix 0.7 x 0.5 x 4 [mm]) was performed for diagnosis, planning purposes, and localization of the VOI for spectroscopic imaging.

Water– and lipid–signal–suppressed 3D MRSI (PRESS [point–resolved spectroscopy] sequence with measurement parameters: TR = 650 ms, TE = 120 ms, nominal voxel size 6 × 6 × 6 mm^3, matrix 16 × 16 × 16, total acquisition time 10–12 min [ScKR04]) yielded 4096 ^{1}H MR spectra per examination. Owing to the elliptical shape of the prostate and the position of surrounding signal saturation bands usually about 900 spectra (22%) were localized within the prostate. All patients had MRSI data of good quality without artifacts or poor signal–to–noise ratio (SNR) except two patients (data c and h, Figure 2).

Preparation of spectroscopic imaging data

The water peak and lipid resonances outside the spectral range of 0–5 ppm were removed before further analysis using Hankel–Lanczos–singular–value decomposition (HLSVD) [BeBO92], [PiBO92]. After Fourier transformation, frequency spectra were phased automatically for display of absorption lines.

Evaluation procedures

First spectra were evaluated visually by two MR spectroscopy (MRS) experts (a), employing automatic line fitting *and* anatomical information, defining the "gold standard"; in a second step (b), after random permutation, data were analysed by the experts without knowledge of the anatomical context, defining a 'blinded' or single–voxel gold standard for comparison with the automated procedures, which used fitting of line–shape models (c) and pattern recognition (d).

(a) Visual–plus–anatomical evaluation

MRSI data was analyzed by both readers in consensus using software provided by the manufacturer (MetaboliteMapper; Siemens Medical Solutions, Erlangen, Germany). All spectra were phased for display of absorption lines and labeled according to a five–point scale with 1 (= "tumor"), 2 (= "possibly tumor"), 3 (= "undecided"), 4 (= "possibly no tumor") to 5 (= "no tumor") and the "reject" class. The contrast of the morphological T_2–weighted images was reduced in order to eliminate bias by T_2–hypointense areas indicating tumor (Figure 1) without compromising visibility of the outer margins of the prostate. This procedure was performed voxel–by–voxel for all slices covering the prostate (usually 10–12 out of 16). Each spectrum was analysed by two radiologists (C.Z., P.Z) with 4–years and 1.5–years experience in MRS, respectively, and labels were assigned *in consensus* (referred to as *anatomical* evaluation 'an'). First, data quality was assessed. Poor SNR or artifacts resulted in an assignment of the spectra to the "reject" class. Relative signal intensities of cholines, creatine (Cr), and citrate as well as results from spectral fitting were examined (without using the "CC/C" value = [Cho+Cr]/Ci). Finally, spectra were classified according to the five–point scale. Spectra localized outside the prostate were identified and excluded from further processing.

(b) Randomized visual evaluation

The order of the spectra was first randomized over all patients, then the spectra were phase– and baseline–corrected and finally displayed without the corresponding T_2–weighted MR images. Additional post–processing, like calculation of CC/C ratios was not possible. The displayed spectra were evaluated again by the same MRS experts and assigned to the five classes and the 'reject' class as described above. This time the evaluation was performed separately for all spectra by each reader (referred to as *random* evaluation by expert 'e1' and 'e2') according to the method described above.

(c) Automated spectral–fit–based evaluation

Three different algorithms were used for fitting of resonance lines (referred to as *spectral fitting*): AMARES (as implemented in jMRUI, [NaCD01]) for fit in the time domain (referred to as 'ft') and two algorithms for fit in the frequency domain: QUEST (referred to as 'f1'), implemented in jMRUI [NaCD01], and CSITOOLS [SiTools ?] (referred to as 'f2') described in [KeMZ07]. Spectra were evaluated in terms of the CC/C value. Results of spectra with inadequate linewidths [Kreis04] or spectra with an estimated Cramér–Rao bound (CRB) of more than 30% of the amplitude in the time domain were rejected.

(d) Automated pattern–recognition–based evaluation.

A software for *automated pattern recognition* applicable to in vivo MRSI data (referred to as 'pr', Figure 1) [KeMZ07], developed at the Interdisciplinary Center for Scientific Computing (IWR) at the University of Heidelberg, was installed at the German Cancer Research Center (DFKZ). The MRSI data and the corresponding T_2–weighted MR images were read from the DICOM data set and evaluated in a single process using the CLARET software [KeMN06]. Spectra with artifacts or poor SNR were rejected using a nonlinear classification approach (NoN–Score [MeKW08]). Magnitude spectra were classified by a logistic linear model, trained on 30 MRSI data sets in an earlier study [KeMZ07], [KeMN06] and ranked according to the five–point scale mentioned above.

Statistical evaluation

For comparison of the different MRSI data evaluation strategies, we pursued the following procedure: We transformed all results from all evaluations (a–d) to the same score, *i.e*., to ranked labels between 1 and 5, and determined that subset of data which could be used for a comparison of *all* methods. A distance metric based on the correlation of their outcomes was chosen to measure the similarity between the different results. Similiar evaluation approaches were grouped in a hierarchical mode. Characteristic and stable groupings were visualized and reasons for (dis–) similarities were sought in the original data. Details on the classification of the

metabolite intensity ratios and the methods used in the comparisons are given in the following.

In both the anatomical and the random evaluation (a, b), the experts assigned values between 1 (definitely healthy) and 5 (definitely tumor) to the spectra. Likewise the automated pattern recognition (d) returned labels between 1 and 5. In contrast, the result of the spectral fitting (c) was a continuous score, the CC/C value [KuVH96]. In order to allow an unbiased comparison among all evaluation approaches, this score was transformed to discrete class labels in a first step. The results of the visual inspections (a, b) were used to find thresholds on the ratio–score, allowing for class assignments that match optimally the experts' decisions. The results of the anatomical and the two random evaluations were averaged and rounded to the nearest integer. Thresholds between two classes were determined using an equal amount of samples from both classes. To this end each pair of neighbouring classes were subsampled to the size of the smaller of the two; the threshold was fixed to the value minimizing the classification error (Figure 4 [THRESHOLDS], Figure 7 [RATIOS]). This procedure was repeated for each threshold 100 times; values were recorded and averaged finally.

A validated principal data set is indispensable in a clinical validation of any diagnostic method. In a technical evaluation of different algorithms, however, already a pairwise comparison of results allows to gain insight into quality and reliability of these methods [BoMa07]. A distance has to be defined which measures dissimilarities between results, while distance matrices can be visualized in low dimensional projections. We extend the method of Ref. [BoMa07] by a subsequent test for stability of the observed (dis–)similiarites.

Distance. To compare the similarity of the results of two evaluation approaches, Kendall's test for correlation was used [Kend48]. *Kendall's tau* measures the degree of concordance between two rankings and estimates the significance of their correspondence. The measure is proportional to the number of concordant pairs between the two ordered lists and thus allows to measure the closeness of agreement in a cross–tabulation (hit matrices) of the classification results. As a rank correlation criterion, it ignores transformations which do not affect the ranking and

which can easily be removed or corrected for by, for example, a rescaling of the score, or by an implicit adaptation of the interpreting expert. While the test seeks for linear relationships between two sets of ranked, ordinal data, it assigns minimal penalties to monotonous shifts (*e.g.*, samples are systematically assigned to the higher class), and maximally penalizes extreme deviations from the data (*e.g.*, a class–5 sample which is assigned to class 1). A value of 0 indicates independence (and hence a complete random assignment of labels in the given task), while a value of 1 indicates a perfect correlation between the two distributions.

Visualization. The comparison of all seven different methods including the average 'ea' of 'e1' and 'e2' leads to an 8×8 matrix with 28 off–diagonal elements (correlation coefficients) (Table 2 DISSIMILARITIES). Since the correlation coefficients of the *tau*–test define a metric, distance–based methods are useful to summarize and interpret the results. While the distance matrix of the comparison of all evaluation methods spans a space of up to six dimensions, multidimensional scaling enables the most accurate (in terms of the least–squares error) projection of this space to lower dimensions. (Dis–)similarities between different evaluation strategies can be visualized as distance in a, for example, two–dimensional plane. To visualize effects/interactions which cannot be projected into a two–dimensional subspace, a hierarchical splitting or segmentation approach can be used in addition.

Here, a model–free hierarchical cluster algorithm is used to visualize relations in the full space. Since we were interested in the connectedness of single methods, *i.e.*, in the question which method B is most closely related to method A, the 'single–linkage' method is used in the grouping. It separates those neighbours on the graph spanned by a 'minimal spanning tree', which are most dissimilar. The topology of this procedure can be described by a tree, indicating what evaluation methods are most similar. The higher a split in the tree, the more significant is the difference between members of the left and the right node.

Stability. To test the results for stability and consistence over all different data sets, *i.e.*, all ten MRSI data volumes, we used a bootstrapping strategy to determine variance of the measured distances and homogeneity of the resulting groupings. Two bootstrapping strategies were chosen. To assess the general stability of the

dissimilarity matrix, the full distribution was bootstrapped in a first approach. To assess the influence of inter–patient variation, sampling was performed blocked over patient labels in a second approach.

The bootstrapping was repeated 100 times, *i.e.*, 100 data sets were randomly sampled with replacement from the original data. Correlations were calculated for each of them and the variance of the observations was determined for each entry of the dissimilarity matrix. The hierarchical clustering was performed on each matrix, and finally a consensus tree [Holm07] was determined (using the “ape” package with the statistical programming language R), representing the topology of the most frequent splits.

Results

The ‘reject’ option limited the evaluation to the smallest common subset of all methods. This, together with the transformation of intensity ratios from spectral fitting to class labels will be presented first, before results of the test for correlation will be explained. Dissimilarities expressed in Kendalls *tau* are proportional to the percentage of concordant pairs in a cross tabulation. Cross tabulations, or hit matrices, are a mean for the paired comparison of two distributions. Few of such comparisons stand out from Table 2 [DISSIMILARITY] and Figure 6 [HIERARCH] and shall finally be looked at in detail.

Data preparation

Evaluation times. The manual evaluation including anatomical information (‘an’) of a single patient data set with up to 4096 spectra lasted 120–180 min. When the experts considered anatomical information and excluded all spectra outside the prostate this time was cut down to 30–45 min for each dataset. The “blinded” evaluation of a single dataset (‘e1’ and ‘e2’) without the anatomical context also lasted 30–45 min, since there only the subset from within the prostate had to be evaluated. The CLARET tool performed the pattern recognition (‘pr’) of a complete MRSI data set within 7–11 min depending on the number of noisy spectra that were excluded in advance. The evaluation using the automated spectral fitting lasted XX min for fit in the frequency domain (‘f1’ and ‘f2’) and XX min for fit in the time domain (‘ft’). [Das weiss Michael.]

Evaluable data. Out of altogether 40960 spectra from all patients about 9000 were localized within the prostate. Overall we found 1018 spectra deemed evaluable by all methods (for the individual patient: 63, 224, 8, 69, 244, 42, 294, 0, 11, 63). Concerning the employed methods (‘an’, ‘pr’, ‘e1’/’e2’, ‘f1’, ‘f2’, ‘ft’) significantly more spectra were deemed evaluable when the anatomical context (‘an’) of a spectrum was known (Table 1 [OVERLAP], first row; Figure 2 [DATA], Figure 3 [METHOD]). Only 40–51% of these spectra were assigned a label in an evaluation without anatomical evaluation. Spectra chosen by the automated pattern recognition (‘pr’) or by the experts (‘e1’, ‘e2’) were most likely (77–84%, first column, Table 1 [OVERLAP]) those chosen in the anatomical evaluation while no more than 50% of

these spectra deemed evaluable in the spectral fitting ('f1', 'f2', 'ft'). The following analysis is restricted to the 1018 spectra deemed evaluable by all methods.

Class labels from ratios. Average scores from visual inspection and the (median) results of the spectral fitting follow a linear trend for low score values (classes 1–4, Figure 4 [THRESHOLD]). High ratios for spectra in class 4 and 5, from spectra with low or unresolved citrate signal, lead to a distinct overlap of these classes (Figure 4 [THRESHOLD]) and a variation of the threshold between these two groups. The average thresholds were 0.89, 1.29, 1.96 and 5.34, when fitting spectra in the time domain ('ft'); 0.89, 1.30, 1.90, and 3.90, when fitting in the frequency domain using algorithm 1 ('f1'); and 0.81, 1.10, 1.80, 4.50 when using algorithm 2 ('f2').

Dissimilarities

Tau values. Correlations between the results of different groups range from *tau* = 0.51 (all values according to Table 2 [DISSIMILARITIES]) between expert 2 ('e2') and the time domain fitting ('ft') to *tau* = 0.95 between the two frequency domain fits ('f1' and 'f2'). The two experts ('e1' and 'e2') reach a value of 0.84, similar to the difference between the spectral fitting routines of frequency and time domain ('f1' and ft: 0.81; 'f2' and ft: 0.85). The experts' average score ('ea') reaches 0.67 when compared with the anatomical evaluation and 0.77 when compared with the automated pattern recognition ('pr'). 'pr' correlates well with both the spectral fitting (0.81, 'f1') and the more experienced of the experts (0.83, 'e1'). It is also the method reaching the highest correlation with the anatomical evaluation ('pr' and an: 0.73).

Grouping. A projection of the results into two dimensions using multidimensional scaling shows a similar pattern (Figure 5 [MDS]). While the anatomical evaluation ('an') clearly differentiates from the single–voxel–based evaluation, results from the visual inspection ('e1', 'e2') and results from the spectral fitting ('ft', 'f1', 'f2') group together, with the automated pattern recognition ('pr') in between. A more detailed analysis of this grouping – also considering (dis–)similarities beyond the two–dimensional projection – provides the hierarchical clustering of the entries as demonstrated in Table 2 [DISSIMILARITIES]. It shows a similar pattern, i.e., the data groups in three different classes: anatomical evaluation ('an'), single–voxel inspection

(automated pattern recognition, pr, and visual inspection, ‘e1’, ‘e2’), and spectral fitting. The highest correlation – *i.e*., the lowest, least significant split – was observed between the two implementations of spectral fitting in the frequency domain (‘f1’, ‘f2’). The two visual inspections (‘e1’, ‘e2’) were also related, although they frequently grouped with the automated pattern recognition ‘pr’ (24% in both bootstrapping approaches). Comparing both readers, ‘e1’ often labeled spectra that seemed not evaluable to reader ‘e2’ (‘e1’ evaluated 422 spectra that deemed not evaluable by ‘e2’, while ‘e2’ evaluated only 53 spectra which seemed not evaluable to ‘e1’, Table 1 [OVERLAP]). Nevertheless there is a high agreement (98%) between the labels of ‘e2’ and ‘e1’ (Table 1 [OVERLAP]).

Single groups

Single voxel. The different spectral fitting approaches show a high similarity corresponding to expectation, assigning nearly identical intensity ratios to most of the spectra (‘ft’ and ‘f1’, Figure 7 [RATIOS]). This leads to highly correlated cross–tabulations with almost identical labels (90.1 %, ‘f1’ and ‘f2’, Table 4 [QUANT]). The two implementations of spectral fitting in the frequency domain (‘f1’, ‘f2’) show significant differences of more than 2 classes in 2.7 % of all spectra, whereas 4.3% and 6.9% of the spectra show these differences when comparing the results of the time domain fitting ‘ft’ with those of ‘f1’ and ‘f2’, respectively. Overall the spectral fitting is highly reproducible (*tau* = [0.81, 0.85, 0.85]), although also a number of gross errors (differences of two or more classes) can be observed.

Differences between spectral fitting and visual inspection (‘e1’, ‘e2’) can be observed for those spectra which are affected by technical artifacts (Figure 3 [METHODS]), as a close inspection of Figure 2 [DATA] (*e.g*., slice 12) reveals.

The assignments by the MRS experts were not as consistent as the fitting routines (*tau* = 0.83 between experts, *tau* = [0.95, 0.85, 0.81] between ‘f1’, ‘f2’ and ‘ft’, Table 2 [DISSIMILARITIES]), due to slight differences in the assignments of definitely–normal *vs*. probably–normal (classes 1 and 2) and definitely tumor *vs*. probably tumor (classes 4 and 5). Differences of 2 or more classes, however, only occurred in 2.1 %

of the spectra, a value which is less than the inter–algorithm variation of the spectral fitting (Table 3 [EXPERTS]).

The automated inspection of the spectral pattern ('pr') is biased towards the extreme ends of possible class labels, but has – as the high correlation coefficient in Table 2 [DISSIMILARITIES] indicates – a similar ordering as the visual inspection. It yields more 'conservative' assignments than the human expert (Figure 9 [CLARET]). Among all single–voxel methods it is the one which is closest to the results of the anatomical evaluation (*tau* = 0.73).

Anatomical. A similar binomial behaviour is the most prominent difference between anatomical and 'blinded' evaluation ('an' and 'e1', 'e2'; Figure 8 [SPATIAL]). A systematic shift to either end, towards a binary decision ('tumor'/'healthy') is enforced in the anatomical evaluation (see deviation from polynomial regression in Figure 8 [SPATIAL]). Besides this, differences of two or more classes can be observed in certain areas of the MRSI volume. Most of them occur at the border of the data volume which was deemed evaluable in the blinded, single–voxel evaluation (Figure 8 [SPATIAL]). These regions typically show low SNR and are often also located in the periphery of the prostate, or distant from the endorectal coil.

Discussion

General limitations of the study

One limitation of our study could be the missing whole mount section and correlation of the tumor histology. Since emphasis was laid on the comparison of manual, automated, and blinded MRSI data evaluation, the knowledge about an existing tumor however which in this study has been proven in all patients by biopsy appears to be sufficient.

The quality of spectral data plays an important role with respect to reproducibility and objectivity in the comparison of different evaluation methods. A manual and unblinded approach certainly copes better with worse spectral quality since an experienced MR spectroscopist is able to determine suspicious patterns also in noisy

spectra. Since a blinded or automated evaluation with a randomized data “stock” would have been prejudiced we excluded not representative MRSI data from further evaluation.

We were also able to demonstrate that the more experienced reader (‘e1’) labeled more spectra in total (Table 2, Table 3) than the less experienced reader (‘e2’). However, it can not be proven that these spectra were labeled correctly, because we had no cross section histology. There were also differences between the experts in the extreme classifications of “tumor/possibly tumor” and “possibly no tumor/no tumor” which we attribute to the fact that no cut–off values, *e.g.*, from CC/C, were used, but only a visual pattern. Since only 2.1 % labels differed in more than 2 classes both experts rated very similar.

Visual inspection of the MRSI volumes

It is highly time consuming for a reader to classify each spectrum according to a tumor probability. This is one of the reasons why MR spectroscopy has up to now not found its way into broad routine diagnostic imaging. The enormous number of spectra provided by MRSI requires software that manages the data load. Since only a fraction of the acquired spectra is usable – a part of the spectra is compromized by poor quality or is localized outside the region of interest – a pre–selection is desirable. This pre–selection is usually performed by the experienced MR spectroscopist in a routine manner because he/she is able to include the anatomical information provided by the underlying MR image. Anatomical images certainly help, especially for voxels near the prostate capsule and the central gland around the urethra and the ejaculatory duct, to decide whether the spectra can be evaluated. Nevertheless, this knowledge can also be misleading since tumor areas and prostatitis are hypointense on T_2–weighted MR images. Therefore we tried to avoid this bias in our visual evaluation by modifying the contrast of the image. By this means the signal changes in the prostate itself were masked, while the outer margins remained visible.

On the other hand the rapid pre–selection of the relevant spectra performed by software allows the reader to focus on regions marked as suspicious. The exclusion of not relevant spectra outside the prostate can cut down markedly the time for

screening a whole data set. In doubtful cases the corresponding spectrum should be easily inspected and also conspicuous regions in the T_2–weighted images be considered simultaneously. The inspection of critical areas like tumors near the capsule or the seminal vesicles – which influence the surgical strategy (*e.g.*, nerve sparing, endoscopic technique etc.) – remains necessary.

One must also expect a bias by spectra from the surrounding tissue that indicate cancer which can lead the MR spectroscopist to label a spectrum suspicious while on the other hand he would discard it due to poor quality in a different context or a blinded situation. In a routine evaluation of spectra this bias can not be excluded and is even sometimes welcomed, particularly in patients where a slight decrease of citrate levels over a larger area is identified which an automated tool would not consider suspicious. Also spectra of poor quality benefit from a manual approach, where an expert can detect single usable spectra within a whole MRSI data set.

Single–voxel evaluation

The spectral fitting routines were highly consistent, indicating that the general parameterization and application of these algorithms was appropriate and correct. Nevertheless, some 'noise' can be observed, leading to a certain number of gross misclassifications. Experts are not as accurate as the spectral fitting (in terms of *tau*, Table 2 [DISSIMILARITIES]), but less susceptible to gross errors. In general, both approaches follow a linear relationship (Figure 7 [RATIO]), leading to the same classification of the data. Differences can be observed, when certain artifacts are present in the spectrum, leading to the overall differences observed in the hierarchical grouping (Figure 6 [HIERARCH]).

Spectral fitting is the current standard approach in the analysis of MRSI data of the prostate, using the ratio of choline+creatine *vs.* citrate ^{1}H MR signal intensities (CC/C) for diagnosis. The ratio itself does not reflect a tumor grade and allows only a probabilistic interpretation with respect to the threshold. So this puts down the question of correct scaling or transformation of CC/C ratio. Typically a linear relation is assumed to hold between the extreme ends of the CC/C distribution, with spectra from healthy tissue on the one side and tumor spectra on the other. Consequently, the CC/C ratios loose sensitivity near their threshold between cancer and benign

lesions, raising the question where to set the threshold. In the present study, for example, we found CC/C thresholds of approximately 1.1/1.3 indicating critical changes, as opposed to values >0.8 in earlier studies [ScHV99], [FüSH07]. These differences can presently be explained only by low citrate levels particularly in one of our patients with unresolved citrate peak. Also different thresholds are required in the peripheral zone and central gland [FüSH07], and some authors rise their classificator score when the choline resonance is clearly resolved [JuCV04]. This, however, requires information on anatomical localization which is not available in an automated analysis of results from spectral fitting.

It might be expected that a MR spectroscopist – visually inspecting the pattern of the metabolic signal and being aware of this possibly unphysiological linearity – seeks for more deliberate decisions. However, the concordance of results of spectral fitting and visual inspection shows that the expert was unwittingly looking for linear relationships in the single–voxel evaluation (Figure 4 THRESHOLDS).

Results from the anatomical evaluation done by the experts were different. Here no linear relations were observed, but rather binary decisions. In the classification of the whole MRSI volume it was more natural to follow the task to “find and locate the tumor” – yielding binary decisions rather than to “score the presented spectrum” with a linear relation as in a single–voxel examination.

Interestingly, the automated pattern recognition also sought for binary decisions (Figure 9 CLARET). Its classifier, a logistic regression, had originally been learned from completely labeled MRSI data volumes [KeMZ07]. In this, it is closest to the anatomical evaluation, explaining why pattern recognition and anatomical inspection showed high similarity (Figure 6 HIERARCH).

Anatomical evaluation

While we observed that differences in the automated analysis of single spectra were quite moderate and in the same order of magnitude as the inter–operator variation on blinded evaluation, we still observe a wide gap to the anatomical analysis of a spectroscopic image and the evaluation of a single spectrum.

First, the analysis of a spectroscopic image focuses on localizing and outlining a possibly suspicious area, requiring a more binary evaluation function. This, as the pattern recognition shows, can easily be learned from spectroscopic images and then be used in a single–voxel processing. Outlining a tumor, however, requires to focus on the transition between spectra from healthy tissue and tumor spectra. At the margin of a tumor, the “undecided” spectra of class 3 will clearly be suspicious, while being “normal” in other areas of the prostate (*e.g.*, around the urethra).

Second, a much larger data volume could be labeled in the anatomical evaluation (Figure 2 [DATA], Figure 2 [METHOD]). Differences between single–voxel and anatomical evaluation typically occurred in voxels with weak signals. Random fluctuations or artifacts (such as chemical–shift artifacts) were interpreted as changes of the spectral signal, which could be identified as such owing to the anatomical context, *e.g.*, the presence of spectra from the surrounding region affected likewise. In this case the experts were able to classify these spectra, but they were unable to classify these spectra when they were presented to them without this additional information.

Overall, interpreting the anatomical context of a spectrum and interpreting the physiological background led to a more reliable analysis of the data which, of course, corresponds to expectation. Two directions of using anatomical context might be followed in the future in an automated analysis of the MRSI data set.

First, a localization of the spectrum in its anatomical context, *i.e.*, considering that the voxel signal originates from within the prostate, could allow to adjust the analysis for the anatomical heterogeneity of the CC/C value of normal tissue. Anatomical atlases are available for the prostate [CoDe07] and might be a useful means for this localization task. Second, and in addition to this global localization, the anatomical context of a spectrum should be considered. Training classifiers on cliques, rather than on single voxels, is a straightforward approach here [LaPe05]. Markov random fields can be used to trade confidence in the information of the single spectrum with the spectral information of its neighbourhood. A fixed coupling term between these two domains allows, for example, a semi–supervised classification of MRSI data [GoMe07] based on few labeled spectra and segmenting the whole volume.

Discriminative random fields even allow inferring the spatio–spectral coupling from the data, adjusting it optimally to the SNR of the specific instrumental setting [KeMW07].

While we observed advantages in analysing whole MRSI slices instead of single voxels, it remains difficult for a human reader to make use of the full information from all three dimensions. Thus, beside the general benefit of an automated processing – facilitating analysis and increasing objectivity – a main virtue of automation in the anatomical analysis might be its potential to be easily extended to higher dimensions of the complete MRSI volume.

Conclusions

This study demonstrates the potential role and the need for pattern recognition methods for the diagnostic evaluation of data obtained in MRSI examinations. While the human reader is better in identifying the anatomical borders and the morphological context of spectra the manual evaluation lacks objectivity and reproducibility as indicated by the higher amount of spectra assigned to benign tissue in manual evaluation. On the other hand, the blinded reader is as good as the automated tool. Therefore a combination of manual and automated methods seems to be an optimal approach for the MR spectroscopist to cut down time constraints in clinical routine, without completely abandoning manual evaluation of MRSI data with respect to tissue–specific knowledge.

A machine–based processing is indispensable in the analysis of MRSI data. Robustness is a requirement for automated algorithms. In particular, MRSI of the prostate is well suited for such an approach: spectra have lower signal intensities compared to MRSI spectra of the human brain and it is highly desirable to take the anatomical context of the prostate into account. The organ has a simple shape, but an inhomogeneous distribution of normal–state concentrations of the different metabolites that are detectable by ^{1}H MRS. It is not even necessary to include the anatomical knowledge since the CLARET tool already proved a good separation between prostate and surrounding tissue by using a nonlinear classification approach [MeKW08]. This way the number of spectra to be evaluated is considerably cut down to the relevant within the organ.

Finally we see a significant advantage of 2D spectroscopic imaging over single–voxel MRS, since an automated algorithm considering the anatomical context can naturally be extended to the complete information of a 3D MRSI data set. Automated approaches have the potential to include anatomical context into the evaluation of the MRSI volume. Although to our knowledge no current software provides this service, this comparison of a manual, pattern–recognition–based and blinded evaluation emphasizes the need for such an automated approach.

References:

[ScHV99] Scheidler J, Hricak H, Vigneron DB, et al. (1999) Prostate cancer: Localization with Three–dimensional Proton MR Spectroscopic Imaging – Clinicopathologic Study. Radiology 213: 473–480.

[KeMN06] CLARET: a tool for fully automated evaluation of MRSI with pattern recognition methods. Kelm BM, Menze BH, Neff T, Zechmann CM, Hamprecht FA; in: Bildverarbeitung für die Medizin 2006 – Algorithmen, Systeme, Anwendungen Springer (2006), p. 51–55.

[KeMZ07] Kelm BM, Menze BH, Zechmann CM, Baudendistel KT, Hamprecht FA. (2007) Automated Estimation of Tumor Probability in Prostate Magnetic Resonance Spectroscopic Imaging: Pattern Recognition vs Quantification. Magn Reson Med 57:150–159.

[RePR07] Stefan A. Reinsberg, Geoffrey S. Payne, Sophie F. Riches, Sue Ashley, Jonathan M. Brewster, Veronica A. Morgan, Nandita M. deSouza. Combined Use of Diffusion-Weighted MRI and 1H MR Spectroscopy to Increase Accuracy in Prostate Cancer Detection. AJR 2007; 188:91–98.

[MuHK87] Mukamel E, Hannah J, de Kernion JB (1987) Pitfalls in preoperative staging in prostate cancer. Urology 30: 318–321.

[AnCM89] Andriole GL, Coplen DE, Mikkelsen DJ, Catalona WJ (1989) Sonographic and pathological staging of patients with clinically localized prostate cancer. J Urol 142: 1259–1261.

[TeXZ94] Tempany CM, Xiao Z, Zerhouni EA et al. (1994) Staging of prostate cancer: results of Radiology Diagnostic Oncology Group project comparison of three MR imaging techniques. Radiology 192: 47–54.

[BaML96] Bartoluzzi C, Menchi I, Lencioni R et al. (1996) Local staging of prostate carcinoma with endorectal coil MRI: correlation with wholemount radical prostatectomy specimens. Eur Radiol 6: 339–345.

[DASW98] D'Amico AV, Schnall M, Whittington R et al. (1998) Endorectal coil magnetic resonance imaging identifies locally advanced prostate cancer in select patients with clinically localized disease. Urology 51: 449–454.

[QuFD94] Quinn SF, Franzini DA, Demlow TA et al. (1994) MR imaging of prostate cancer with an endorectal surface coil technique: correlation with whole–mount specimens. Radiology 190: 323–327.

[PeKJ96] Perrotti M, Kaufman RP, Jennings TA et al. (1996) Endorectal coil magnetic resonance imaging in clinically localized prostate cancer: is it accurate? J Urol 156: 106–109.

[ScYT92] Schiebler ML, Yankaskas BC, Tempany C et al. (1992) MR imaging in adenocarcinoma of the prostate: interobserver variation and efficacy for determining stage C disease. AJR 158: 559–562.

[IkKK98] Ikonen S, Kärkkäinen P, Kivisaari L et al. (1998) Magnetic resonance imaging of clinically localized prostatic cancer. J Urol 159: 915–919.

[PrHN96] Presti JC, Hricak H, Narayan PA, Shinohara K, White S,Carrol PR (1996) Local staging of prostatic carcinoma: comparison of transrectal sonography and endorectal MR imaging. AJR 166: 103–108.

[RiZG90] Rifkin M, Zerhouni E, Gatsonis C et al. (1990) Comparison of magnetic resonance imaging and ultrasonography in staging early prostate cancer.Results of a multi–institutional cooperative trial. N Engl J Med 323: 621–626.

[EpPW93] Epstein JI, Pizov G, Walsh PC (1993) Correlation of pathologic findings with progression after radical retropubic prostatectomy. Cancer 72: 3582–3593.

[OuPS94] Outwater EK, Petersen RO, Siegelman ES et al. (1994) Prostate carcinoma: assessment of diagnostic criteria for capsular penetration on endorectal coil MR images. Radiology 193: 333–339.

[DrFT99] Drew PJ, Farouk R, Turnbull LW, Ward SC, Hartley JE, Monson JR (1999) Preoperative magnetic resonance staging of rectal cancer with an endorectal coil and dynamic gadolinium enhancement. Br J Surg 86: 250–254.

[BuGB94] Buist MR, Golding RP, Burger CW et al. (1994) Comparative evaluation of diagnostic methods in ovarian carcinoma with emphasis on CT and MRI. Gynecol Oncol 52: 191–198.

[Kend48] Kendall M. (1948) *Rank Correlation Methods*, Charles Griffin & Company Limited

[JeSW07] Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ (2007). Cancer statistics, 2007. CA Cancer J Clin 57: 43–66.

[HrCE07] Hricak H, Choyke PL, Eberhardt SC, Leibel SA, Scardino PT (2007) Imaging Prostate Cancer: A Multidisciplinary Perspective. Radiology 243: 28–53.

[MeKW08] Menze BH, Kelm BM, Weber MA, Bachert P, Hamprecht FA (2008) Mimicking the human expert: a pattern recognition approach to score the data quality in MRSI. MRM, in press (2008)

[HrWV94] Hricak H, White S, Vigneron D, et al. (1994) Carcinoma of the prostate gland: MR imaging with pelvic phased array coil versus integrated endorectal-pelvic phased-array coils. Radiology 193: 703–709.

[ScKR04] Scheenen TWJ, Klomp DWJ, Röll SA, Fütterer JJ, Barentsz JO, Heerschap A. (2004) Fast Acquisition-Weighted Three-Dimensional Proton MR Spectroscopic Imaging of the Human Prostate. Magn Reson Med 52:80–88.

[KuVH96] Kurhanewicz J, Vigneron DB, Hricak H, Narayan P, Caroll P, Nelson SJ. (1996) Three-dimensional H-1 MR spectroscopic imaging of the in situ human prostate with high (0.24-0.7 cm^3) spatial resolution. Radiology 198: 795–805.

[JuCV04] Jung JA, Coakley FV, Vigneron DB, Swanson MG, Qayyum A, Weinberg V, Jones KD, Carroll PR, Kurhanewicz J. Prostate depiction at endorectal MR spectroscopic imaging: investigation of a standardized evaluation system. Radiology 233:701–708.

[FüSH07] Fütterer JJ, Scheenen TW, Heijmink SW, Huisman HJ, Hulsbergen-Van de Kaa CA, Witjes JA, Heerschap A, Barentsz JO. (2007) Standardized threshold approach using three-dimensional proton magnetic resonance spectroscopic imaging in prostate cancer localization of the entire prostate. Invest Radiol. 42:116-122.

[PiBO92] Pijnappel WWF, A. Van den Boogaart, R. de Beer, and D. Van Ormondt. (1992) SVD-based quantification of magnetic resonance signals. J. Magn. Reson. 97, 122–134.

[BeBO92] de Beer R, van den Boogaart A, van Ormondt D, Pijnappel WW, den Hollander JA, Marien AJ, Luyten PR. (1992) Application of time-domain fitting in the quantification of in vivo 1H spectroscopic imaging data sets. NMR Biomed. 5:171–178.

[NaCD01] Naressi A, Couturier C, Devos JM, et al. (2001) Java-based graphical user interface for the MRUI quantitation package. MAGMA 12:141–52. http://www.mrui.uab.es/mrui/

[Kreis04] Kreis R. Issues of spectral quality in clinical 1H magnetic resonance spectroscopy and a gallery of artifacts. *NMR Biomed* 2004;17:361–381.

[Holm07] Holmes S. Bootstrapping Phylogenetic Trees: Theory and Methods. Statist. Sci. Volume 18, Issue 2 (2003), 241-255.

[BoMa07] Bouix S, Martin-Fernandez M, Ungar L, Nakamura M, Koo M-S, McCarley RW Shentona ME. On evaluating brain tissue classifiers without a ground truth, NeuroImage 36 (2007) 1207–1224

[LaPe05] Laudadio T, Pels P, De Lathauwer L, Van Hecke P, Van Huffel S, Tissue segmentation and classification of MRSI data using Canonical Correlation Analysis, Magnetic Resonance in Medicine,Vol. 54, 1519-1529, 2005.

[GoMe07] Görlitz L*, Menze BH*, Weber MA, Kelm BM, Hamprecht FA. Semi-supervised tumor detection in magnetic resonance spectroscopic images using discriminative random fields. In: FA Hamprecht, C Schnörr, B Jähne (eds.) Proc 29th Symposium of the German Association for Pattern Recognition (DAGM 07), Heidelberg, Germany, *Pattern Recognition. Lecture Notes in Computer Science 4713*. Springer, Heidelberg and Berlin, 2007 224-233

[KeMW07] Kelm BM, Menze BH, Weinman J, Henning A, Görlitz L, Hamprecht FA. Trading resolution against noise in NMR spectroscopic images with conditional random fields. Technical Report, IWR, University of Heidelberg, 2007

[CoDe07] Costa, J.; Delingette, H.; Novellas, S., Ayache, N. Automatic Segmentation of Bladder and Prostate Using Coupled 3D Deformable Models.
MICCAI, 2007

[NaTI04] Nakashima J, Tanimoto A, Imai Y, Mukai M, Horiguchi Y, Nakagawa K, Oya M, Ohigashi T, Marumo K, Murai M. Endorectal MRI for prediction of tumor site, tumor size, and local extension of prostate cancer. Urology 64 (2004) 101-105.

Figures and Tables

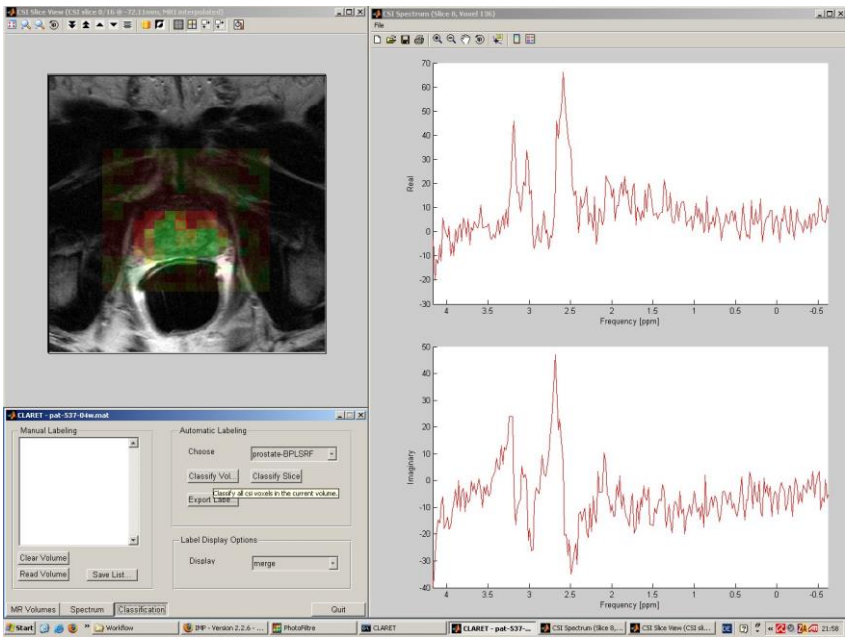
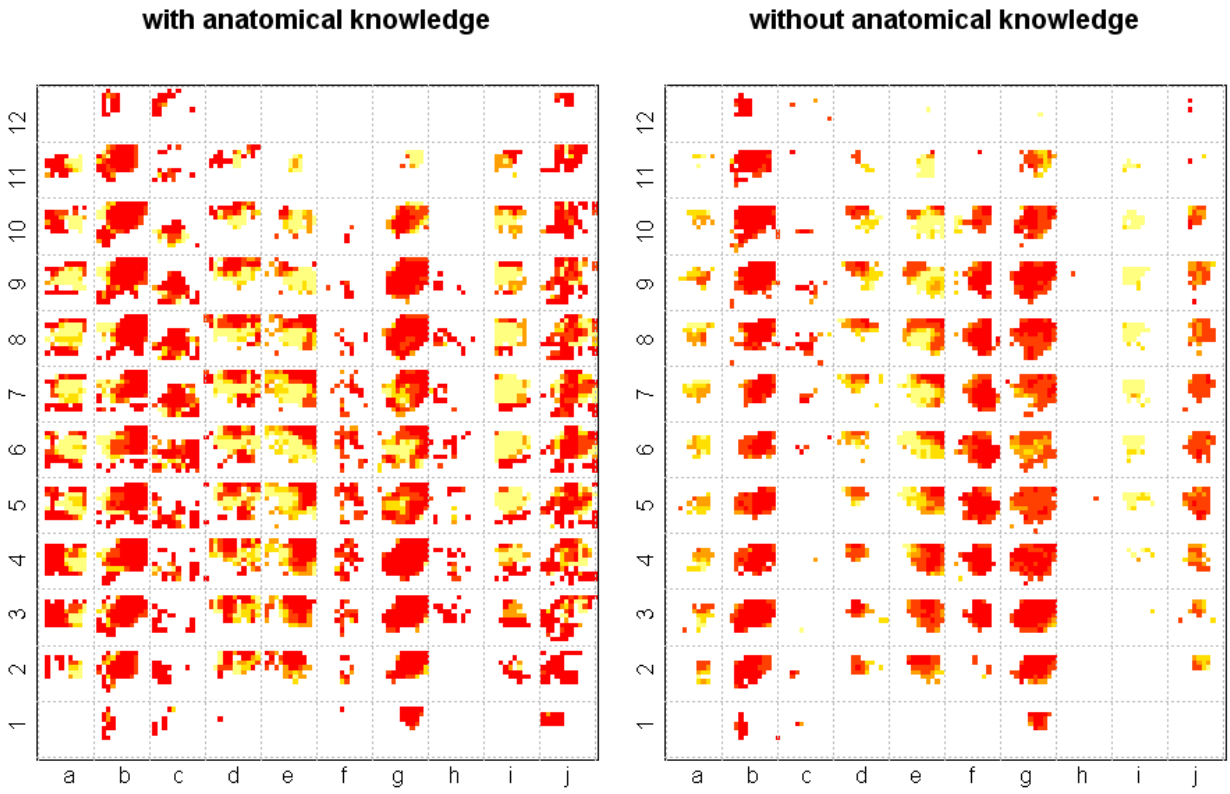


Figure 1

Color–coded tumor probability map of the prostate of a patient (pat No, Age) with adenocarcinoma, calculated and displayed by CLARET software with tumor voxel in red and areas without pathological findings in green.

Figure 2



[DATA]: *In vivo* prostate ^{1}H MRSI (1.5 T) data evaluation with (left) and without (right) inclusion of anatomical information showing maps for twelve central slices (slices 1–12) of ten different MRSI data volumes (patients 'a' – 'j'). Spectra were labeled according to a five–point scale with 1 (= "tumor"), 2 (= "possibly tumor"), 3 (= "undecided"), 4 (= "possibly no tumor"), and 5 (= "no tumor"). Voxel of spectra that identify healthy prostate tissue are marked in red (class 1), while bright yellow voxel label tumor (class 5). White voxel could not be evaluated due to poor spectral quality or localization outside the prostate.

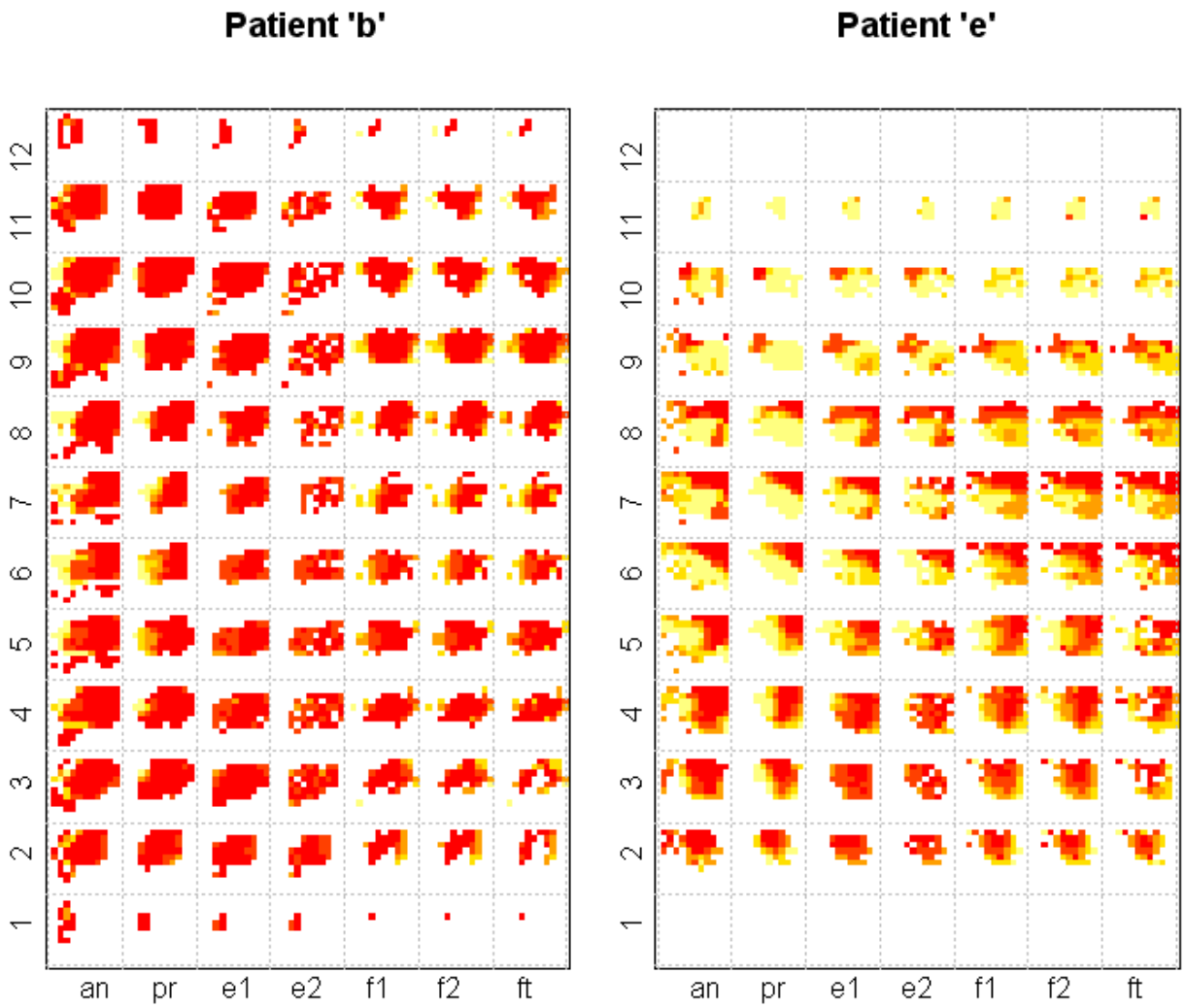


Figure 3

[METHODS]. Results of the evaluation of two exemplary MRSI data volumes (from patients ‘b’ and ‘e’, see Figure 2[DATA]) by all seven processing methods employed in this study (‘an’, …, ft.). Central slices 1–12 are shown which were evaluated by experts’ consensus with anatomical knowledge (‘an’), automated pattern recognition (‘pr’), expert 1 (‘e1’) and expert 2 (‘e2’) without anatomical knowledge, classification based on fitting in the frequency (‘f1’, ‘f2’) and time domain (‘ft’).

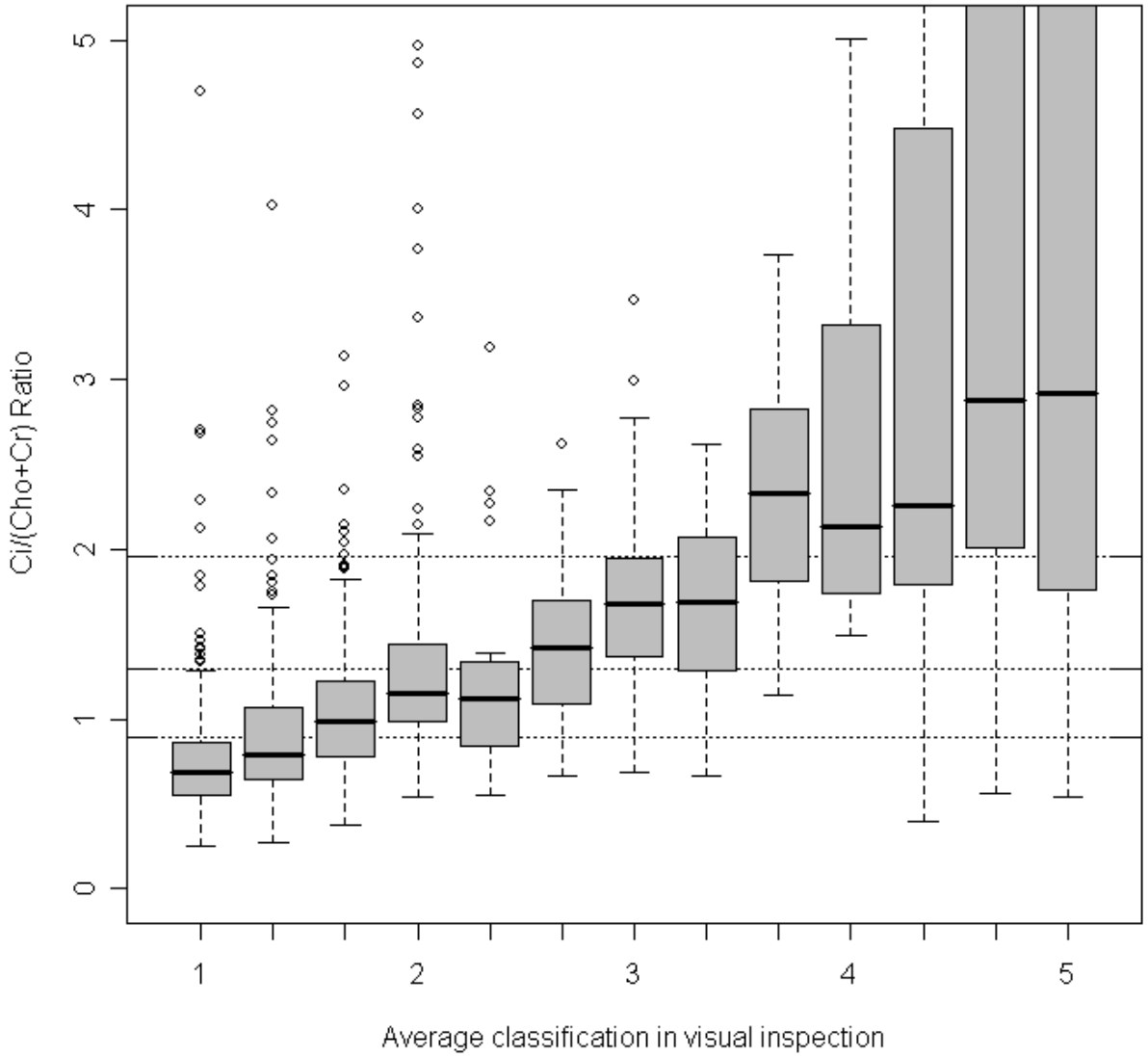


Figure 4

[THRESHOLDS]: Classification of spectra between 0 (normal) and 5 (tumor) based on the Ci/(Cho+Cr) signal intensity ratio (y axis, truncated at y = 5) obtained by fit of ^{1}H MRSI signals in the time domain. Observations are grouped along the x axis according to the average label assigned to the spectrum in the visual inspections performed by two MRS experts (evaluation with and without anatomical information). Boxplots show median (thick black lines), quartiles (box extensions) and outliers (notches and points) for the distribution of the samples in each of the group. Horizontal lines (----, at y = 0.89, 1.29, 1.96) indicate optimal thresholding for the transformation of the ratios to classes 1–5 (class 5 above y = 5.34, not shown). Average score from visual inspection and results from the spectral fitting follow a linear trend for low score values.

		‘an’	‘pr’	‘e1’	‘e2’	‘f1’	‘f2’	(‘ft’
Expert anatomical	‘an’	4516 (100%)	2093 (46.3%)	2108 (46.7%)	1786 (39.6%)	2259 (50.0%)	2306 (51.1%)	2014 (44.6%)
Pattern recogn.	‘pr’	2093 (84.0%)	2493 (100%)	1897 (76.1%)	1589 (63.7%)	1785 (71.6%)	1785 (71.6%)	1633 (65.5%)
Expert 1	‘e1’	2108 (78.8%)	1.897 (71.0%)	2674 (100%)	2252 (84.2%)	1897 (70.9%)	1906 (71.3%)	1906 (64.1%)
Expert 2	‘e2’	1786 (77.5%)	1589 (68.9%)	2252 (97.7%)	2305 (100%)	1588 (68.9%)	1599 (69.4%)	1432 (62.1%)
Fitting freq 1	‘f1’	2259 (47.3%)	1785 (37.3%)	1897 (39.7%)	1588 (33.2%)	4777 (100%)	4723 (98.9%)	4007 (83.9%)
Fitting freq 2	‘f2’	2306 (47.1%)	1785 (36.4%)	1906 (38.9%)	1599 (33.6%)	4777 (96.4%)	4900 (100%)	4010 (81.8%)
Fitting time 1	‘ft’	2014 (49.7%)	1633 (40.3%)	1715 (42.3%)	1432 (35.3%)	4007 (98.8%)	4010 (98.9%)	4055 (100%)

Table 1

[OVERLAP]: Numbers of spectra deemed evaluable in the different approaches (numbers in thousands) and overlap between the different evaluation methods. Percentages (in parentheses) indicate the amount of overlap between the methods in the respective row. As an example: among the 4516 spectra evaluated in the anatomical inspection of the data (‘an’, first row), a subset of 44.6 % (2014 spectra) could be evaluated by spectral fitting in the time domain (‘ft’). Expert 1 and expert 2 labeled 2674 and 2305 spectra, respectively with agreement in 2252 spectra.

		‘an’	‘pr’	‘e1’	‘e2’	‘ea’	‘f1’	‘f2’	‘ft’
Expert anatomical	‘an’	100 (0/0)	73 (2/11)	72 (2/9)	62 (2/11)	67 (1/10)	73 (1/6)	68 (2/6)	58 (2/6)
Pattern recogn.	‘pr’	–	100 (0/0)	83 (1/6)	68 (2/8)	77 (2/7)	81 (1/4)	75 (2/5)	64 (2/5)
Expert 1	‘e1’	–	–	100 (0/0)	84 (1/5)	93 (1/3)	74 (2/6)	68 (2/4)	59 (2/6)
Expert 2	‘e2’	–	–	–	100 (0/0)	93 (1/3)	63 (2/8)	58 (2/6)	51 (3/7)
Expert avg.	‘ea’	–	–	–	–	100 (0/0)	69 (2/7)	64 (2/6)	54 (2/5)
Fitting freq 1	‘f1’	–	–	–	–	–	100 (0/0)	95 (1/1)	81 (1/4)
Fitting freq 2	‘f2’	–	–	–	–	–	–	100 (0/0)	85 (1/4)
Fitting time 1	‘ft’	–	–	–	–	–	–	–	100 (0/0)

Table 2

[DISSIMILARITIES]: Similar performance of the different processing methods, quantified by Kendall’s *tau* and over all ten MRSI data volumes. Values are given in percent, 100 % indicating perfect correlation and 0% complete randomness between two methods. Values in parentheses show the standard deviation of Kendall’s *tau* in a patient–wise bootstrapping (first value) or bootstrapping over the full data set (second value). Data are visualized in Figure 5 [MDS] and Figure 6 [HIERARCH].

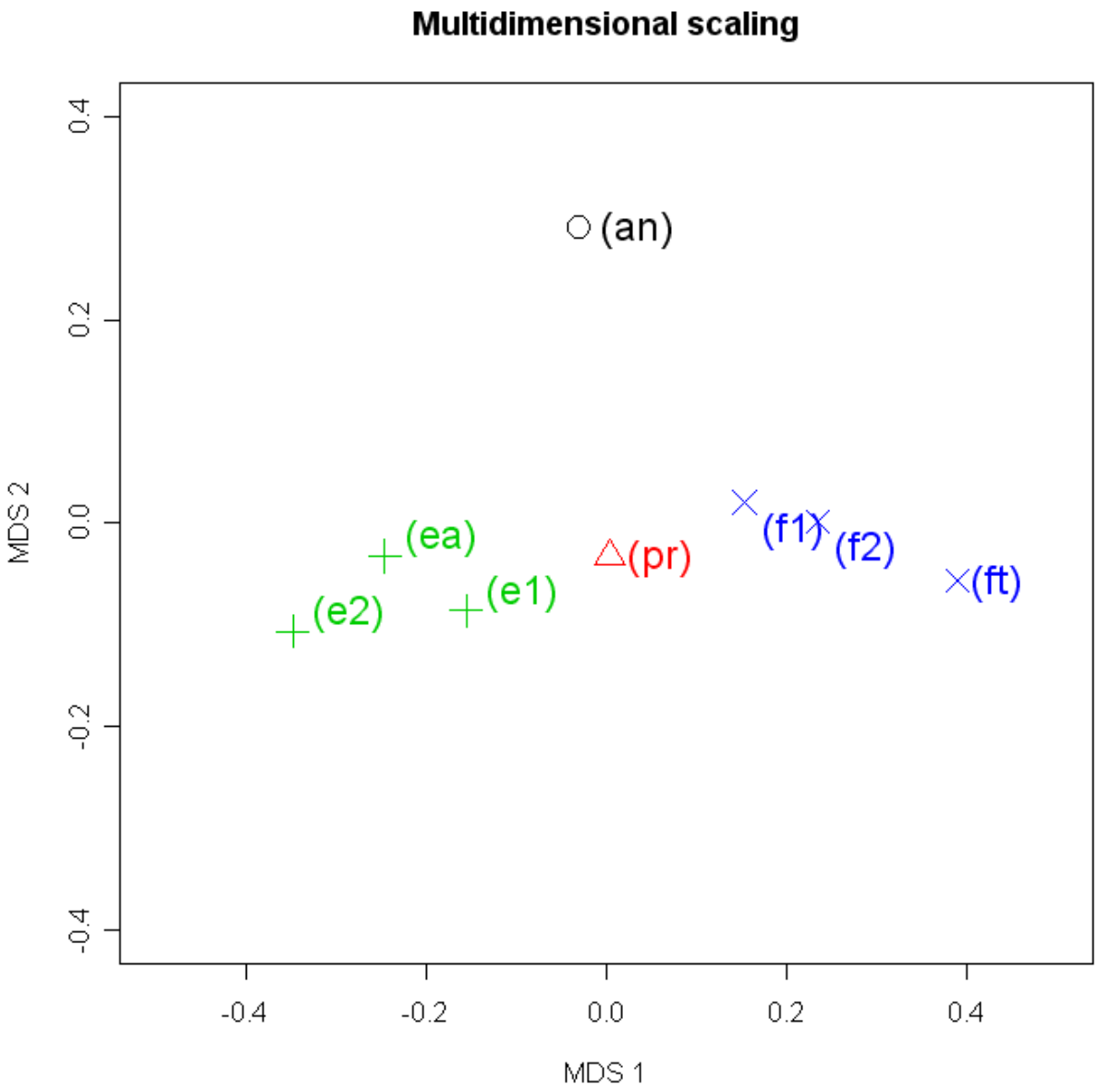


Figure 5

[MDS]: Projection of entries of Table 2 [DISSIMILARITIES] into two dimensions by multidimensional scaling. Distances in the plane encode the (dis–)similarity of the different MRSI data processing methods. While results of fitting in the frequency domain ('f1', 'f2') are at nearly identical positions, the anatomical evaluations ('an') separate from the other post–processing methods. Automated pattern recognition ('pr') is located between visual inspection ('e1', 'e2', 'ea') and spectral fitting ('ft', 'f1', 'f2').

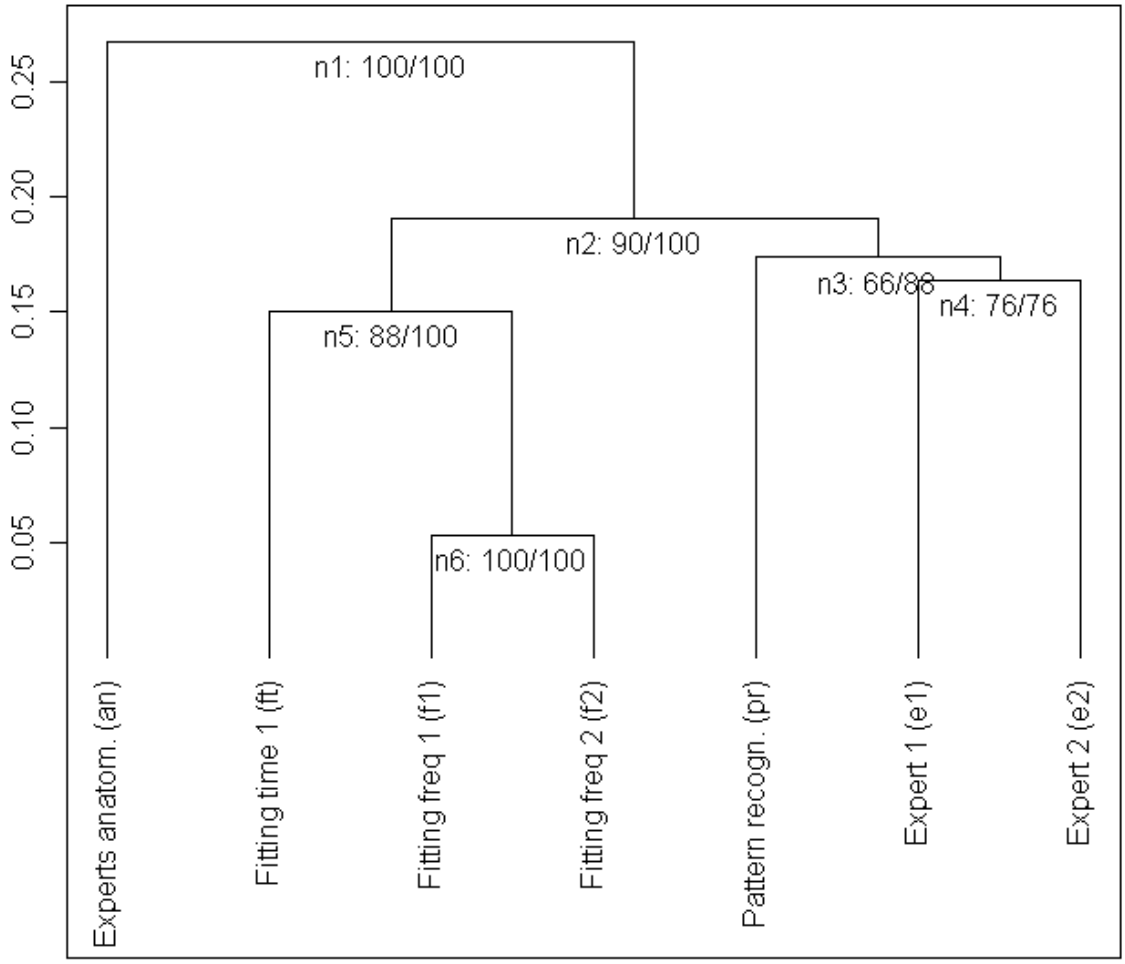


Figure 6

[HIERARCH]: Similarity of different post–processing methods of in vivo ^{1}H MRSI data in a hierarchical segmentation, based on data in Table 2 [DISSIMILARITIES]. The higher the split in the dendrogram, the more dissimilar are the members of the nodes. Evidence for a certain grouping is determined in a bootstrapping (first value: patient–wise sampling; second value: random sampling). As expected results of anatomical analysis are different from all methods evaluating spectra without anatomical information, ‘f1’ and ‘f2’ are in the same node, and finally a grouping in spectral fitting (‘ft’, ‘f1’, ‘f2’) pattern and inspection (‘pr’, ‘e1’, ‘e2’) is observed.

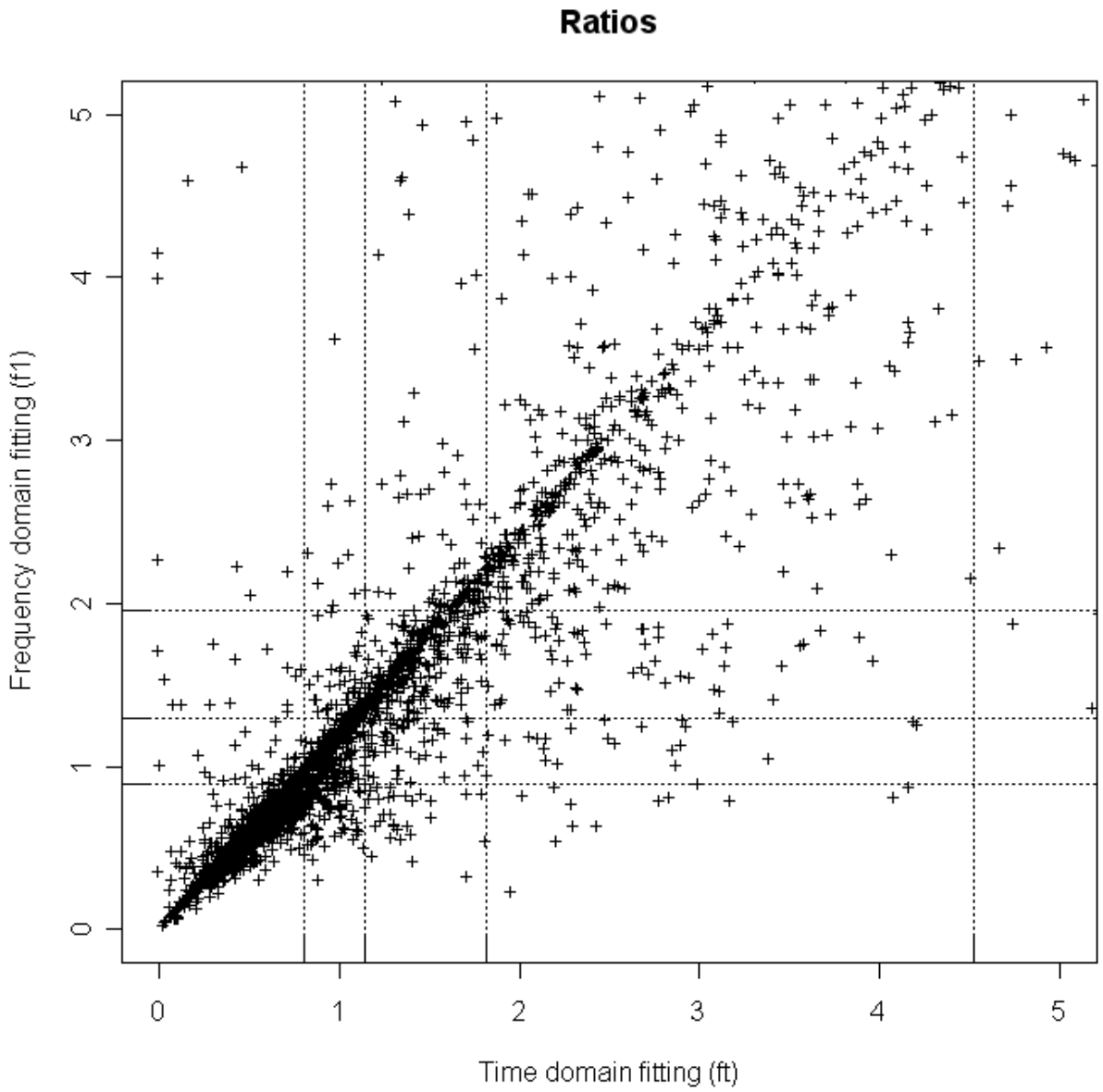


Figure 7

[RATIOS]: Results for Ci/(Cho+Cr) ratios from fitting in frequency (y–axis) and time domain (x axis, 'AMARES' implementation); each cross indicates the result for a single spectrum. Ranges are truncated at 5.xxx for both axes. Dotted lines indicate the thresholds transferring the continuous ratios to discrete classes 1–5 (Figure 4 THRESHOLDS).

	e2 1	e2 2	e2 3	e2 4	e2 5
e1 1	315	86	15	0	0
e1 2	28	317	7	2	0
e1 3	1	11	83	0	1
e1 4	0	2	13	44	6
e1 5	0	0	0	7	80

Table 3

[EXPERTS]: Comparison of estimations by the two experts independently evaluating single spectra without anatomical knowledge. The cross tabulation shows a high coincidence of the assessments. Disagreement occurred most frequently between classes 1–2 and 4–5. Differences of 2 or more classes were found in 21 spectra (2.1 % of total 1018, which is less than the inter–algorithm variation of the fitting in the frequency domain Table 4 [QUANT]).

	f2 1	f2 2	f2 3	f2 4	f2 5
f1 1	392	3	7	0	0
f1 2	9	228	9	9	0
f1 3	1	8	144	7	0
f1 4	3	5	21	109	1
f1 5	0	0	2	16	44

Table 4

[QUANT]. Comparison of the classification based on fitting in the frequency domain. The algorithms show a high coincidence, identical labels are assigned to most spectra (917 out of 1018 spectra, 90.1%). 27 spectra (2.7%) show differences of more than 2 classes, whereas 4.3 % (‘f1’ vs. ft) / 6.9 % (‘f2’ vs. ft) of the spectra show this differences when compared with the results of the time domain fits.

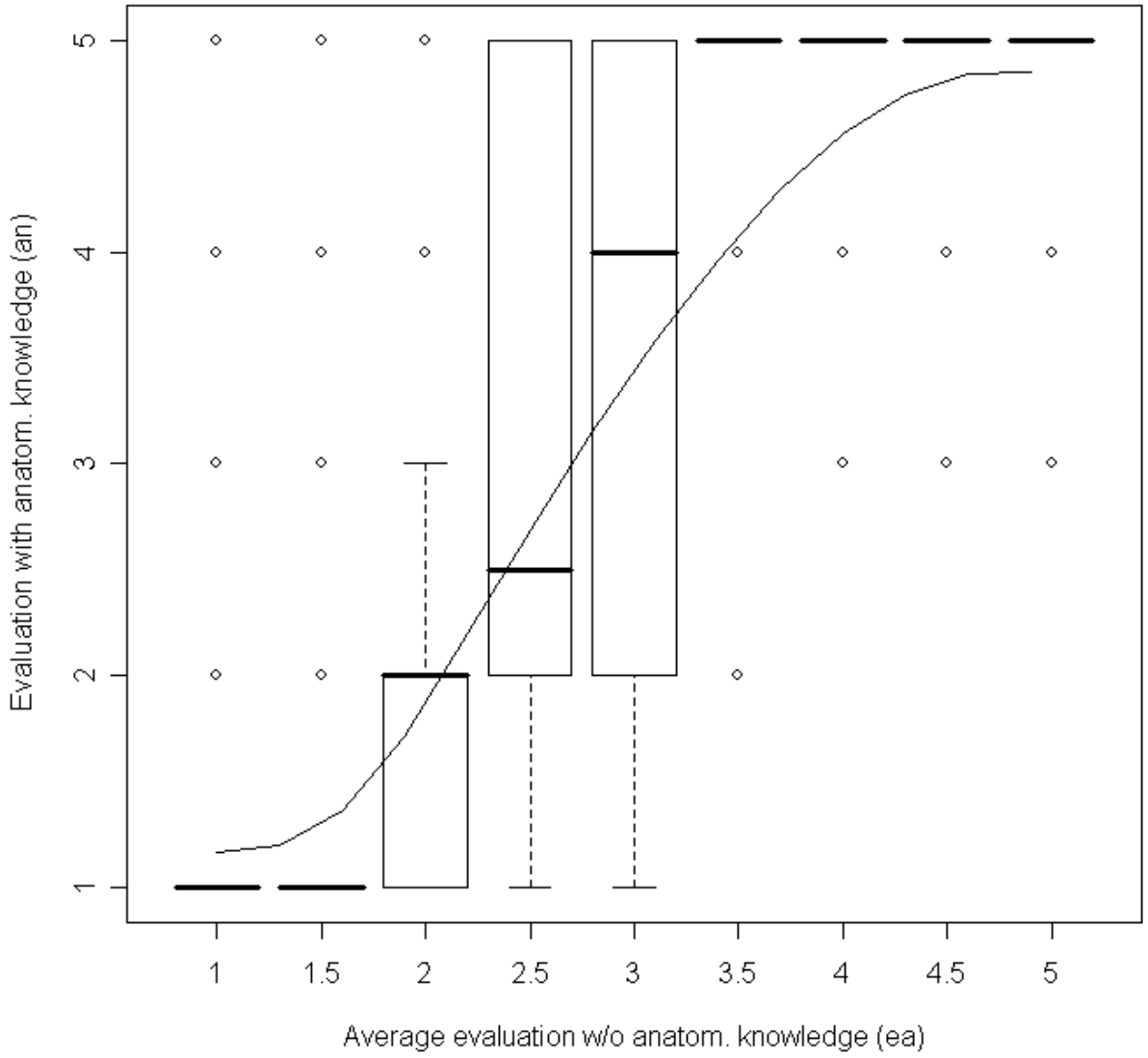


Figure 8

[SPATIAL]: Cross tabulation of results from the evaluation *with* anatomical information (y axis), grouped by the decisions from visual inspection *without* this information (x axis, average of estimations by both experts). Boxplots show median (thick black line), quartiles (box extensions), and outliers (notches and points) of the distributions. The curved black line indicates trend determined by local polynomial regression. Results deviate systematically, a binary decision (‘tumor’/’healthy’) is enforced in the anatomical evaluation.

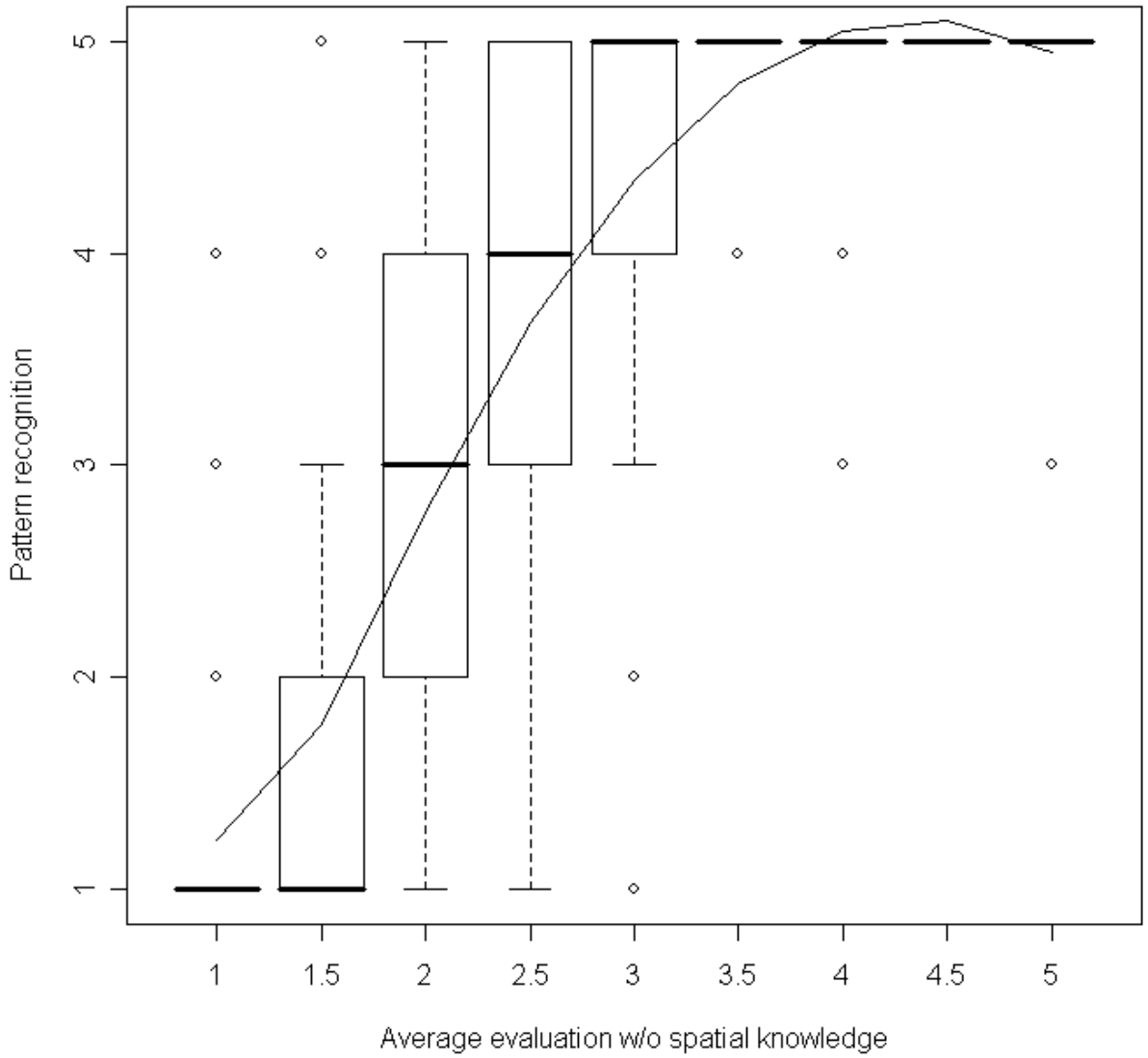


Figure 9

[CLARET]: Results from the inspection of the spectral pattern by the automated pattern recognition (y axis), grouped by the results of a visual inspection of the spectral pattern (x axis, average of estimations by both experts). Boxplots show median (thick black line), quartiles (box extensions), and outliers (notches and points). The curved black line indicates trend determined by local polynomial regression. The automated algorithm yields more ‘conservative’ assessments than the human expert; and a more binary (‘tumor’/’normal’) decision is enforced.