Mimicking the Human Expert: Pattern Recognition for an Automated Assessment of Data Quality in Magnetic Resonance Spectroscopic Images (MRSI)

Bjoern H. Menze[1,2] B. Michael Kelm[1]
Marc-André Weber[3] Peter Bachert[2,4]
Fred A. Hamprecht[1,2]

1: Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany

2: Department of Physics and Astronomy, University of Heidelberg, Germany

3: Department of Radiology, German Cancer Research Center (dkfz), Heidelberg, Germany

4: Department of Medical Physics in Radiology, German Cancer Research Center (dkfz), Heidelberg, Germany

published in *Magnetic Resonance in Medicine 59:1457–1466 (2008)*
requests for reprints to *fred.hamprecht@iwr.uni-heidelberg.de*

Abstract

Besides the diagnostic evaluation of a spectrum, the assessment of its quality and a check for plausibility of its information remains a highly interactive and thus time-consuming process in MR spectroscopic imaging (MRSI) data analysis. In the automation of this quality control, a score is proposed that is obtained by training a machine learning classifier on a representative set of spectra that have previously been classified by experts into evaluable data and non-evaluable data. In the first quantitative evaluation of different quality measures on a test set of 45 312 long echo time spectra in the diagnosis of brain tumor, the proposed pattern recognition (using the random forest classifier) separated high- and low-quality spectra comparable to the human operator (area-under-the-curve of the receiver-operator-characteristic, AUC > 0.993), and performed better than decision rules based on the signal-to-noise-ratio (AUC < 0.934) or the estimated Cramér-Rao-bound on the errors of a spectral fitting (AUC < 0.952). This probabilistic assessment of the data quality provides comprehensible confidence images and allows filtering the input of any subsequent data processing, i.e., quantitation or pattern recognition, in an automated fashion. It thus can increase robustness and reliability of the final diagnostic evaluation and allows for the automation of a tedious part of MRSI data analysis.

Key words: magnetic resonance spectroscopic imaging; quality classification; artifact recognition; automated diagnostic systems; expert systems

1 Introduction

Ever since the advent of in vivo MR spectroscopy (MRS), MR spectroscopic imaging (MRSI) has been attributed a high potential in clinical diagnostics (1). Unfortunately, the easy access of diagnostic information from MRS images and thus their routine clinical use is hindered by a time-consuming, highly interactive process of data analysis. Different methods exist to automate parts of the analysis and to facilitate the operators diagnostic evaluation of a spectrum. Spectral fitting, the most widely used method, estimates the MR signal intensities of different metabolite resonances and allows interpreting the spectral pattern on the basis of a low number of estimated parameters of the resonance line model in a subsequent step (2-6). Other methods derive the information on pathologic alterations of tissue directly from characteristic changes of the spectral pattern, without the prior estimation of resonance line features. In conjunction with MRS data analysis, these approaches are often referred to as pattern recognition methods (7-11).

Ideally, spectral fitting or pattern recognition methods should allow for a completely automated processing of MRSI data. In reality, artifacts may interfere with the employed resonance line model, the pattern recognition may inadvertently be applied to spectra originating from tissue that the algorithm has not been "trained" for, or a low signal-to-noise ratio (SNR) may lead to unreliable results. Thus, the task of providing an automatic self-control together with the diagnostic procedure itself (12,13) remains a relevant issue when a robust automation of the data processing is desired. In the context of MRSI, such a *quality assurance* is possible at two different stages: Either by monitoring the spectral quality and thus inspecting the *input* of the data analysis, or by controlling the *output* of the processing and testing for plausibility of the results. Linewidth and SNR serve as physical measures of the data quality of the input (e.g., discussed in Ref. 14). In conjunction with pattern recognition, these measures have provided the only means to assure the reliability of the diagnostic procedure so far (15). In an evaluation of the output, estimated Cramér-Rao bounds (CRBs) (16) provide a means to score the validity of the estimated model parameters, such as area or amplitude of a resonance line. If the specified spectral model applies – and no artifacts or additional spurious peaks are present – the estimated CRBs indicate the reliability of the parameter estimates that are used for the diagnostic interpretation of the spectrum. In the automated processing of MRSI data, estimated CRBs have repeatedly been suggested to provide confidence or error images of the output parameters (17-21).

In clinical practice, however, the control of these quality measures is only part of a more complex decision process: In addition to ensuring that estimated CRBs and linewidths are below certain thresholds (14) and that the SNR is sufficiently high (15), a check for abnormal changes of the spectral pattern is mandatory. In spectral fitting, this check allows ensuring that the specified spectral model is appropriate for the acquired spectrum, and in pattern recognition it ensures that the spectral pattern will be recognized by the employed algorithm. Criteria for

rejection of a spectrum are, for example, the presence of doubled or asymmetric resonance lines, large baseline artifacts and residual water peak, or the presence of other, spurious peaks (14,15). Consequently, the tedious visual inspection of spectra remains an integral part of the data processing. Extensive knowledge and experience not only in the diagnostic interpretation of MR spectra, but also in the technical evaluation of the spectral pattern remains indispensable to the operator when using MRSI.

In the present work an approach is proposed which aims at the automation of this visual control, the current "gold standard" of MRSI quality assurance. Its central idea is to mimic the decision process of a human operator and to embody his/her knowledge and expertise in artifact recognition in a machine classifier. Conceptually, this approach is closely related to pattern recognition in the classification of MR spectra of different tissue types (7-11). In practice, the proposed artifact recognition is applicable as follows: In a specific diagnostic task, a representative spectral dataset is collected from clinical routine, comprising both good spectra and spectra containing a variety of spectral artifacts. An expert labels the data, assuring that all spectra with artifacts or insufficient signal quality are rejected and classified as non-evaluable, while good spectra, e.g., of tumorous or healthy tissue, are assigned to the class of evaluable data. Then a multivariate classifier is trained to follow the decisions and the quality assessment of the human expert: In a high-dimensional space spanned by the feature vector of the spectral pattern, a nonlinear decision boundary is learned that separates the training samples of the two classes. When applied to previously unseen data, the classifier establishes on which side of the learned decision boundary a spectrum lies, and thus assigns it to one of the two groups (e.g., class "0" or "1"). Beyond a crisp 0/1 classification, it is also possible to provide a probability between 0 and 1, thus making a fuzzy or "smooth" prediction that better reflects the uncertainty that is inherent when basing a decision on very noisy observations. This approach is pursued in the following, with the 0/1 probability interpreted as a "quality score." Finally, this score can be displayed as a confidence image to allow for a comprehensive overview of the quality of the spectral data at a single glance. In a fully automated processing, the score can be used to discard spectra characterized by artifacts or insufficient signal intensity, and thus to increase the reliability and robustness of any subsequent diagnostic analysis on the remaining data.

The following section describes the design and implementation of this approach in detail. The method is evaluated on clinical data used for the diagnosis of brain tumor and the monitoring of its recurrence and compared quantitatively with the estimated CRB and SNR criteria, two alternative measures in the automated MRS(I) quality assessment. For the sake of clarity, all of these quality measures are applied individually to the test data, and not in combination (Results section). It should be noted that the proposed method can – or ideally should – be used *in addition* to, or combined with these other quality measures. Such possible extensions and aspects of the implementation in the clinical routine will be discussed and conclusions will be offered in the closing section.

2 Material and Methods

Data labeling procedure

Training a classifier requires the availability of a verified dataset providing a "ground truth" on the task to be learned. In the given case, a sufficient number of spectra were labeled in a visual inspection, both to obtain a training set for the design of the classifier and to test the algorithm on a second independent set of labeled data.

The two datasets for training and testing were acquired at the German Cancer Research Center ("dkfz"), Heidelberg, with a routine protocol in the pretherapeutic diagnostics of brain tumors and after therapy, using two 1.5T MR scanners (Magnetom Vision [training set] and Magnetom Symphony [test set]; Siemens Medical Solutions, Erlangen, Germany) with commercially available MRSI pulse sequences and the standard head coil (Siemens CP Head coil). The MR spectra were obtained with a double spin-echo sequence with water signal suppression and long echo time (2D PRESS, TR [pulse repetition time] 2000 ms, TE [echo time] 135 ms, 512 data points). The training dataset comprised 36 spectral images from 36 patients with a resolution of 24*24 voxels at a size of approx. 1 cm^3, interpolated to 32*32, a total of 36,864 spectra. The test set comprised 26 spectral images from 23 patients, acquired at different timepoints before therapy and during follow-up control, with multi-slice 2D spectroscopic imaging and a resolution of 16*16*8 voxels. Out of these datasets, 21 were used completely, while for one dataset only a subset of five slices (5*16*16 voxels) was available, and for the remaining four only one slice (16*16*1) was available, a total of 45 312 spectra. The approximate size of the voxels was 1 cm^3. Differences in the training and test sets were selected on purpose to investigate the methods ability to account for acquisition sequence variability. Approximately 20% of the spectra were acquired from within the PRESS box. For all spectra, resonances with chemical shift larger than $\delta = 3.5$ppm, in particular the water peak, were removed by HLSVD using jMRUI (22). All spectra were corrected for frequency shifts. Magnitude spectra from the spectral region between 0.5 ppm and 3.6 ppm were used as feature vectors for the pattern recognition. This region comprises resonances assigned to cholines (Cho), creatine (Cre), N-acetyl-aspartate (NAA), lactate, and lipids. The spectral feature vectors (of length 101) were normalized to unit area (L1-norm) in the region between 1.9 ppm and 3.4 ppm. For the spectral fitting, AMARES (as implemented in jMRUI 2.1 [2]) was used with a fixed shift between Cho, Cre, and NAA, as well as soft constraints on lactate (1.3 ± 0.3 ppm) and lipids (0.9 ± 0.5 ppm). No constraints were applied to the linewidth, and the initialization was chosen to be 10 Hz for Cho, Cre, and NAA, 30 Hz for lipids, and 20 Hz for lactate.

Both training and test datasets were labeled based on a visual inspection of each spectrum. The training data were labeled by one expert. Rejection criteria for a spectrum were the presence of artifacts (due to spurious echoes, water peaks, strong baseline distortions) and poor SNR of Cho, Cre, and NAA peaks. Spectra that could still be interpreted in terms of the diagnostic task remained

in the evaluable class even if their quality was poor, i.e., tumor spectra with high Cho intensities and vanishing NAA and vice versa for spectra of healthy tissue. This procedure assigned 2724 spectra to the high-quality/evaluable class (termed "nice" for short) and 34 140 spectra to the artifact/ non-evaluable class (termed "noise" for short). To ensure operator independence in the crucial definition of a "ground truth" for the test data, all 45 312 spectra were labeled twice and independently by two experts (one of them having labeled the training set), according to the same criteria as above. In total, 38 998 spectra were assigned to the "noise" class by both experts, and 5311 to the "nice" class. For a set of 1003 spectra – nearly all of ambiguous quality – no consensus was obtained, resulting in a third, "intermediate" class. As a consequence, two binary partitions were defined and statistically evaluated: first, "nice" versus "intermediate & noise" representing a conservative assessment of the data quality; second, "nice & intermediate" versus "noise" under less rigid requirements on the data quality.

Spectral quality assessment using the NoN score

The quality measure proposed in this work (termed *nice-or-noise* score, NoN) was obtained from a multivariate classifier, called "random forest" (23), which was trained to distinguish automatically between the spectral patterns of the "nice" and "noise" classes. Random forest is a recently proposed ensemble classifier based on decision trees. In contrast to conventional classification trees, a whole ensemble – or "forest" – of decision trees is trained, and their individual decisions are pooled in the decision process. To obtain this tree ensemble from a single dataset, the classifier uses "random splits" (23) to randomize over the input features in the induction of the tree. In addition, a bootstrapping – as in "bagging" (23) – is pursued in the training process, generating slightly perturbed training data subsets and thus allowing for a degree of independence among the trees in the ensemble. Advantages of this algorithm are its high performance, which is comparable to other popular multivariate classifiers such as support vector machines or neural networks and, more important, its ease of training (standard parameters often perform close to optimal). In the given task the feature vectors of the magnitude spectra (see data section above) were used as inputs, without further feature transformation or selection to a freely available implementation of random forests (24) (with standard parameter settings *ntree* = 500 and *mtry* = 10).

When applied to new data, the random forest is able to return a probability on the membership of either the "nice" or "noise" class. This probability, termed NoN score, is the quality measure proposed in the present work and used in the following.

Spectral quality assessment using established scores

For evaluation the classifier was applied to the second dataset. The NoN score was compared with standard spectral quality measures based on the SNR and

estimated CRBs. Although Cramér-Rao theory only provides lower bounds on the actual standard deviation (SD) of the estimated parameters, asymptotically the employed least squares approach (AMARES) is known to yield a minimum variance unbiased estimator (25). Assuming that the model assumptions underlying the line fitting are valid, this means that the CRBs are close to the true SDs and it is customary to use them as an estimate of these. Here, approximate CRBs were determined for the amplitudes of the Cho, Cre, and NAA resonances using jMRUI (estimating the noise SD from the residue). The results were divided by the absolute value of the respective amplitude, providing lower bounds on the normalized SDs. Since all three metabolites are deemed important in the detection of recurrent tumor (26), the minimum of the three values served as a Cramér-Rao quality measure (CR). As estimates on the CRBs are only provided by jMRUI when the spectral fitting succeeds for *all* defined metabolites (here: including lactate and lipids, relevant in the evaluation of tumor spectra), this measure could only be calculated for 26 747 voxels (59.0%, 3221/536/22 990 nice/intermediate/noise). The SNR was calculated from magnitude spectra as the quotient between the maximum intensity of either the Cho, Cre, or NAA resonance, and the SD of the signal in a spectral region distant from known metabolites, at chemical shift larger than $\delta = 6$ ppm.

Quantitative comparison of spectral quality scores

All three quality measures resulted in continuous output scores and had to be compared with a binary ground truth. For a quantitative comparison the following performance measures were applied to the results on the test data (using the statistical programming language R (27) with the ROCR package (28)): First, the receiver-operator-characteristic (ROC) was calculated, allowing determination of sensitivity (= true positives/all positives) and specificity (= true negatives/all negatives) for every possible threshold on the continuous score. Summarizing the curve by its integral and in a single number which can be compared more easily, the area-under-the-curve (AUC) was also calculated (AUC = 1 indicating perfect discrimination and AUC = 0.5 indicating no discrimination at all). Second, the convex precision-recall curves were determined. This measure is similar to the ROC curve, but allows focusing on the classification errors of one of the two classes (here: positives = "nice" or "nice & intermediate") using precision (= true positives/(true positives + false positives)) and recall (= true positives/(true positives + false negatives) = sensitivity). Finally, the maximum F-measure was determined, summarizing the precision-recall curve in a single number by calculating the evenly weighted harmonic mean of precision and recall (F = 2 * precision * recall/(precision + recall)). The measures were assessed on both binary partitions of the test set ("nice" vs. "intermediate & noise," "nice & intermediate" vs. "noise"), globally on the whole data and individually for each single acquisition, thus allowing assessing the inter-patient variability.

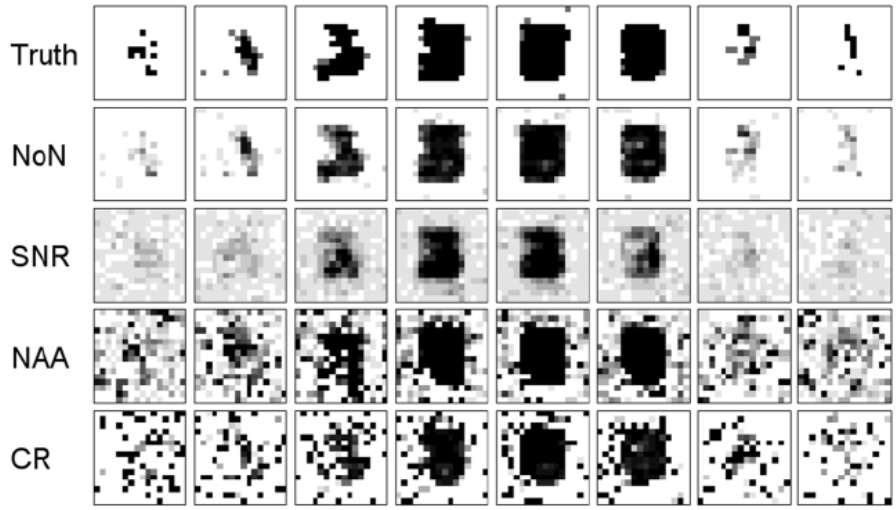

Figure 1: Experts' labels ("ground truth") and quality measures on exemplary test data. From left to right: adjacent MRSI slices of one 3D detection volume. From top to bottom: 1) ground truth, black corresponds to high quality ("nice"), white corresponds to low quality ("noise"), gray shows intermediate voxels where human experts disagreed; 2) NoN score proposed here; 3) SNR, black pixels have a value above 30%; 4) NAA line amplitudes from spectral fitting; 5) Cramér-Rao score (white indicates also: no standard deviation available, see text).

3 Results

An overview of the performance of the different measures is provided for an exemplary 3D MRSI detection volume of the test set (Fig. 1). In the central regions of the data volume that are properly excited by the pulse sequence, the ground truth maps indicate a high spectral quality, deteriorating toward lateral voxels (row 1). Spectra of the intermediate/ undecided quality class are found between high and low quality. Notably, the experts labels also indicate spectra of low quality within the excited volume (slices 3, 4), illustrating that regions of non-evaluable spectra also occur in thoroughly planned acquisitions with well-placed excited volumes. Regions that are assigned a high quality by the proposed nice-or-noise classification (NoN score displayed as confidence images, row 2) tend to follow the assessment of the human operator in many details (e.g., slices 2, 3). The SNR (row 3) also emphasizes these regions, but lacks the sharp contrast of the NoN score between regions of high quality and non-evaluable background (slices 2, 3, 7). Amplitudes from the spectral fitting and normalized SD of the amplitudes (rows 4, 5) allow for a rough localization of the evaluable parts of the spectral volume, but reject many of the spectra at the border of the high-quality areas. As visible in the amplitudes of NAA (row

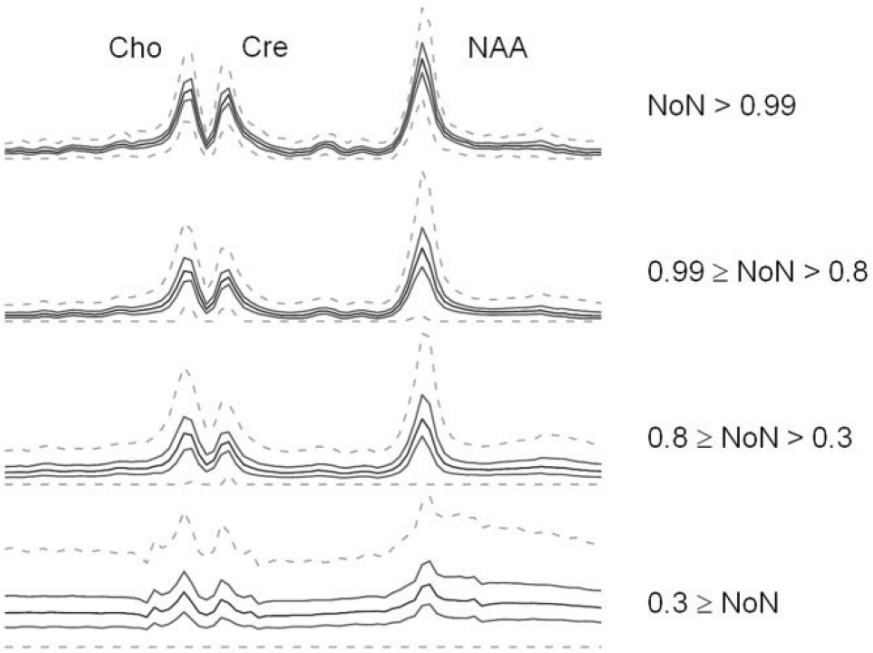

Figure 2: Average spectra of the test dataset: median (black) along with 25% and 75% quantiles (gray) and outliers (dashed). Peaks from left to right, Cho, Cre, and NAA. Spectra grouped according to the quality score proposed (NoN). From top to bottom: NoN > 0.99, $0.99 \geq$ NoN > 0.8, $0.8 \geq$ NoN > 0.3, $0.3 \leq$ NoN. All spectra from the last group are rejected as non-evaluable.

4), the failure of some of the spectral fits results in a number of voxels without confidence measure when using the CR rule (row 5) and thus diminishes the diagnostic value of the measure in the given setting.

Although the MRSI detection volume of Fig. 1 is a representative example of the test set, each of the different measures can be studied qualitatively in more detail. When grouping the spectra of the test set according to the NoN score (Fig. 2), high scores coincide with a stable spectral pattern showing low variance (top). The average pattern in the lowest quality group ("noise," rejected) is obscured by noise and artifacts (Fig. 2, bottom). Figure 3 visualizes the estimated Cramér-Rao lower bound on the error on the amplitudes, normalized by the absolute value of the amplitude for each spectrum (14). Evaluable and non-evaluable data separate, although the rigid ratio-rule with a threshold of 30% or more results in a high number of erroneously accepted spectra. The figure demonstrates that the application of a more flexible decision rule, e.g., a nearest-neighbor classification on the "nice" data, will yield a considerably higher specificity. The scatter-plots (Figs. 3 and 4) allow comparing the SNR

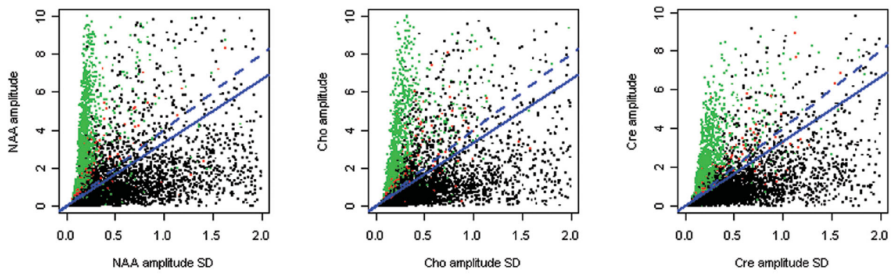

Figure 3: CR ratio rule for NAA, Cho, and Cre. Each point in the plot represents amplitude (vertical axis) and corresponding SD (horizontal axis) of one spectrum in the test set. The colors reflect the operators assessment of the data quality (green: "nice," black: "noise," red: intermediate/undecided). The blue lines visualize the recommendation to reject spectra with a ratio of more than 25% (dashed line) and 30% (solid lines), respectively; in the figure these spectra are below the blue lines.

and CR ratio against the NoN score. Here, spectra can easily be separated according to NoN score and SNR, whereas the CR ratio of the two classes shows considerable overlap both in the scatter-plot and when studying the corresponding marginal distributions of the different groups (Fig. 5).

The ROCs and precision-recall-curves are statistical rank-order measures that summarize the different distributions under objective criteria (Fig. 6). AUCs and F-values provide concise summaries of these curves (Table 1). For all three methods it is easier to separate the high quality group ("nice") from intermediate and noisy spectra than to separate both high and intermediate quality data from noise (Fig. 6; Table 1). No general differences can be observed between the full dataset and the subset of successfully fitted spectra, or between the full dataset and the subset of spectra from within the PRESS box. The ROCs of the different measures show that the SNR and CR perform similarly well, while the NoN score performs better than either. The same ranking of the three methods can be observed when studying the precision-recall curve that quantifies the recovery of the high-quality data (Fig. 6). If a false rejection of, e.g., 5% evaluable spectra (nice & intermediate) from within the PRESS box is tolerated (recall 95%, dotted line in the right box of Fig. 6), this will result in 4% falsely accepted noise spectra (96% precision) when relying on the NoN score. The false acceptance rate is 15for SNR and 19.5% for the CR rule. These results stem from a pooled analysis of the data of all patients; applying the same analysis to single patients allows assessing the inter-patient variation and the robustness of the measures (Figs. 7 and 8). For both binary separations of the test dataset ("nice" vs. "intermediate & noise," "nice & intermediate" vs. "noise"), the ROCs of the NoN score reveal a high sensitivity for nearly all of the 26 test datasets. In a few cases the NoN score is able to separate the

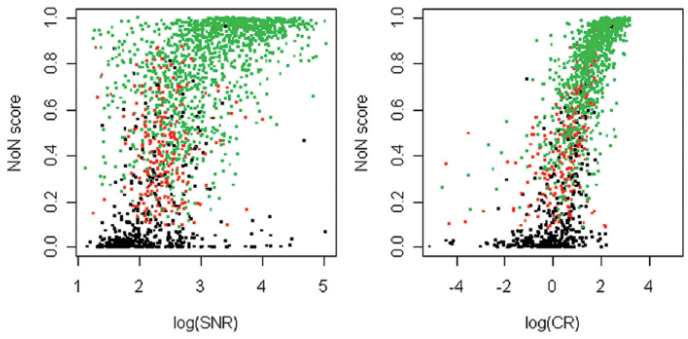

Figure 4: Scatter-plot of quality score (NoN) vs. SNR (left) and CR ratio (right). Shown are spectra from within the PRESS box only. Green dots indicate "nice" spectra, red dots represent intermediate/undecided, and black dots correspond to "noise" spectra. The separation according to either SNR or NoN score is clearer than according to the CR ratio, as evidenced by the estimated marginal distributions shown in Fig. 5.

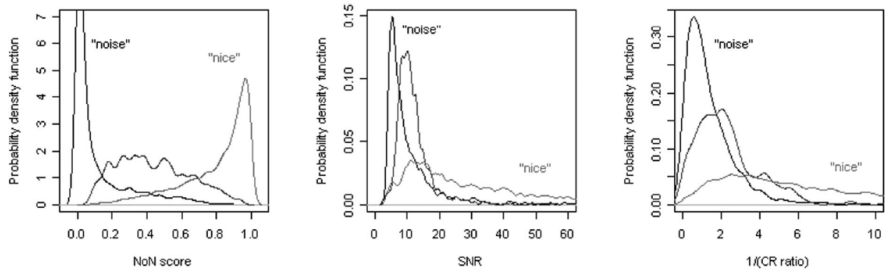

Figure 5: Separation of the three quality groups according to the different measures showing the estimated marginal distributions of the different groups ("nice," undecided/intermediate, "noise") on the vertical axis ("probability density function"). Shown are spectra from within the PRESS box only. The probability density function of the NoN function has been truncated in the left box.

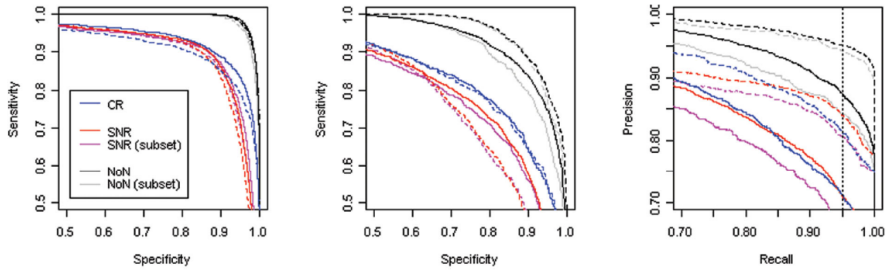

Figure 6: Performance of the three quality measures. ROC of all spectra (left), ROC of spectra from within the PRESS box (middle), and precision-recall-curve of the same subset (right). The solid lines indicate performance for the classification of "nice" vs. "intermediate & noise," dashed lines indicate "nice & intermediate" vs. "noise." The dotted vertical line is discussed in the results section. All curves are summarized by the measures in Table 1. The NoN score performs better than the other two measures both on the whole data and on the spectra within the PRESS box.

data perfectly. SNR and CR perform well on some datasets, but have a weaker performance both on average and for most individual patients (as visible in the density plot of the AUCs, Fig. 8).

4 Discussion

Fields of application

The proposed NoN score is a measure to automatically control the input of a spectral analysis. In pattern recognition, which is attributed to a high potential in certain applications (9-11,29), a control of the spectral signal is a necessary element to ensure the reliability of the diagnostic process. Here the proposed quality measure might be of the highest relevance. Considering the similarity in the methodology of diagnostic pattern recognition and NoN score, the latter can be seen as an extension of the former.

A multi-class decision system, such as in Ref. (15), could easily be extended by a "noise" class. In a spectral fitting the estimated Cramér-Rao lower bound on the error of the amplitudes provides a confidence measure on the output of the processing. However, it is only an assessment of a lower bound on the error, while the real error can be of arbitrary magnitude, if the spectrum cannot be modeled appropriately by the spectral model used, i.e., if artifacts or other deviations of the model were present in the data. This might also be a possible reason for the comparatively weak performance of the CR rule on the given dataset. To circumvent this problem the additional application of the proposed artifact recognition can be used to automatically flag spectra that do not correspond to the expected spectral pattern, and thus to increase the robustness of

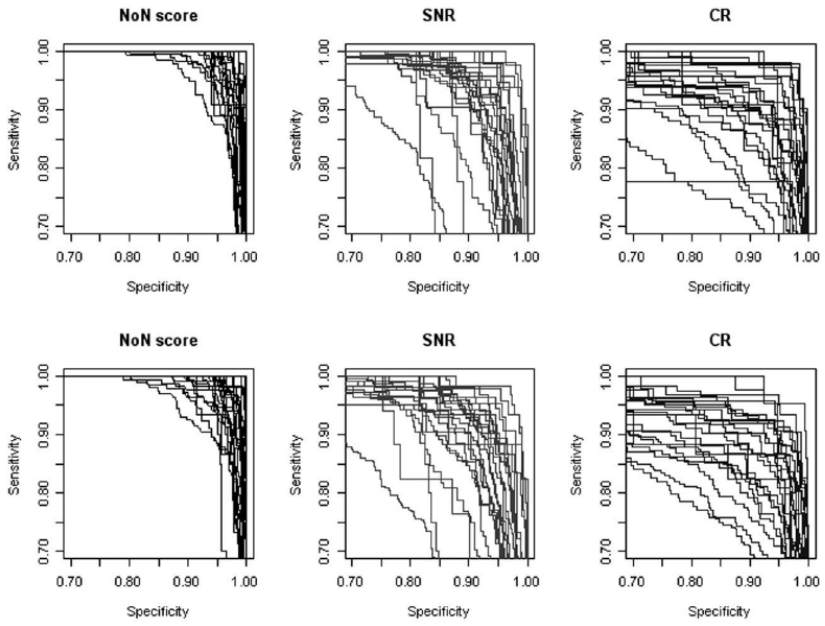

Figure 7: ROCs for single patients (N = 26). Top: "nice" vs. "intermediate & noise." Bottom: "nice & intermediate" vs. "noise." Quality measures as indicated.

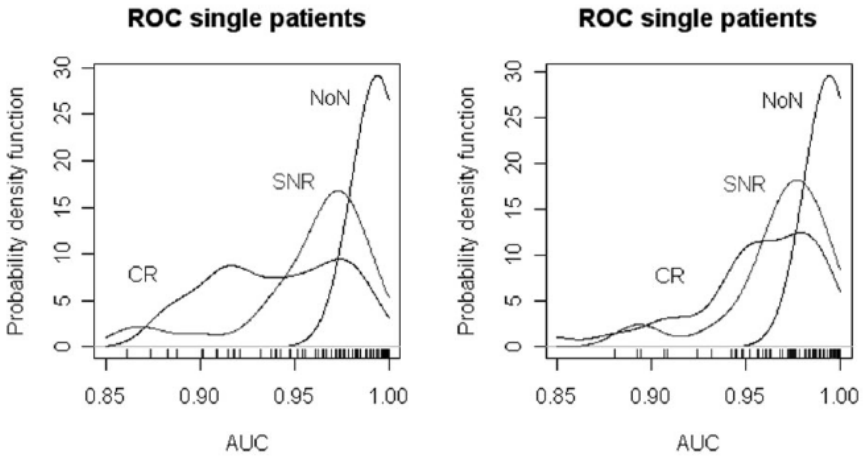

Figure 8: AUCs of ROCs for single patients: distribution of NoN score, SNR, and CR rule. Left: "nice" vs. "intermediate & noise." Right: "nice & intermediate" vs. "noise." Graphs indicate kernel density estimates (kernel width = 0.05), while vertical lines below the graphs indicate single observations, corresponding to one patient each.

a subsequent spectral fitting and the reliability of its output.

In applications in which a high degree of automation is not required, the NoN score can be displayed as a standard confidence image (Fig. 1). As a probability, the NoN score is always bound between zero and one and can thus easily be displayed with a fixed color scale, in contrast to the other measures which often range over several orders of magnitude (see Figs. 4 and 5) and need further transformations for an optimal display (e.g., truncation, scaling (21), two-dimensional color-maps, and other operator-dependent actions). The NoN score provides one comprehensive confidence value resulting in a single image, which is easier to represent and survey than a whole set of different maps, displaying parameters like SNR plus linewidth, or the estimated CR bounds on several parameters of the resonance line model (17,21).

Generalization of the approach

Although we have relied on one particular classifier – random forests – we also expect that other, preferably nonlinear, classifiers such as mixture models, neural networks, or support vector machines will do as well. The collection of a representative test set with a high number of spectra is mandatory for any application, although it is neither as difficult as the preparation of such a collection for a diagnostic classification, nor overly time-consuming. Keeping acquisition parameters such as TE and TR fixed, it was easily possible to change the instrumentation (from Magnetom Vision to Magnetom Symphony), indicating a

robustness of the pattern classification against some parameter changes. Changing acquisition parameters which only changed the quality of the spectral signal did neither affect the process of training the classifier nor its general applicability. It was possible, for example, to have different point spread functions during the acquisitions (24*24 resolution during training, 16*16 for the test) or to interpolate the spatial resolution after the acquisition of the data (from 24*24 to 32*32 in the training set). Such operations will, however, affect the resulting NoN score in the same way that they affect the general quality of the spectral signal.

Changes in acquisition parameters which modify the general appearance of the spectral pattern, on the other hand, may require the collection of a new set of training samples and retraining the classifier. Such a retraining under different clinical settings (30) was performed to obtain a quality assessment of long echo time (TE 135 ms) 1H-MR spectra of the human prostate, resulting in a quality score that performed as well as the one in the present study. An evaluation of the quality classification on short echo time data is open to further studies, although general limitations of the proposed approach are not expected (compare Refs. 9,31).

Automated processing

The ultimate goal remains to obtain the metabolic information of MR spectra in a completely automated fashion, in order to enable the integration of MRSI analysis with the standard set of techniques for routine clinical diagnostics (1). This requires high robustness and reliability of the applied MRS evaluation routines and demands high performance in the monitoring of the procedures. In the present test set the proposed quality score did not reach 100% accordance with the experts' labels (although the accuracy of "ground truth" always remains questionable to some degree when derived from human operators). However, the precision-recall plot indicates (Fig. 6) that a readjustment of the acceptance threshold on the quality score allows trading between rejected data volume and quality of the remaining data. First results on data acquired under the same protocol as in the present study, using a pattern recognition approach (11) to detect the presence of recurrent brain tumor, indicate that this tradeoff increases the reliability of an automated processing. Here, an automated pattern recognition had been trained on the data of an earlier medical study (26) using standard chemometric methods (partial least-squares regression) to discriminate between the spectral patterns of different tissue classes. In a test, this classifier was applied to another set of 269 spectra (of 31 patients), labeled according to follow-up examinations as "normal" (151 spectra) or "tumorous" (118 spectra, including tumor border). The automated classifier reached an accuracy of 93.3% on the full test dataset, but a rejection of 60 spectra labeled as non-evaluable by the NoN score (34/26 normal/tumor) increased the overall performance to 100% (32). In this case misclassification between the tissue types occurred on low-quality spectra only, and these were successfully removed by using a predefined threshold on the NoN score.

Implementation and extensions

The central idea of the presented approach is to classify the quality of a spectrum by an algorithm that has learned its decisions from previously labeled data. This idea can be extended into several directions: The present approach checks the pattern of the spectrum and thus the input to the processing. Alternatively, a classifier could be used to check for remaining peaks or structures in the output of the processing, i.e., in the residue of a spectral fitting (33). Obviously, this procedure would allow indirectly evaluating whether the line model of the spectral fitting was chosen appropriately. So far, this evaluation is a major reason for the tedious visual inspection of a spectrum and its line fit. Also, as mentioned at the beginning, standard quality measures like SNR, or linewidth and CR bound of the spectral fitting, could easily be integrated into an automated quality classification: Either all measures – SNR, linewidth, CR bound, and NoN score of the spectral pattern – could be used as input to a second classifier, or SNR, linewidth, CR bound, and so forth could be used to augment the spectral pattern, thus adding extra dimensions to the feature vector of the space in which the nonlinear classifier operates. As mentioned above (Fig. 3; Results section), a classifier which automatically checks the outcome of a spectral fitting – amplitudes and standard deviations – for plausibility on a certain diagnostic task without a classification of the spectral pattern itself, can be designed as well. Once multi-modal decision support systems become available (34), external information, e.g., originating from other MR imaging modalities, can also be considered in the test for plausibility of the spectral information, and decision rules on the data quality as implemented in Ref. (34) can be extended. Finally, it is possible to extend the basic "nice" versus "noise" decision to a more detailed classification of "nice" versus "noise (lipids)"/"noise (unremoved peaks)"/"noise (low signal intensity)." Training sets for all of these classifiers are either available from the library of the decision support system (15) or can be acquired with limited effort specifically for a diagnostic routine at any clinical center.

5 Conclusion

The approach of using a classifier in the quality control of MRSI data allows to rank the spectra according to their quality (Fig. 5). When displayed in confidence images, the proposed NoN score has a high contrast as a result of a superior class separation between high- and low-quality data, and mimics the decisions of the human operator in great detail (Fig. 1). In the quantitative comparison against standard measures (CR, SNR), the proposed NoN score could recover high-quality data at a lower number of false positives, and performed better than the other approaches both globally and in a per-patient evaluation.

For pattern recognition in particular, the proposed approach allows designing rules for an automated quality assurance that allows for more complex decisions than a simple threshold operation on linewidth and SNR. In a spectral fitting

the proposed quality score could be used to automatically separate data that are deemed non-evaluable by a human operator and thus to increase the reliability of the output. A combination of the proposed pattern recognition on the spectral pattern with quality measures such as CR, SNR, or linewidth is possible in both cases.

The presented algorithm has the potential to automate technical overhead and to ease the work of the human operator significantly. In a fully automated processing it can help to narrow the evaluation to data that are safe to interpret and thus are likely to be processed without error.

References

1. Maudsley AA. Can MR spectroscopy ever be simple and effective? Am J Neurorad 2005;26:2167.

2. 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;129:35-43.

3. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. Magn Reson Med 1993;30:672-679.

4. 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;18:1-13.

5. Soher BJ, Young K, Govindaraju V, Maudsley AA. Automated spectral analysis III: application to in vivo proton MR spectroscopy and spectroscopic imaging. Magn Reson Med 1998;40:822-31.

6. Jansen JF, Backes WH, Nicolay K, Kooi ME. 1H MR spectroscopy of the brain: absolute quantification of metabolites. Radiology 2006;240:318- 32.

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

8. Tate AR, Majo s C, Moreno A, Howe FA, Griffith JR, Aru s C. Automated classification of short echo time in vivo 1H brain tumor spectra: a multicenter study. Magn Reson Med 2003;49:29-36.

9. Devos A, Lukas L, Suykens JA, Vanhamme L, Tate AR, Howe FA, Majos C, Moreno-Torres A, van der Graaf M, Aru s C, van Huffel S. Classification of brain tumours using short echo time 1H MR spectra. J Magn Reson 2004;170:164-75.

10. Laudadio T, Pels P, de Lathauwer L, van Hecke P, van Huffel S. Tissue segmentation and classification of MRSI data using canonical correlation analysis. Magn Reson Med 2005;54:1519-29.

11. 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 2006:19:599-60.

12. Saitta L, Neri F. Learning in the "real world." Machine Learn 1998;30: 133-163.

13. Brodley CE. The need for diagnostics for classification algorithms. International Conference on Machine Learning Research (ICML), 1997. Workshop on Machine Learning Applications in the Real World; Methodological Aspects and Implications, Nashville, Tennessee, July 8-12, 1997.

14. Kreis R. Issues of spectral quality in clinical 1H magnetic resonance spectroscopy and a gallery of artifacts. NMR Biomed 2004;17:361-381.

15. Tate AR, Underwood J, Acosta DM, Julia'-Sape M, Majo s C, Moreno-Torres A, Howe FA, van der Graaf M, Lefournier V, Murphy MM, Loosemore A, Ladroue C, Wesseling P, Bosson JL, Caban as ME, Simonetti AW, Gajewicz W, Calvar J, Capdevila A, Wilkins PR, Anthony Bell B, Remy C, Heerschap A, Watson D, Griffiths JR, Aru s C. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed 2006;19:411-434.

16. Cavassila S, Deval S, Huegen C, van Ormondt D, Graveron-Demilly D. Cramér-Rao bounds: an evaluation tool for quantitation. NMR Biomed 2001;14:278-83.

17. Young Y, Khetselius D, Soher BJ, Maudsley AA. Confidence images for MR spectroscopic imaging. Magn Reson Med 2000;44:537-545.

18. Vanhamme L, Lemmerling P, van Huffel S. Comments on "Confidence Images for MR Spectroscopic Imaging" by Young, Khetselius, Soher and Maudsley. Magn Reson Med 2001;46:1254-56.

19. Ebel A, Soher BJ, Maudsley AA. Assessment of 3D proton MR echoplanar spectroscopic imaging using automated spectral analysis. Magn Reson Med 2001;46:1072-1078.

20. Elster C, Schubert F, Link A, Walzel M, Seifert F, Rinneberg H. Quantitative magnetic resonance spectroscopy: Semiparametric modeling and determination of uncertainties. Magn Reson Med 2005;53:1288- 1296.

21. Jiru F, Skoch A, Klose U, Grodd W, Hajek M. Error images for spectroscopic imaging by LCModel using Cramér Rao bounds. MAGMA 2006; 19:1-14.

22. Naressi A, Couturier C, Devos JM, Janssen M, Mangeat C, de Beer R, Graveron-Demilly D. Java-based graphical user interface for the MRUI quantitation software package. MAGMA 2001;12:141-152.

23. Breiman L. Random forests. Machine Learn J 2001;45:5-32.

24. Liaw A, Wiener M. Classification and regression by random Forest. R News 2001;2:18-22.

25. Seber GAF, Wild CJ. Nonlinear regression. New York: John Wiley & Sons; 1989.

26. Schlemmer HP, Bachert P, Herfarth KK, Zuna I, Debus J, van Kaick G. Proton MR spectroscopic evaluation of suspicious brain lesions after stereotactic radiotherapy. AJNR Am J Neuroradiol 2001;22:1316-1324.

27. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat 1996;5:299-314.

28. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940-3941.

29. Nelson S. Novel approaches to spectral processing and quantification. In: ISMRM 14th Scientific Meeting, Seattle, Washington, 6-12 May 2006.

30. Kelm BM, Menze BH, Zechmann CM, Baudendistel KT, Hamprecht FA. Automated estimation of tumor probability in prostate magnetic resonance spectroscopic imaging: pattern recognition vs quantification. Magn Reson Med 2007;57:150-159.

31. Lukas L, Devos A, Suykens JA, Vanhamme L, Howe FA, Majos C, Moreno-Torres A, van der Graaf M, Tate AR, Aru s C, van Huffel S. Brain tumor classification based on long echo proton MRS signals. Artif Intel Med 2004;31:73-89.

32. Menze BH, Kelm BM, Heck D, Lichy MP, Hamprecht FA. Machine based rejection of low-quality spectra, and estimation of brain tumor probabilities from magnetic resonance spectroscopic images. In: Handels H, Ehrhardt J, Horsch A, Meinzer HP, Tolxdorf T, editors. Proceedings of the BVM-Workshop, Informatik aktuell. New York: Springer; 2006. p 31-35.

33. Guha R, Jurs PC. Determining the validity of a QSAR model – a classification approach. J Chem Inf Model 2005;45:65-73.

34. Maudsley AA, Darkazanli A, Alger JR, Hall LO, Schuff N, Studholme C, Yu Y, Ebel A, Frew A, Goldgof D, Gu Y, Pagare R, Rousseau F, Sivasankaran K, Soher BJ, Weber P, Young K, Zhu X. Comprehensive processing, display and analysis for in vivo MR spectroscopic imaging. NMR Biomed 2006;19:492-503.