Classifying Multiple Sclerosis courses by combining clinical data with Magnetic Resonance metabolic features and lesion loads

Adrian Ion-Mărgineanu [1,2,3], Gabriel Kocevar[1], Claudio Stamile [1,2,3], Diana M Sima[2,3,4], Françoise Durand-Dubief[1,5], Sabine Van Huffel[2,3], and Dominique Sappey-Marinier[1,6]

[1] CREATIS CNRS UMR5220 & INSERM U1206; Université de Lyon, Université Claude Bernard-Lyon 1, INSA-Lyon, Villeurbanne, France
[2] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium
[3] imec, Leuven, Belgium
[4] ico**metrix**, R&D department, Leuven, Belgium
[5] Service de Neurologie A, Hôpital Neurologique, Hospices Civils de Lyon, Bron, France
[6] CERMEP - Imagerie du Vivant, Université de Lyon, Bron, France

Abstract.

Purpose. The purpose of this study is classifying multiple sclerosis (MS) patients in the four clinical forms as defined by the McDonald criteria using machine learning algorithms trained on clinical data combined with lesion loads and magnetic resonance metabolic features.

Materials and Methods. Eighty-seven MS patients (12 Clinically Isolated Syndrome (CIS), 30 Relapse Remitting (RR), 17 Primary Progressive (PP) and 28 Secondary Progressive (SP)) and eighteen healthy controls were included in this study. Longitudinal data available for each MS patient included clinical (e.g. age, disease duration, Expanded Disability Status Scale), conventional magnetic resonance imaging and spectroscopic imaging. We extract N-acetyl-aspartate (NAA), Choline (Cho), and Creatine (Cre) concentrations, and we compute three features for each spectroscopic grid by averaging metabolite ratios (NAA/Cho, NAA/Cre, Cho/Cre) over good quality voxels. We built linear mixed-effects models to test for statistically significant differences between MS forms. We test nine binary classification tasks on clinical data, lesion loads, and metabolic features, using a leave-one-patient-out cross-validation method based on 100 random patient-based bootstrap selections. We compute F1-scores and BAR values after tuning Linear Discriminant Analysis (LDA), Support Vector Machines with gaussian kernel (SVM-rbf), and Random Forests.

Results. Statistically significant differences were found between the disease starting points of each MS form using four different response variables: Lesion Load, NAA/Cre, NAA/Cho, and Cho/Cre ratios. Training SVM-rbf on clinical and lesion loads yields F1-scores of 71-72% for CIS vs. RR and CIS vs. RR+SP, respectively. For RR vs. PP we obtained good classification results (maximum F1-score of 85%) after training LDA on clinical and metabolic features, while for RR vs. SP we obtained slightly higher classification results (maximum F1-score of 87%) after training LDA and SVM-rbf on clinical, lesion loads and metabolic features.

Conclusions. Our results suggest that metabolic features are better at differentiating between relapsing-remitting and primary progressive forms, while lesion loads are better at differentiating between relapsing-remitting and secondary progressive forms. Therefore, combining clinical data with magnetic resonance lesion loads and metabolic features can improve the discrimination between relapsing-remitting and progressive forms.

Keywords: multiple sclerosis, longitudinal analysis, magnetic resonance spectroscopic imaging, lesion load, machine learning

1. Introduction

Multiple sclerosis (MS) is an inflammatory disorder of the brain and spinal cord in which focal lymphocytic infiltration leads to damage of myelin and axons [1]. MS affects approximately 2.5 million people worldwide, with an onset age commonly between 20 and 40 years, and an incidence more than twice as high in women compared to men [2].

The majority of MS patients (85%) usually experience a first attack defined as Clinically Isolated Syndrome (CIS), and will develop a relapsing-remitting (RR) form [3]. Two thirds of the RR patients will develop a secondary progressive (SP) form, while the other third will follow a benign course [4]. The rest of MS patients (15%) will start directly with a primary progressive (PP) form.

The criteria to diagnose MS forms was originally formulated by McDonald in 2001 [5] and revised by Polman in 2005 [6] and 2011 [7]. They all rely on using conventional magnetic resonance imaging techniques (MRI) such as T1-weighted, gadolinium-enhanced T1-weighted MRI, as well as T2-weighted and FLAIR, due to a high sensitivity for visualizing MS lesions. Conventional MRI is also used for quantifying lesion load (LL), a marker of inflammation process but only a moderate predictor of MS evolution [8].

More recently, advanced magnetic resonance techniques such as ^{1}H-Magnetic Resonance Spectroscopic Imaging (MRSI), Diffusion Tensor Imaging (DTI) and Magnetization Transfer Imaging have been shown [9] to provide a better characterization of the normal appearing white matter (NAWM) and thus a better understanding of the pathological mechanisms. Decrease of N-acetyl-aspartate (NAA) was observed in both chronic lesions and NAWM, reflecting a neuronal integrity loss [9]. Choline (Cho) and Creatine (Cre) contents were found to be increased in WM lesions and in NAWM, indicating the presence of severe demyelination, and cell proliferation in relation with inflammatory processes [10, 11].

In this study, we investigate the added value of combining routinely acquired clinical MS data (e.g. patient age, disease duration (DD), Expanded Disability Status Scale (EDSS)) with advanced magnetic resonance features (e.g. lesion loads (LL) and three metabolic features: NAA/Cho, NAA/Cre, Cho/Cre). To this purpose, we build multiple binary classifiers to automatically discriminate between different clinical forms of MS patients, by training each classifier on combinations of clinical data, lesion loads and metabolic features.

2. Materials and Methods

2.1. Patient population

Eighty-seven MS patients (12 CIS, 30 RR, 28 SP and 17 PP) were included in this study, while 18 volunteers without any neurological disorders served as healthy control (HC) subjects. Diagnosis and disease course were established according to the McDonald criteria [5,12]. This prospective study was approved by the local ethics committee (CPP Sud-Est IV) and the French national agency for medicine and health products safety (ANSM) and written informed consents were obtained from all patients and control subjects prior to study initiation. More details for each MS group, such as average age at first scan, average disease duration, median EDSS and average lesion loads can be found in Table 1.

2.2. Longitudinal MS data

The MS patients involved in this study were scanned multiple times over a different period for each patient, ranging from 2.5 to 6 years. The minimum number of scans is 3, while the maximum is 10. The gap between two consecutive scans is either 6 months

	CIS	RR	PP	SP
Number of patients (Male/Female)	12 (6/6)	30 (6/24)	17 (6/11)	28 (17/11)
Age at first scan [years]	31.8 (6.4)	33.2 (7)	39.5 (6)	41.1 (4.8)
Disease duration [years]	2.9 (1.9)	8.3 (4.8)	7.5 (2.9)	14.9 (6.1)
EDSS median [range]	1 (0-4)	2 (0-5.5)	4 (2-7.5)	5 (3-8.5)
Lesion Load [ml]	6.6 (3.5)	16.7 (12.6)	20.8 (13)	31 (12.9)
Total number of scans	62	226	125	206

Table 1: Patient population: Age - average value (standard deviation); Disease duration - average value (standard deviation); EDSS - median (minimum - maximum); Lesion Load - average value (standard deviation).

or 1 year. In total there are 619 MS scans, but because of missing lesion loads and metabolic features, there are 592 (95.6%) scans with full complete data, leading to an average of 6-7 complete scans/patient.

2.3. MRI acquisition and processing

All patients and control subjects underwent MR examination using a 1.5 Tesla MR system (Sonata Siemens, Erlangen, Germany) and an 8 elements phased-array head-coil.

2.3.1. Conventional MRI

Conventional MRI protocol consisted of a 3 dimensional T1-weighted (magnetization prepared rapid gradient echo-MPRAGE) sequence with repetition time/echo time/time for inversion (TR/TE/TI)= 1970/3.93/1100 ms, flip angle= 15°, matrix size= 256×256, field of view (FOV)= 256×256 mm, slice thickness= 1 mm, voxel size= $1 \times 1 \times 1$ mm, acquisition time= 4.62 min, and a fluid attenuated inversion recovery (FLAIR) sequence with TR/TE/TI= 8000/105/2200 ms, flip angle= 150°, matrix size= 192×256, field of view (FOV)= 240×240 mm, slice thickness= 3 mm, voxel size= $0.9 \times 0.9 \times 3$ mm, acquisition time= 4.57 min.

2.3.2. MRSI acquisition

MRSI data was acquired from one slice of 1.5 cm thickness, placed above the corpus callosum and along the anterior commissure - posterior commissure (AC-PC) axis, encompassing the centrum semioval region, and took 5 minutes and 20 seconds. A point-resolved spectroscopic sequence (PRESS) with TR=1690 ms and TE=135 ms was used to select a volume of interest (VOI) of $105 \times 105 \times 15$ mm^3 during the acquisition of 24×24 (interpolated to 32×32) phase-encodings over a field of view (FOV) of 240×240 mm^2.

2.3.3. MRSI processing

MRSI data processing was performed using SPID [13, 14] in MatLab 2015a (MathWorks, Natick, MA, USA). AQSES-MRSI [15,16] was used to quantify *N*-acetyl-aspartate, Choline (Cho), and Creatine (Cre), using a synthetic basis set. The basis set incorporates prior knowledge of the individual metabolites in the quantification procedure. MPFIR (maximum-phase finite impulse response) filtering [17] was included in the AQSES-MRSI procedure for residual water suppression, with a filter length of 50 and spectral range from 1.9 to 3.4 ppm. A band of two voxels at the outer edges of each VOI was discarded in order to avoid chemical shift displacement artifacts and lipid contamination artifacts.

2.3.4. Quality control

After quantifying metabolites from all MRSI grids, a quality control was performed. Voxels with Cramer-Rao Lower Bounds (CRLBs) lower than 10% for *each* of the three metabolites (NAA, Cho, and Cre) were kept as having "good quality" to perform feature extraction. If the number of "good quality" voxels is lower than 50% of the total amount of voxels in the MRSI grid, then the acquisition is discarded. All 18 Control subjects had MRSI data with a number of "good quality" voxels higher than 50% of the total amount of voxels, and 606 out of 619 (97.9%) MRSI data from MS patients had good quality as defined earlier.

2.4. Feature extraction

In this study we use three types of features: clinical (e.g. patient age, disease duration, and EDSS), lesion loads, and metabolic features. The clinical features are routinely acquired in the hospital. The lesion loads were computed based on T1 and FLAIR, using the MSmetrix software [18] developed by ico**metrix** (Leuven, Belgium). The computation of metabolic features was performed in two steps: three metabolic ratios (NAA/Cho, NAA/Cre, Cho/Cre) were computed for each "good quality" voxel and then averaged, leading to three metabolic features extracted from each MRSI grid.

2.5. Training approach

Nine binary classification tasks were studied: HC vs. CIS, HC vs. RR, HC vs. PP, HC vs. RR+SP, HC vs. PP+SP, CIS vs. RR, CIS vs. RR+SP, RR vs. PP, RR vs. SP. The first three tasks investigated differences between HC and the starting MS forms (CIS, RR, and PP). The next task investigated differences between HC and MS patients that are likely to evolve or had evolved into secondary progressive form (RR+SP). Afterwards, we investigated differences between HC and definite progressive forms (PP+SP). The next two tasks investigated differences between CIS patients and the most likely progression of CIS, namely RR and RR+SP. From a neurological point

of view, the last two tasks were the most intriguing, as they were discriminating between the most common inflammatory MS form (RR) and the two progressive forms, PP and SP.

For each task, data normalization was performed. We use a leave-one-patient-out cross-validation (LOPOCV) setup combined with 100 random patient-based bootstrap selections for the training set. In this way, the test set has all data points of one patient, while the training set has $n-1$ data points corresponding to $n-1$ patients, where n is the total number of patients, different for each classification task (e.g. for HC vs. CIS, $n = 30$). Basically, to construct the training set, we randomly select one data point from each patient assigned to the training set. The test set always includes all data points of the test patient. We repeat the procedure 100 times and store the results. Each data point from the test set will be assigned 100 times to either class 1 or class 2, and in the end it will be assigned to one of the classes according to majority voting. This procedure is repeated until all patients from each classification task have been tested.

By using this random patient-based bootstrap selection, the two classes in the training set have a more balanced distribution of points (18 HC, 12 CIS, 30 RR, 17 PP, 28 SP), compared to using the total number of points of each class (18 HC, 61 CIS, 214 RR, 121 PP, 196 SP).

2.6. Performance measures and statistical testing

For each task, we computed and reported the F1-score and balanced accuracy rate (BAR) in percentage. F1-scores were computed for the positive class, which in our case was the first class from each of the nine binary classification tasks: HC for the first 5 tasks, CIS for the 6th and 7th tasks, and RR for 8th and 9th tasks. F1-score is defined as the harmonic mean between recall (number of correctly classified positive results divided by the total number of positive results) and precision (number of correctly classified positive results divided by the total number of results classified as positive). BAR is defined as the average between sensitivity (or recall) and specificity (or true negative rate).

In order to correctly assess if there are significant differences between the four MS groups, we built several linear mixed effects models which were able to incorporate the temporal evolution of each patient's MS course. We used five fixed effects and two random effects. The fixed effects are: MS course, gender, disease onset age, disease duration, and the interaction between MS course and disease duration. The random effects are set for each patient allowing an individual starting point and an individual disease evolution. The most interesting fixed effect for this study is the first one, which represents the average of the response variable at the beginning of the MS course, or when 'disease duration' = 0. We built four linear mixed effects models, one for each response variable: NAA/Cho, NAA/Cre, Cho/Cre, and lesion load. All statistical models were built in the 'R' software using the "lme4" package [19], statistical testing was done using the "lmerTest" package [20] and post-hoc analysis was done using the

"multcomp" package [21]. All tests were done for a significance level (α) of 0.05.

2.7. Classifiers

Three supervised classifiers implemented in Python 2.7.11 with scikit-learn 0.17.1 [22] have been used: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forest (RF). We tuned each classifier's parameters by optimizing the F1-score over a 5-fold cross validation on the training set within a grid search of individual parameters, specified further for each particular classifier. Because the classes were not perfectly balanced, we also added prior information to automatically adjust the classifiers' parameters according to this unbalance.

Fisher's LDA [23] is a classifier that finds a linear combination of input features that best separates the two classes. The Python implementation of LDA allows for three different solvers: singular value decomposition, least squares solution, and eigenvalue decomposition, which are deemed appropriate in various situations. Shrinkage is allowed for the last two solvers, therefore we tuned over the three type of solvers combined, where possible, with shrinkage varying from 0 to 1 in steps of 0.1. To automatically adjust for class unbalance, we set the *priors* parameter equal to class probabilities.

SVM [24,25] is among the most popular machine learning models. Given a training set with points from two classes, SVM tries to find the best hyperplane to differentiate between the two types of points. It can be used in the original feature space or the points can be mapped to another space by using kernel transformations. In this study we use SVM with a radial basis function kernel (SVM-rbf), which can be defined by two parameters: C, or the misclassification cost, and γ, which can be seen intuitively as the inverse of a support vector's radius of influence. The Python implementation of SVM-rbf allows tuning values of C and γ, therefore a logarithmic grid search was performed between 0.00001 and 100000 for both C and γ. To automatically adjust for class unbalance, we set the *class_weight* parameter to *balanced*.

Random Forests [26] is one of the most famous classifiers belonging to the ensemble learning family. Ensemble learning involves the combination of several weak models to solve a single prediction problem. RF is based on a group of decision trees: when a new testing point is given to the forest, each decision tree independently classifies the new point. The output of the forest is the class voted by the majority of all trees. One of the most important parameters of a forest is the number of decision trees, which we select after tuning between 200, 400, 600, 800, and 1000. To automatically adjust for class unbalance, we set the *class_weight* parameter to *balanced_subsample*. In this way, each tree's weights are computed inversely proportional to class frequencies based on the bootstrap sample selected for that tree.

3. Results

Figure 1 shows boxplots comparing MR metabolic features (A, B, C) and lesion loads (D) extracted from HC and each MS course. Boxplots are drawn using default style in MatLab, meaning the middle line inside the box represents the median value, the vertical limits are the 25th and 75th percentiles (q_1 and q_3), each whisker covers 1.5 the interquartile range (i.e. $q_3 - q_1$), and the crosses outside the whiskers represent outliers. Figures 2, 3, 4, and 5 from Appendix show the MS data points in various 2-D feature spaces.

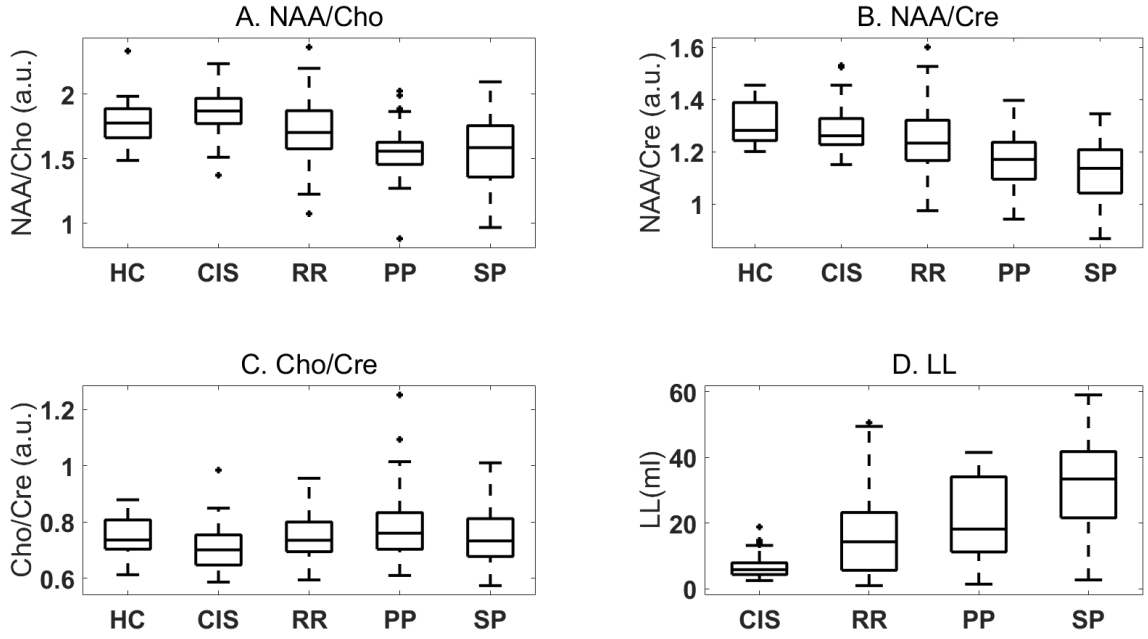

Figure 1: Boxplots of MR metabolic features and lesion loads extracted from HC and MS patients: A. NAA/Cho; B. NAA/Cre; C. Cho/Cre; D. Lesion load (LL).

Using the previously described (Section 2.6) linear mixed-effects models we found that the fixed effect MS course is statistically significant in the evolution of NAA/Cho, NAA/Cre, Cho/Cre, and LL, with corresponding p-values of: $3.4 \times 10^{-6}, 2 \times 10^{-4}$, 2.3×10^{-2}, and 2.6×10^{-4}. Table 2 provides adjusted p-values for multiple comparisons between the MS groups.

Table 3 shows F1-scores after training LDA using only metabolic ratios, as clinical data and lesion loads were not available for healthy controls. Corresponding BAR values of this table can be found in Table 5 in Appendix. If F1-scores are missing, then the classifier assigned all data points to the negative class (second MS group).

Surprisingly, the F1-scores for separating HC from any MS course are very low, and the same holds true for separating very early MS form (CIS) and the most likely MS evolution, RR and RR+SP. In contrast, for RR vs. PP we find that all three metabolic ratios have F1-scores higher than 75, with a maximum of 78 for NAA/Cre. For RR vs. SP the F1-scores are lower, with a maximum of 69 after combining all metabolic

	CIS - RR	RR - PP	RR - SP
NAA/Cho	-	**	**
NAA/Cre	-	-	*
Cho/Cre	-	-	-
LL	-	-	*

Table 2: Adjusted p-values for multiple comparisons between MS groups modelled by linear mixed effects model, tested using the "multcomp" package in 'R' (* for $p < 0.05$ and ** for $p < 0.01$).

	NAA/Cho	NAA/Cre	Cho/Cre	All 3 metabolic ratios
HC vs. CIS	35	33	43	36
HC vs. RR	6	16	-	14
HC vs. PP	47	45	19	49
HC vs. RR+SP	8	19	-	16
HC vs. PP+SP	21	26	-	28
CIS vs. RR	15	-	-	21
CIS vs. RR+SP	3	-	-	19
RR vs. PP	75	78	75	74
RR vs. SP	60	67	58	69

Table 3: F1-scores for all nine classification tasks (rows) after training LDA using only metabolic ratios. Values above 75 are coloured in light gray.

features.

Table 4 shows F1-scores of classification tasks involving only MS patients. Training was done on seven different combinations of features to evaluate the classification power of clinical data, lesion loads, and metabolic features. Corresponding BAR values can be found in Table 6 in Appendix. If F1-scores are missing, then the classifier assigned all data points to the negative class (second MS group).

The highest F1-scores for CIS vs. RR and CIS vs. RR+SP, respectively 71 and 72, were achieved by SVM-rbf trained on clinical data and lesion loads. Training any classifier only on metabolic features yielded very low F1-scores.

The highest F1-score for RR vs. PP (85) was achieved by LDA using patient age, disease age, and EDSS. Adding all spectroscopic information maintained the F1-score at 85, while adding lesion load lowered the F1-score at 79. LDA outperformed SVM-rbf and RF in all RR vs. PP cases, always achieving an F1-score higher than 70.

The highest value for RR vs. SP (87) was first achieved after training SVM-rbf on clinical and metabolic features, but also with LDA trained on all features combined (clinical data, lesion loads, and metabolic features). SVM-rbf outperfomed LDA in the majority RR vs. SP cases, but only with 1 to 2%.

	CIS vs. RR			CIS vs. RR+SP			RR vs. PP			RR vs. SP		
	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF
M	21	48	11	19	31	-	74	52	73	69	70	67
LL	-	51	27	-	40	24	71	19	73	75	77	68
Age + DD	48	58	51	44	56	50	79	64	74	76	75	71
Age + DD + EDSS	55	65	49	57	66	48	85	81	79	84	85	84
Age + DD + EDSS + LL	67	71	59	63	72	60	79	75	79	86	86	86
Age + DD + EDSS + M	56	59	48	60	59	51	85	83	80	86	87	85
Age + DD + EDSS + LL + M	65	64	57	65	63	57	83	81	78	87	87	86

Table 4: F1-scores for classification tasks involving only MS patients (columns). Abbreviations: M = all three average metabolic ratios; Age = patient age; DD = disease duration; LL = lesion load; EDSS = Expanded Disability Status Scale. Values between 75 and 79 are coloured in light gray, values between 80 and 84 are coloured in medium gray, while values larger than 85 are coloured in dark gray.

4. Discussion

In this paper, we present results for nine binary classification problems using clinical data, lesion loads and metabolic features extracted from MS patients and healthy controls. We focused on metabolic features as numerous studies showed significant metabolic alterations in MS patients of different MS forms. It has been demonstrated that metabolic abnormalities in MS patients are not restricted to lesions alone [27–33] and NAWM tissue is well known to be altered in MS [34, 35]. Concentrations of NAA in NAWM were shown to be significantly lower in MS patients [36–42]. Concentrations of Cho and Cre in NAWM were shown to be significantly higher in MS patients [10, 27, 39, 40, 43]. Concentrations of NAA/Cre in NAWM were shown to be significantly lower in MS patients [44, 45]. Multiple studies also report significant differences between metabolite concentrations in lesions vs. NAWM of HC: lower NAA and increased Cho and Cre [27, 29, 46–50].

Our findings are in agreement with these previous reports as decreased NAA and increased Cho and Cre contents were measured in NAWM and lesions of MS patients. After building linear mixed-effects models to properly analyze the statistical difference between the four clinical courses, we observed significant differences at the disease starting points of all MS courses using four response variables, namely the lesion load, NAA/Cre, NAA/Cho, and Cho/Cre ratios. A cross-sectional study [51] based on a large subset of our MRSI data found statistical differences in the NAA/Cre and NAA/Cho ratios between HC and RR, PP, SP, and RR+PP+SP patients. Another cross-sectional study, based on 77 MS patients [52], analyzing classification tasks based on DTI data found very high F1-scores (91.8 for both HC-CIS and CIS-RR) after training SVM-rbf on six global brain connectivity metrics. For RR vs. PP their maximum F1-score was 75.6, which is lower than our results based on metabolic features, while for RR vs. SP, their maximum F1-score was 85.5, which is comparable to our results. It is also worth mentioning that they focused only on DTI metrics without adding any clinical data, which might improve their results.

In this study, we analyzed the added value of combining standard clinical data with quantitative magnetic resonance features. To this purpose, we trained linear and non-linear classifiers only on advanced MR features, and then only on clinical data. Afterwards we train the classifiers on clinical data combined with lesion loads and metabolic features.

Although MS patients are expected to have significantly different WM metabolism compared to healthy controls, this difference was not reflected in the metabolic average obtained from "good quality" voxels (Figure 2, A and B). This result is not entirely surprising, considering that we averaged over a high number of voxels, and the subtle lesion information could be lost in the average. However, we can visually see in Figure 2:C&D that the two progressive MS courses tend to have lower NAA/Cho and NAA/Cre ratios than healthy controls.

CIS and RR patients' distribution over the NAA/Cho and NAA/Cre feature space do not differ too much, as seen in Figure 3:A. Disease duration interval for RR patients is much larger than for CIS patients, as most of CIS patients have a disease duration lower than 5 years, which can be seen in Figure 4:A. Because RR patients have more relapses than CIS patients, the number of lesions will be higher and the lesion volume as well, while EDSS scores are mainly in the same range, as seen in Figure 5:A. BAR values in Table 6 show a maximum of 85, when combining patient age, disease duration, EDSS, and lesion load. However, the corresponding maximum F1-score of 71 is much lower because the dataset is unbalanced (61 CIS vs. 214 RR), heavily influencing the classifier's precision. In this case the F1-score reflects better than BAR the difficulty of discriminating CIS from RR forms.

CIS and SP patients' distribution over different features is visible in Figure 3:B, Figure 4:B, and Figure 5:B, and it is clear that these two are the least and most advanced forms of MS. Because RR patients will eventually evolve into SP forms during their lifetime, we grouped together RR and SP patients for a separate classification task versus CIS patients. BAR values in Table 6 show a maximum of 92, when combining patient age, disease duration, EDSS and lesion load. The same discussion as for CIS vs. RR apply: the corresponding maximum F1-score is only 72 because the dataset is very unbalanced (61 CIS vs. 410 RR+SP) and the precision will be very low.

RR and PP patients can be discriminated using only EDSS by visually inspecting Figure 5:C. Training a linear classifier on clinical data (patient age, disease duration, and EDSS) gives the maximum F1-score of 85. Adding the 3 metabolic features keeps the score at 85, while adding lesion load information lowers the score to 79. This drop in the F1-score suggests that lesion load is not a good feature to use in differentiating RR from PP patients. Indeed, these two MS forms have the closest lesion load averages (16.7 ml and 20.8 ml), as shown in Table 1. In contrast, the clinical status of RR and PP patients are very different, as reflected by the EDSS values of 2 for RR and 4 for PP. Training LDA on individual metabolic features always provided higher F1-scores than lesion load, therefore we can conclude that for RR vs. PP, metabolic features have a higher discrimination power than LL. BAR values in Table 6 are also closer to the

F1-scores in Table 4 because the dataset is more balanced compared to previous cases.

RR and SP patients can also be discriminated using only EDSS by visually inspecting Figure 5:D. Our results showed that EDSS is very important in differentiating RR patients from primary or secondary progressive patients. We also report consistent higher F1-scores for classifiers trained only on lesion load compared to classifiers trained only on metabolic features. Furthermore, it is clearly visible in Table 3 that we obtain higher F1-scores for this classification task using multiple features, compared to the rest of 8 tasks. These findings suggest that in the future it might be possible to build a decision support system using clinical data combined with lesion loads and metabolic features.

When comparing classification tasks from a computational point of view, LDA is clearly the winner as the training period last only 3 hours using a computer with 8 threads. Training both SVM-rbf and RF took around 20 days in total and it was done using 60 threads, meaning LDA is approximately 600 times faster than SVM-rbf or RF. Also, the maximum F1-scores for RR vs. PP and RR vs. SP were obtained with LDA and SVM-rbf, suggesting that a linear classifier performs equally good as a non-linear classifier in these cases.

Averaging metabolite ratios over the entire MRSI grid, even if only the "good quality" voxels are kept, provides only moderate markers for discriminating RR vs. PP. Combining patient age, disease duration, EDSS, and averaged metabolic ratios, leads to the highest classification results. Extracting metabolic information from specific brain sub-regions of the MRSI grid (e.g NAWM or lesions) should provide more accurate information and help the classification tasks. Further investigations about the MS patients' evolution will be done in the future and will focus on sub-regions metabolite quantification, patient treatment, multi-class classification, and DTI-based brain connectivity metrics.

5. Conclusions

In this paper, we performed nine binary classification tasks and report F1-scores and BAR values after learning linear and non-linear classifiers on combinations of clinical data, lesion loads, and metabolic features. We presented a simple method to compute metabolic features by averaging metabolite ratios over "good quality" voxels of a MRSI grid. Using linear mixed-effects models we found that the MS course is statistically significant in the evolution of four response variables: Lesion Load, NAA/Cre, NAA/Cho, and Cho/Cre ratios. Our results showed that the best classifier for discriminating CIS from RR or RR+SP is SVM-rbf trained on clinical data and lesion loads. We also showed that discriminating RR from PP or SP with high accuracy is possible when training LDA on clinical data. For RR vs. PP, adding metabolic features will not change the results, while for RR vs. SP, adding metabolic features and lesion loads will slightly improve the results.

6. Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was funded by European project EU MC ITN TRANSACT 2012 (no. 316679) and the ERC Advanced Grant BIOTENSORS nr.339804. EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

7. References

[1] A. Compston and A. Coles, "Multiple sclerosis," *The Lancet*, vol. 372, pp. 1502–1518, Oct. 2008.
[2] D. McAlpine and A. Compston, *McAlpine's multiple sclerosis*. Elsevier Health Sciences, 2005.
[3] D. H. Miller, D. T. Chard, and O. Ciccarelli, "Clinically isolated syndromes," *The Lancet Neurology*, vol. 11, no. 2, pp. 157–169, 2012.
[4] A. Scalfari, A. Neuhaus, A. Degenhardt, G. P. Rice, P. A. Muraro, M. Daumer, and G. C. Ebers, "The natural history of multiple sclerosis, a geographically based study 10: relapses and long-term disability," *Brain*, vol. 133, no. 7, pp. 1914–1929, 2010.
[5] W. I. McDonald, A. Compston, G. Edan, D. Goodkin, H.-P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, *et al.*, "Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis," *Annals of neurology*, vol. 50, no. 1, pp. 121–127, 2001.
[6] C. H. Polman, S. C. Reingold, G. Edan, M. Filippi, H.-P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O'Connor, *et al.*, "Diagnostic criteria for multiple sclerosis: 2005 revisions to the mcdonald criteria," *Annals of neurology*, vol. 58, no. 6, pp. 840–846, 2005.
[7] C. H. Polman, S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, *et al.*, "Diagnostic criteria for multiple sclerosis: 2010 revisions to the mcdonald criteria," *Annals of neurology*, vol. 69, no. 2, pp. 292–302, 2011.
[8] M. Filippi, M. Horsfield, S. Morrissey, D. MacManus, P. Rudge, W. McDonald, and D. Miller, "Quantitative brain mri lesion load predicts the course of clinically isolated syndromes suggestive of multiple sclerosis," *Neurology*, vol. 44, no. 4, pp. 635–635, 1994.
[9] À. Rovira, C. Auger, and J. Alonso, "Magnetic resonance monitoring of lesion evolution in multiple sclerosis," *Therapeutic advances in neurological disorders*, vol. 6, no. 5, pp. 298–310, 2013.
[10] M. Tartaglia, S. Narayanan, N. De Stefano, R. Arnaoutelis, S. Antel, S. Francis, A. Santos, Y. Lapierre, and D. Arnold, "Choline is increased in pre-lesional normal appearing white matter in multiple sclerosis," *Journal of neurology*, vol. 249, no. 10, pp. 1382–1390, 2002.
[11] B. R. Sajja, J. S. Wolinsky, and P. A. Narayana, "Proton magnetic resonance spectroscopy in multiple sclerosis," *Neuroimaging clinics of North America*, vol. 19, no. 1, pp. 45–58, 2009.
[12] F. D. Lublin, S. C. Reingold, *et al.*, "Defining the clinical course of multiple sclerosis results of an international survey," *Neurology*, vol. 46, no. 4, pp. 907–911, 1996.
[13] J.-B. Poullet, "Quantification and classification of magnetic resonance spectroscopic data for brain tumor diagnosis," *Katholic University of Leuven*, 2008.
[14] "Spid." `http://homes.esat.kuleuven.be/%7Ebiomed/software.php#SpidGUI`. Accessed: 2015-02-10.

[15] J.-B. Poullet, D. M. Sima, A. W. Simonetti, B. De Neuter, L. Vanhamme, P. Lemmerling, and S. Van Huffel, "An automated quantitation of short echo time mrs spectra in an open source software environment: Aqses," *NMR in Biomedicine*, vol. 20, no. 5, pp. 493–504, 2007.

[16] C. Sava, R. Anca, D. M. Sima, J.-B. Poullet, A. J. Wright, A. Heerschap, and S. Van Huffel, "Exploiting spatial information to estimate metabolite levels in two-dimensional mrsi of heterogeneous brain lesions," *NMR in Biomedicine*, vol. 24, no. 7, pp. 824–835, 2011.

[17] T. Sundin, L. Vanhamme, P. Van Hecke, I. Dologlou, and S. Van Huffel, "Accurate quantification of 1 h spectra: From finite impulse response filter design for solvent suppression to parameter estimation," *Journal of Magnetic Resonance*, vol. 139, no. 2, pp. 189–204, 1999.

[18] S. Jain, D. M. Sima, A. Ribbens, M. Cambron, A. Maertens, W. Van Hecke, J. De Mey, F. Barkhof, M. D. Steenwijk, M. Daams, *et al.*, "Automatic segmentation and volumetry of multiple sclerosis brain lesions from mr images," *NeuroImage: Clinical*, vol. 8, pp. 367–375, 2015.

[19] D. M. Bates, "lme4: Mixed-effects modeling with r," *URL http://lme4. r-forge. r-project. org/book*, 2010.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "Package lmertest," *R package version*, pp. 2–0, 2015.

[21] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical journal*, vol. 50, no. 3, pp. 346–363, 2008.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[25] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] P. A. Narayana, T. J. Doyle, D. Lai, and J. S. Wolinsky, "Serial proton magnetic resonance spectroscopic imaging, contrast-enhanced magnetic resonance imaging, and quantitative lesion volumetry in multiple sclerosis," *Annals of neurology*, vol. 43, no. 1, pp. 56–71, 1998.

[28] T. J. Doyle, R. Pathak, J. S. Wolinsky, and P. A. Narayana, "Automated proton spectroscopic image processing," *Journal of Magnetic Resonance, Series B*, vol. 106, no. 1, pp. 58–63, 1995.

[29] J. He, M. Inglese, B. S. Li, J. S. Babb, R. I. Grossman, and O. Gonen, "Relapsing-remitting multiple sclerosis: Metabolic abnormality in nonenhancing lesions and normal-appearing white matter at mr imaging: Initial experience 1," *Radiology*, vol. 234, no. 1, pp. 211–217, 2005.

[30] L. Fu, P. Matthews, N. De Stefano, K. Worsley, S. Narayanan, G. Francis, J. Antel, C. Wolfson, and D. Arnold, "Imaging axonal damage of normal-appearing white matter in multiple sclerosis.," *Brain*, vol. 121, no. 1, pp. 103–113, 1998.

[31] C. Husted, D. Goodin, J. Hugg, A. A. Maudsley, J. Tsuruda, S. De Bie, G. Fein, G. Matson, and M. Weiner, "Biochemical alterations in multiple sclerosis lesions and normal-appearing white matter detected by in vivo 31p and 1h spectroscopic imaging," *Annals of neurology*, vol. 36, no. 2, pp. 157–165, 1994.

[32] S. Narayanan, L. Fu, E. Pioro, N. De Stefano, D. Collins, G. Francis, J. Antel, P. Matthews, and D. Arnold, "Imaging of axonal damage in multiple sclerosis: spatial distribution of magnetic resonance imaging lesions," *Annals of neurology*, vol. 41, no. 3, pp. 385–391, 1997.

[33] P. Sarchielli, O. Presciutti, G. Pelliccioli, R. Tarducci, G. Gobbi, P. Chiarini, A. Alberti, F. Vicinanza, and V. Gallai, "Absolute quantification of brain metabolites by proton magnetic resonance spectroscopy in normal-appearing white matter of multiple sclerosis patients," *Brain*, vol. 122, no. 3, pp. 513–521, 1999.

[34] P. A. Narayana, "Magnetic resonance spectroscopy in the monitoring of multiple sclerosis," *Journal of Neuroimaging*, vol. 15, no. s4, pp. 46S–57S, 2005.

[35] N. De Stefano and M. Filippi, "Mr spectroscopy in multiple sclerosis," *Journal of Neuroimaging*, vol. 17, no. s1, pp. 31S–35S, 2007.

[36] A. Bitsch, H. Bruhn, V. Vougioukas, A. Stringaris, H. Lassmann, J. Frahm, and W. Brück, "Inflammatory cns demyelination: histopathologic correlation with in vivo quantitative proton mr spectroscopy," *American Journal of Neuroradiology*, vol. 20, no. 9, pp. 1619–1627, 1999.

[37] C. Bjartmar, R. P. Kinkel, G. Kidd, R. A. Rudick, and B. D. Trapp, "Axonal loss in normal-appearing white matter in a patient with acute ms," *Neurology*, vol. 57, no. 7, pp. 1248–1252, 2001.

[38] M. Tiberio, D. Chard, D. Altmann, G. Davies, C. Griffin, M. McLean, W. Rashid, J. Sastre-Garriga, A. Thompson, and D. Miller, "Metabolite changes in early relapsing–remitting multiple sclerosis," *Journal of neurology*, vol. 253, no. 2, pp. 224–230, 2006.

[39] M. Inglese, B. S. Li, H. Rusinek, J. S. Babb, R. I. Grossman, and O. Gonen, "Diffusely elevated cerebral choline and creatine in relapsing-remitting multiple sclerosis," *Magnetic resonance in medicine*, vol. 50, no. 1, pp. 190–195, 2003.

[40] J. Suhy, W. Rooney, D. Goodkin, A. Capizzano, B. Soher, A. A. Maudsley, E. Waubant, P. Andersson, and M. Weiner, "1h mrsi comparison of white matter and lesions in primary progressive and relapsing-remitting ms," *Multiple sclerosis*, vol. 6, no. 3, pp. 148–155, 2000.

[41] M. Wattjes, M. Harzheim, G. Lutterbey, L. Klotz, H. Schild, and F. Träber, "Axonal damage but no increased glial cell activity in the normal-appearing white matter of patients with clinically isolated syndromes suggestive of multiple sclerosis using high-field magnetic resonance spectroscopy," *American Journal of Neuroradiology*, vol. 28, no. 8, pp. 1517–1522, 2007.

[42] M. P. Wattjes, M. Harzheim, G. G. Lutterbey, M. Bogdanow, H. H. Schild, and F. Träber, "High field mr imaging and 1h-mr spectroscopy in clinically isolated syndromes suggestive of multiple sclerosis," *Journal of neurology*, vol. 255, no. 1, pp. 56–63, 2008.

[43] A. Tourbah, J.-L. Stievenart, A. Abanou, M.-T. Iba-Zizen, H. Hamard, O. Lyon-Caen, E. Cabanis, and L. Stievenart, "Normal-appearing white matter in optic neuritis and multiple sclerosis: a comparative proton spectroscopy study," *Neuroradiology*, vol. 41, no. 10, pp. 738–743, 1999.

[44] S. M. Leary, C. A. Davie, G. J. Parker, V. L. Stevenson, L. Wang, G. J. Barker, D. H. Miller, and A. Thompson, "1h magnetic resonance spectroscopy of normal appearing white matter in primary progressive multiple sclerosis," *Journal of neurology*, vol. 246, no. 11, pp. 1023–1026, 1999.

[45] P. A. Narayana, J. S. Wolinsky, S. B. Rao, R. He, M. Mehta, *et al.*, "Multicentre proton magnetic resonance spectroscopy imaging of primary progressive multiple sclerosis," *Multiple Sclerosis*, vol. 10, no. 3 suppl, pp. S73–S78, 2004.

[46] C. Davie, G. Barker, A. Thompson, P. Tofts, W. McDonald, and D. Miller, "1h magnetic resonance spectroscopy of chronic cerebral white matter lesions and normal appearing white matter in multiple sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 63, no. 6, pp. 736–742, 1997.

[47] D. Arnold, N. De Stefano, S. Narayanan, and P. Matthews, "Proton mr spectroscopy in multiple sclerosis.," *Neuroimaging clinics of North America*, vol. 10, no. 4, pp. 789–98, 2000.

[48] J. S. Wolinsky, P. A. Narayana, and M. J. Fenstermacher, "Proton magnetic resonance spectroscopy in multiple sclerosis," *Neurology*, vol. 40, no. 11, pp. 1764–1764, 1990.

[49] H. Larsson, P. Christiansen, M. Jensen, J. Frederiksen, A. Heltberg, J. Olesen, and O. Henriksen, "Localized in vivo proton spectroscopy in the brain of patients with multiple sclerosis," *Magnetic resonance in medicine*, vol. 22, no. 1, pp. 23–31, 1991.

[50] C. Davie, C. Hawkins, G. Barker, A. Brennan, P. Tofts, D. Miller, and W. McDonald, "Serial proton magnetic resonance spectroscopy in acute multiple sclerosis lesions," *Brain*, vol. 117, no. 1, pp. 49–58, 1994.

[51] S. Hannoun, M. Bagory, F. Durand-Dubief, D. Ibarrola, J.-C. Comte, C. Confavreux, F. Cotton,

and D. Sappey-Marinier, "Correlation of diffusion and metabolic alterations in different clinical forms of multiple sclerosis," *PLoS One*, vol. 7, no. 3, p. e32525, 2012.

[52] G. Kocevar, C. Stamile, S. Hannoun, F. Cotton, S. Vukusic, F. Durand-Dubief, and D. Sappey-Marinier, "Graph theory-based brain connectivity for automatic classification of multiple sclerosis clinical courses," *Frontiers in Neuroscience*, vol. 10, p. 478, 2016.

8. Appendix

	NAA/Cho	NAA/Cre	Cho/Cre	All 3 metabolites
HC vs. CIS	47	46	61	53
HC vs. RR	50	55	50	52
HC vs. PP	76	78	45	77
HC vs. RR+SP	52	60	50	59
HC vs. PP+SP	61	66	50	71
CIS vs. RR	52	50	50	52
CIS vs. RR+SP	51	49	50	54
RR vs. PP	59	63	48	63
RR vs. SP	57	65	39	66

Table 5: BAR values for all 9 classification tasks (rows) after training LDA using only metabolic ratios.

	CIS vs. RR			CIS vs. RR+SP			RR vs. PP			RR vs. SP		
	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF
M	52	68	49	54	63	49	63	28	59	66	66	63
LL	48	70	52	50	73	56	43	12	58	74	75	68
Age + DD	66	75	68	66	83	70	67	38	62	75	76	71
Age + DD + EDSS	71	80	67	77	89	69	81	78	70	84	85	84
Age + DD + EDSS + LL	79	85	73	81	92	76	71	72	69	86	86	85
Age + DD + EDSS + M	72	76	66	81	82	70	80	81	71	86	87	84
Age + DD + EDSS + LL + M	78	80	71	82	83	73	78	78	68	86	86	86

Table 6: BAR values for classification tasks involving only MS patients (columns). Abbreviations: M = all three average metabolic ratios; Age = patient age; DD = disease duration; LL = lesion load; EDSS = Expanded Disability Status Scale. Values between 75 and 79 are coloured in light gray, values between 80 and 84 are coloured in medium gray, while values larger than 85 are coloured in dark and very dark gray.

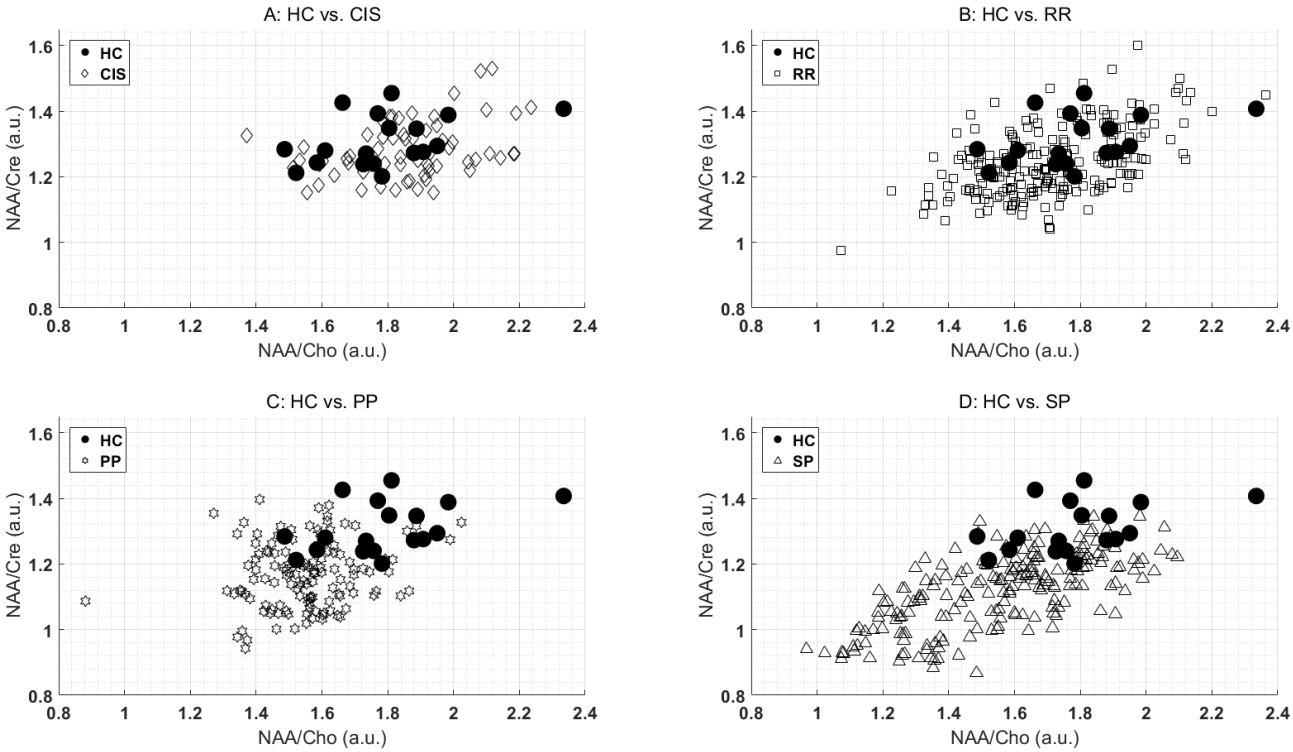

Figure 2: HC vs. MS groups in 2-D feature space: x-axis is NAA/Cho and y-axis is NAA/Cre.

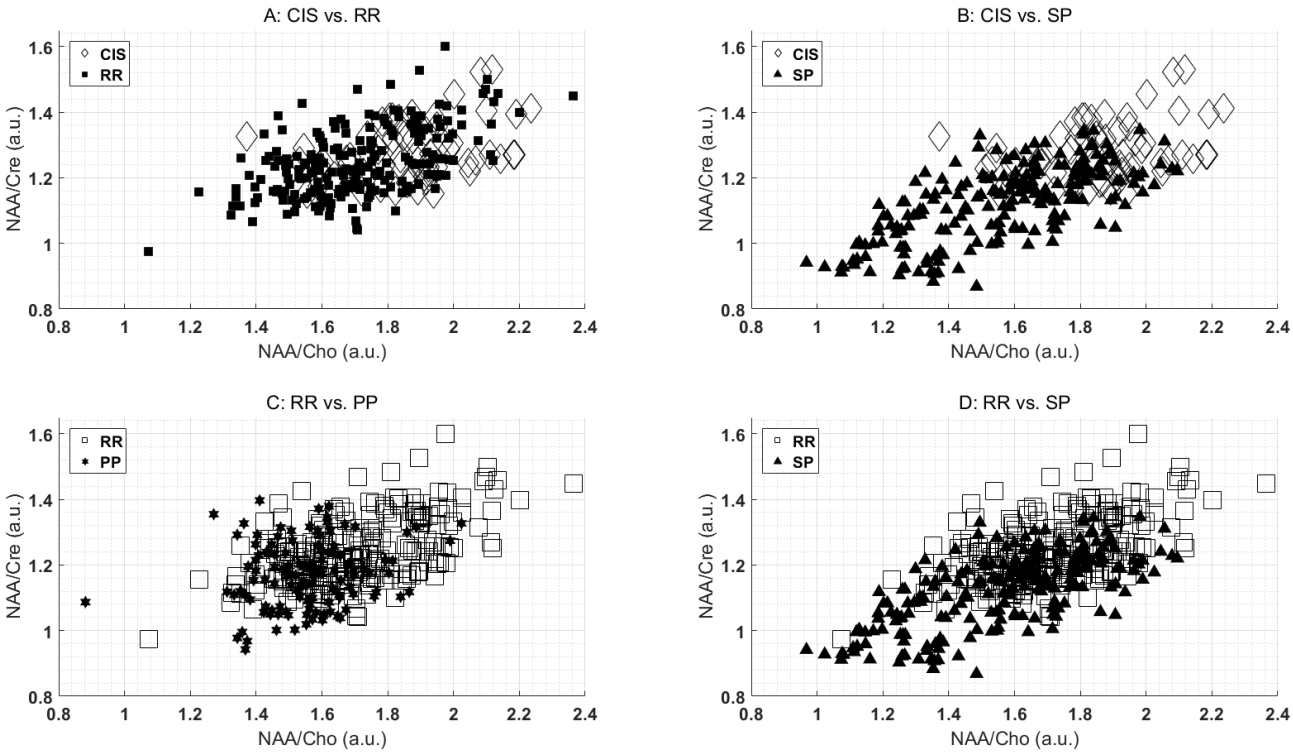

Figure 3: Comparison of MS groups in 2-D feature space: x-axis is NAA/Cho and y-axis is NAA/Cre.

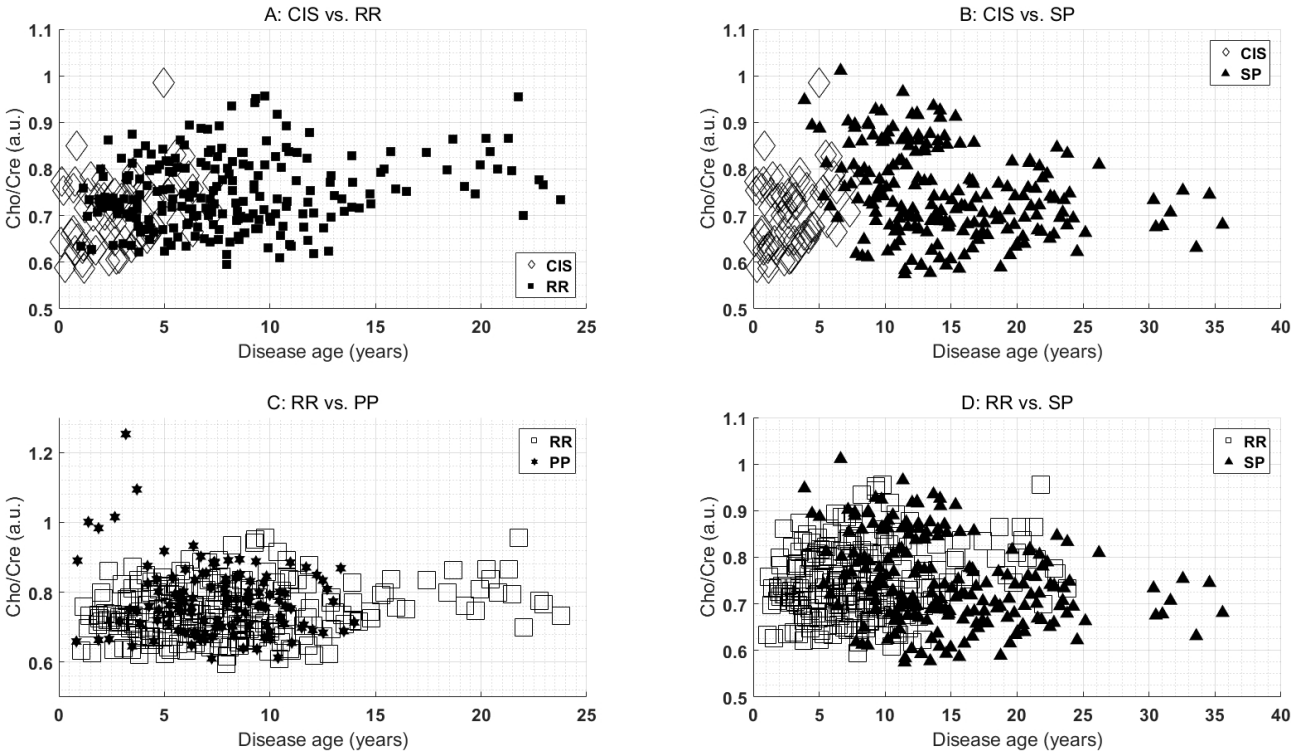

Figure 4: Comparison of MS groups in 2-D feature space: x-axis is disease age and y-axis is Cho/Cre.

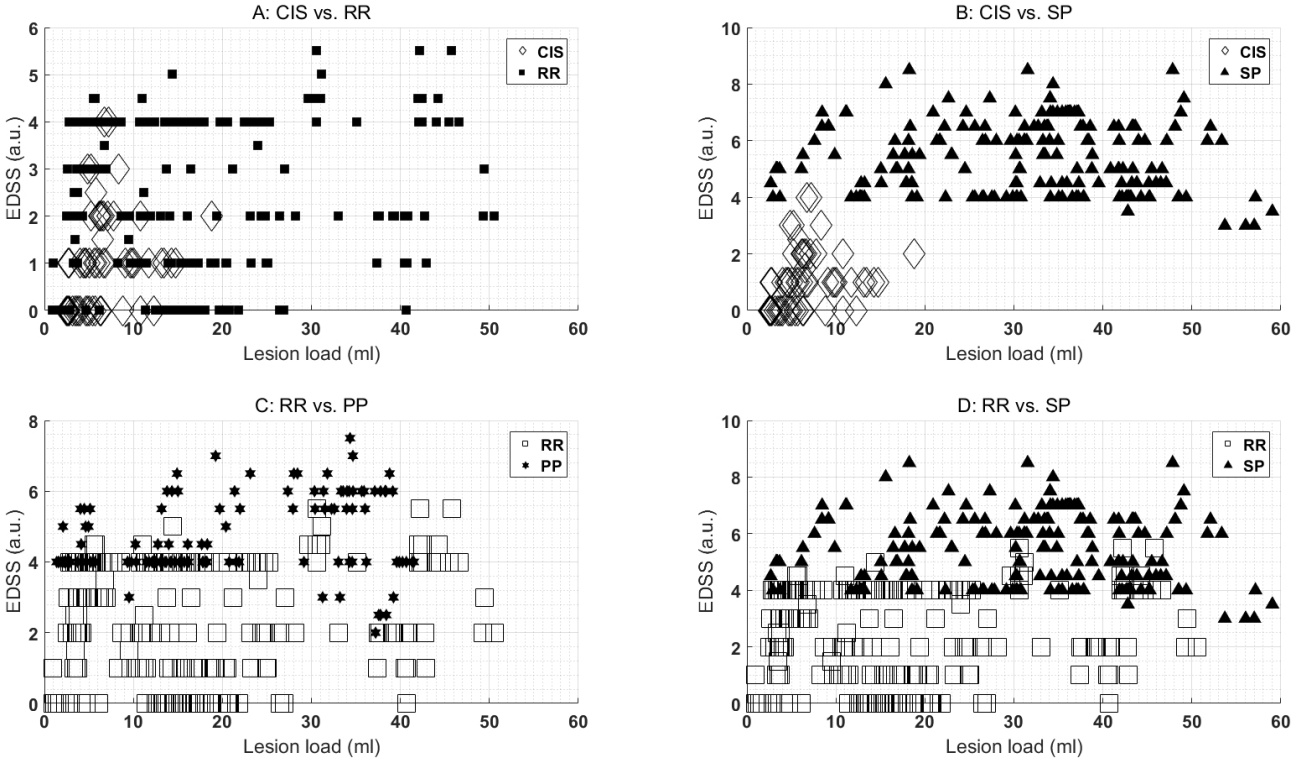

Figure 5: Comparison of MS groups in 2-D feature space: x-axis is lesion load and y-axis is EDSS.