RESEARCH ARTICLE

Early pseudoprogression and progression lesions in glioblastoma patients are both metabolically heterogeneous

Gülnur Ungan[1,2] | Albert Pons-Escoda[3] | Daniel Ulinic[2] | Carles Arús[1,2] | Sandra Ortega-Martorell[4] | Ivan Olier[4] | Alfredo Vellido[1,5] | Carles Majós[1,3] | Margarida Julià-Sapé[1,2]

[1]Centro de Investigación Biomédica en Red (CIBER), Madrid, Spain

[2]Departament de Bioquímica i Biologia Molecular and Institut de Biotecnologia i Biomedicina (IBB), Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

[3]Grup de Neuro-oncologia, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Hospital Universitari de Bellvitge, Barcelona, Spain

[4]Data Science Research Centre, Liverpool John Moores University (LJMU), Liverpool, UK

[5]IDEAI-UPC Research Center, UPC BarcelonaTech, Barcelona, Spain

Correspondence
Margarida Julià-Sapé, Facultat de Biociències, Edifici Cs, Carrer de la Vall Moronta, 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain.
Email: margarita.julia@uab.cat

The standard treatment in glioblastoma includes maximal safe resection followed by concomitant radiotherapy plus chemotherapy and adjuvant temozolomide. The first follow-up study to evaluate treatment response is performed 1 month after concomitant treatment, when contrast-enhancing regions may appear that can correspond to true progression or pseudoprogression. We retrospectively evaluated 31 consecutive patients at the first follow-up after concomitant treatment to check whether the metabolic pattern assessed with multivoxel MRS was predictive of treatment response 2 months later. We extracted the underlying metabolic patterns of the contrast-enhancing regions with a blind-source separation method and mapped them over the reference images. Pattern heterogeneity was calculated using entropy, and association between patterns and outcomes was measured with Cramér's V. We identified three distinct metabolic patterns—proliferative, necrotic, and responsive, which were associated with status 2 months later. Individually, 70% of the patients showed metabolically heterogeneous patterns in the contrast-enhancing regions. Metabolic heterogeneity was not related to the regions' size and only stable patients were less heterogeneous than the rest. Contrast-enhancing regions are also metabolically heterogeneous 1 month after concomitant treatment. This could explain the reported difficulty in finding robust pseudoprogression biomarkers.

KEYWORDS
convex non-negative matrix factorization, glioblastoma, MR, MRS, pseudoprogression

---

Abbreviations: AUC, area under the receiver operating characteristic curve; Cho/NAA, choline/N-acetyl aspartate; CI, confidence interval; cNMF, convex non-negative matrix factorization; GAM, glioma-associated microglia/macrophage; LDA, linear discriminant analysis; LR, logistic regression; MRSI, magnetic resonance spectroscopic imaging; MV, multivoxel; NAC, N-acetyl-aspartate-containing compound; PET, positron emission tomography; PP, pseudoprogression; PUFA, polyunsaturated fatty acid; RF, random forest; SCER, suspicious contrast-enhancing region; SVM, support vector machine; T1w, T1 weighted; TMZ, temozolomide; TP, true progression.

1 | INTRODUCTION

The standard treatment in glioblastoma patients includes maximal safe resection followed by concomitant treatment with radiotherapy plus chemotherapy with temozolomide (TMZ), and adjuvant TMZ.[1] The first MR follow-up study to evaluate treatment response is performed 1 month after concomitant treatment (P1M). The assessment of this MR study is challenging because an abnormal contrast-enhancing growth at this moment can be due to either tumor growth (true progression, TP), or treatment-associated changes (pseudoprogression, PP). Currently, these patients must wait for about 2 months until the next MR examination is performed to confirm the progression status of the abnormal contrast-enhancing growths—that is, whether they grow, shrink, or remain stable. In this respect, two recent studies using advanced imaging techniques that provide metabolic information have shown strong potential for distinguishing between TP and PP. In one of them,[2] microstructural information obtained from diffusion kurtosis MRI was combined with a positron emission tomography (PET) tracer derived from O-(2-[^{18}F]-fluoroethyl)-l-tyrosine (^{18}F-FET) PET, achieving an area under the receiver operating characteristic curve (AUC) of 0.97 on a set of 32 glioblastoma patients at time points between one and 55 months after the end of the radiotherapy. In another study[3] with 27 patients, eight of whom were with a P1M follow-up, the authors used 3D echo-planar spectroscopic imaging and found very promising differences in the choline/N-acetyl aspartate (Cho/NAA) ratio from the abnormal contrast-enhancing growths, with an AUC of 0.83. Other studies using MRS[4–6] are equally limited due to low numbers of cases. A recent multicenter retrospective study[7] (98 patients), using radiomic features extracted from the perfusion images for the classification task using a support vector machine (SVM), achieved an AUC of 0.89. Despite their potential, these techniques are far from the reach of clinical practitioners, and in general the cohorts under study are small even when pooling different time-points after the end of the concomitant treatment. In this respect, studies in cohorts with more extended periods after concomitant treatment would be focusing more effectively on tumor versus post-treatment changes, given that the PP dilemma appears exclusively within 3–6 months after concomitant treatment. For example, in Reference 8, a subset of patients from the *AVAglio* trial (NCT00943826) with early progressive disease were analyzed with radiomics, and a 70.6% accuracy in the prediction of one-year survival after the early progression phenomenon was found.

The purpose of our study was to evaluate whether the metabolic pattern at the P1M timepoint, assessed using MRS of early suspicious contrast-enhancing regions (SCERs), would be predictive of the progression status at the next exam (P3M).

2 | MATERIALS AND METHODS

2.1 | Study design and sample size calculation for machine learning (MLe)

Retrospective, observational MRS data were analyzed from 53 consecutive patients, who were treated at *Hospital Universitari de Bellvitge* between July 2016, when the first magnetic resonance spectroscopic imaging (MRSI) sequence was implemented, and January 2019.

2.2 | Ethics statement

Unspecific informed consent to participate in research projects was obtained from all patients and a waiver of a specific informed consent was provided by the ethics committee for this retrospective study (*Hospital Universitari de Bellvitge*, procedure number PR168/22). Patients' data were anonymized for this analysis. The confidential information of the patients was protected in accordance with national and international norms. The study was conducted in accordance with the Declaration of Helsinki.

2.3 | Inclusion criteria

1. Consecutive patients with histopathologically verified glioblastoma, having undergone resection and treatment with concomitant chemoradiotherapy with TMZ, followed by adjuvant TMZ according to the Stupp protocol.[1]
2. Detection of one or more early SCERs in the routine P1M exploration, an SCER being a contrast-enhancing mass adjacent to the postsurgical cavity.
3. Endpoint availability: follow-up with MR at least 2 months after the P1M and a clinical diagnosis of TP, PP, stable disease (S) or response (R) according to the RANO working group criteria.[9,10]
4. MRSI acquired at the P1M available.

2.4 | Acquisition protocol

The following full MR explorations were acquired at 1.5 T or 3 T with a Philips Ingenia scanner (Philips Healthcare, Best, The Netherlands): pre-operative, early post-operative, P1M, and every 2–3 months afterwards. For this study, the pre-operative, early post-operative, and P1M explorations and the next one (P3M) were evaluated. Each exploration included T1 weighted (T1w) with/without contrast, T2 weighted (T2w), and fluid-attenuated inversion recovery (FLAIR).

All P1M explorations were performed in a 3 T scanner and included MRSI with the following parameters: dimensions, field of view $8 \times 8 \times 1.5$ cm^3, volume of interest $8 \times 8 \times 1.5$ cm^3, with $16 \times 16 \times 1$ resolution; voxel size, $0.5 \times 0.5 \times 1.5$ cm^3; sequence, point resolved spectroscopy (PRESS); echo time (T_E) = 30 ms, repetition time (T_R) = 2000 ms. The MRSI grid was positioned to avoid bone or subcutaneous lipids and to include most of the SCER. One radiologist (AP-E) labelled the voxels over the SCER using the post-contrast T1w image at P1M, with the MRS grid overlaid onto the post-contrast T1w image, marking each individual voxel that was fully included in the SCER region. Progression status was determined according to RANO criteria[9,10] by the participating radiologist (either AP-E or CM).

2.5 | MRS data processing

One 16×16 MRSI grid per patient was available. Processing was performed by an experienced spectroscopist (MJ-S) using the jMRUI2XML[11] plugin of jMRUI,[12] with INTERPRET parameters,[13,14] to ensure that all spectra had been processed exactly in the same way (no manual first- or second-order phasing). All SCER voxels that had been marked by the radiologist (AP-E) were labelled with the same outcome according to the follow-up of each patient as TP, PP, R or S. From the labelled spectra, the same spectroscopist who processed the data performed a final quality control step, discarding bad-quality data. Criteria included the recognition of the following artifacts: bad water suppression with remaining water signal after Hankel–Lanczos singular value decomposition (HLSVD); bad phasing after eddy-current correction and low signal-to-noise ratio (no distinguishable peaks); baseline distortions. Provided that there were visible peaks in the 0–4.2 ppm region and there was no first- or second-order phasing problem and SNR was >20, the spectrum was considered acceptable, and as a final step spectra were UL2 renormalized in the 0–4.2 ppm region.

2.6 | MRS data analysis

2.6.1 | Data analysis design

Distinctive spectral patterns were extracted from the pooled SCER spectra that had passed all inclusion criteria and spectral quality control, using convex non-negative matrix factorization (cNMF).[15] The contribution of the extracted patterns to each spectrum in each patient's SCER was used to create color maps as overlays onto the reference MRI. The association of the sources with the progression status was calculated using Cramér's V, and the degree of heterogeneity of the SCER voxels was measured with Shannon's entropy. The weights obtained for each of the sources extracted for each spectrum were also used as input features for supervised MLe based classification, with the aim of predicting which patients would show progression in the P3M.

2.6.2 | cNMF

cNMF[15] is a variant of NMF, a method for blind source separation that, from a pattern recognition perspective, can also be seen as a feature extraction technique. cNMF can decompose all spectra from a dataset into sources or distinctive spectral patterns, as well as providing a value for the predominance of each source in each spectrum.

NMF methods factorize the data into two non-negative matrices: one being the sources (**F**) with dimensions $d \times k$, where k is the number of sources and d is the observed data dimension, and another (**M**) with the source-mixing coefficients, with dimensions $k \times n$, where n is the number of spectra, such that the product of **F** and **M** gives a good approximation to the original data matrix as $\mathbf{X} \approx \mathbf{FM}$. In this study, we used cNMF in the Matlab implementation from previous studies performed by our group using similar data.[16,17] For each voxel analyzed, the "winning source" was defined as in Reference 17, where the main source contributing to the spectral pattern was inferred from the highest value of the mixing matrix for that spectrum. Since NMF methods tend to converge to local minima, k-means initialization was used for the cNMF algorithm, as proposed in Reference 15. Convergence was considered to be achieved when the reconstruction error of the signals was below the threshold value of 10^{-5}. Given that in a previous study[18] we reported that the application of cNMF to a similar dataset often includes a variable number of

sources that are either artifactual or very unstable, in this study we performed a previous robustness evaluation of the sources extracted. To this end, we extracted $k = 2, 3, 4, 10,$ and 20 sources, and repeated each extraction 100 times.

2.6.3 | Visualization: winning source maps

Color maps of winning sources were obtained using source mixing coefficients and the winning source. Color maps were overlaid onto the reference DICOM image of the MRSI, with an in-house Matlab script based on the code already available in the Gannet software,[19] which can read Philips multivoxel (MV) data in DICOM. Briefly, from the DICOM MV and reference image, slice thickness, mid-slab position, and mid-slab orientation parameters were extracted and used in Gannet to find the exact grid location and to extract a patch of the reference image that matched the MV grid for each case.

2.6.4 | Statistics

Sample size was calculated according to References 20 and 21 for a Fisher's linear discriminant analysis (LDA), which has proven to perform well with this type of data,[13,22] and an asymptotic probability of misclassification P (infinite number of training samples) of $P = [0.001, 0.1]$. For this case the number of samples needed would be $N = [32-72]$ for a dimensionality of eight features, using $N = 4f$ for $P = 0.1$ and $N = 9f$ for $P = 0.001$, where N is the number of samples and f the number of features, features being those used for classification (see the following section).

The association between sources and progression status was calculated with Cramér's V (Reference 23) using jamovi.[24] Cramér's V ranges from 0 to 1: the closer to 1 the higher the association. For two categories, a medium association would be greater than 0.3 and large would be greater than 0.5.

Shannon's entropy (H) was calculated to evaluate the heterogeneity of the winning source outputs for each patient and each outcome (TP, PP, S, R), using the equation

$$\mathbf{H}(\mathbf{X}) = -\sum_{i=1}^{n} \mathbf{P}(x_i) \log_2 \mathbf{P}(x_i)$$

where x_i is the weight value in matrix $\mathbf{H}$ for each SCER voxel in the dataset.

2.6.5 | MLe-based classification

Several binary classifiers were developed to predict outcome: the goal was to distinguish TP (positive class) from response (negative class), defining response as a superclass including PP, R, and S. The following techniques from the scikit-learn Python library[25] were used: logistic regression (LR), random forest (RF), SVM and LDA (details in Supporting Information S3). Two different classification tasks were performed: voxel based and patient based. Class labelling was performed in this way: for the voxel-based classification, a voxel was assumed to be predicted as TP if S3 was the winning source for the three-source extraction. For the patient-based classification, all voxels in each patient were labelled according to the winning source for the patient. The input features for the classifiers were the three values of the mixing matrix $\mathbf{M}$ for each SCER voxel in the dataset. Overfitting was avoided with grid-search of parameters. Validation was performed with a fivefold cross-validation on both tasks, by randomly assigning voxels or patients to each of the folds. Classification performance measures were balanced accuracy, sensitivity, specificity, positive likelihood ratio and AUC.

3 | RESULTS

3.1 | Number of patients and voxels

A total of 53 patients were eligible for the study (32 [58.1%] men, 21 [38.9%] women); median age, 62 ± 10.3 years (61 ± 7.7 men and 62 ± 13.3 years women). Twenty-eight patients were MGMT(−), 24 were MGMT(+), and in one patient MGMT status was not available. Only one patient (number 39) was IDH mutant (also with methylated MGMT), in two other patients the IDH status was not available, and the remaining patients were IDH wildtype. The number of patients and reasons for exclusion, together with the accepted and discarded voxels per patient per

class, are shown in Figure 1 and in Table S1. From the 31 patients analyzed, (21 [67.7%] men, 10 [32.3%] women) the median age was 60 ± 9.1 years (60 ± 7.8 for men and 59 ± 11.4 years for women), and the class distribution was 11 TP, 15 PP, 3 S, and 2 R patients, amounting to a total of 172 individual voxels, with 72, 80, 12, and 8 voxels per class, respectively (Figure S1). The number of SCER voxels per patient ranged from 1 to 13.

3.2 | Source extraction from SCER spectra

The robustness of the set of solutions obtained decreased when increasing the number of sources, similarly to previous studies using the same approach.[18] For two and three sources, the same set was obtained in 100% of the repetitions. Beyond three sources, the solutions were exceedingly unstable. For four sources, there were three different solutions, the first one appearing in 80% of the 100 repetitions, the second one in 15% and the third one in 5%. For 10 sources 10 different solutions resulted with the same frequency (10%), and for 20 sources 20 different solutions with the same frequency (5%) (Supporting Information S1, Figures S2–S6). Figure 2 displays the spectral data used for the cNMF extractions, as well as the result of two- and three-source extractions. As can be observed, the spectral data from the different classes look very similar, with fatty acyl chains of mobile lipids (ML) at 0.9 and 1.28 ppm, lipid macromolecules (MM) at 2.0 ppm, total choline containing compounds (Cho) at 3.21 ppm, and total creatine (Cr) at 3.03 ppm, with variable ratios depending on the class, the S class being mostly composed of lipid signals and the TP class having on average a choline signal of similar intensity to the 1.28 ppm lipid signal, as would be expected from a proliferating tissue, in contrast with the PP class, in which the mean choline signal is about half of the intensity of the mean lipid signal at 1.28 ppm, but with high intraclass variability. Table 1 shows the winning source contribution for each voxel included in the study.

Source extraction resulted in the appearance of interpretable spectral patterns. For two sources (Figure 2, middle row), the first source (S1) corresponded to the typical necrotic tumor pattern, mainly mobile lipids at 0.9, 1.28, and 2.0 ppm, while the second (S2) was characteristic of proliferating glial tumors, still with mobile lipids at 0.9 and 1.28 ppm, lipid macromolecules between 2.0 and 2.5 ppm, and choline/creatine > 1, choline being the most prominent signal at 3.21 ppm. With three-source extraction (Figure 2, bottom row), the S1 of the previous extraction seemingly split into two slightly different patterns, with S2 very similar to the S1 of the two-source extraction, and S1, an intermediate pattern between proliferation and higher lipid signals. Interestingly, this S1 pattern also displayed a 2.8 ppm signal corresponding to polyunsaturated fatty acids (PUFAs), which we have recurrently found to be associated with response to different chemotherapies in previous studies with animal models.[26–28]
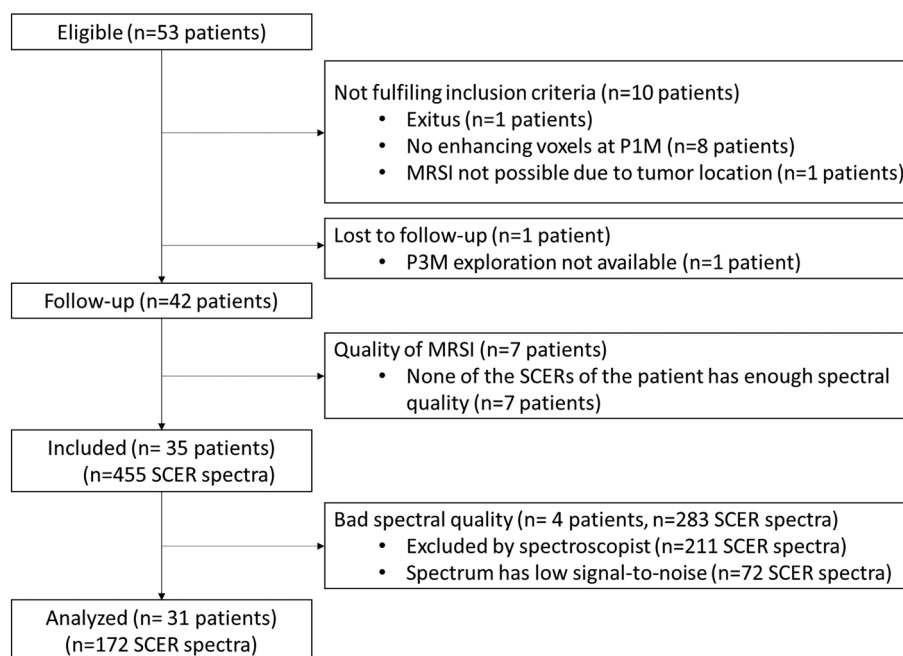

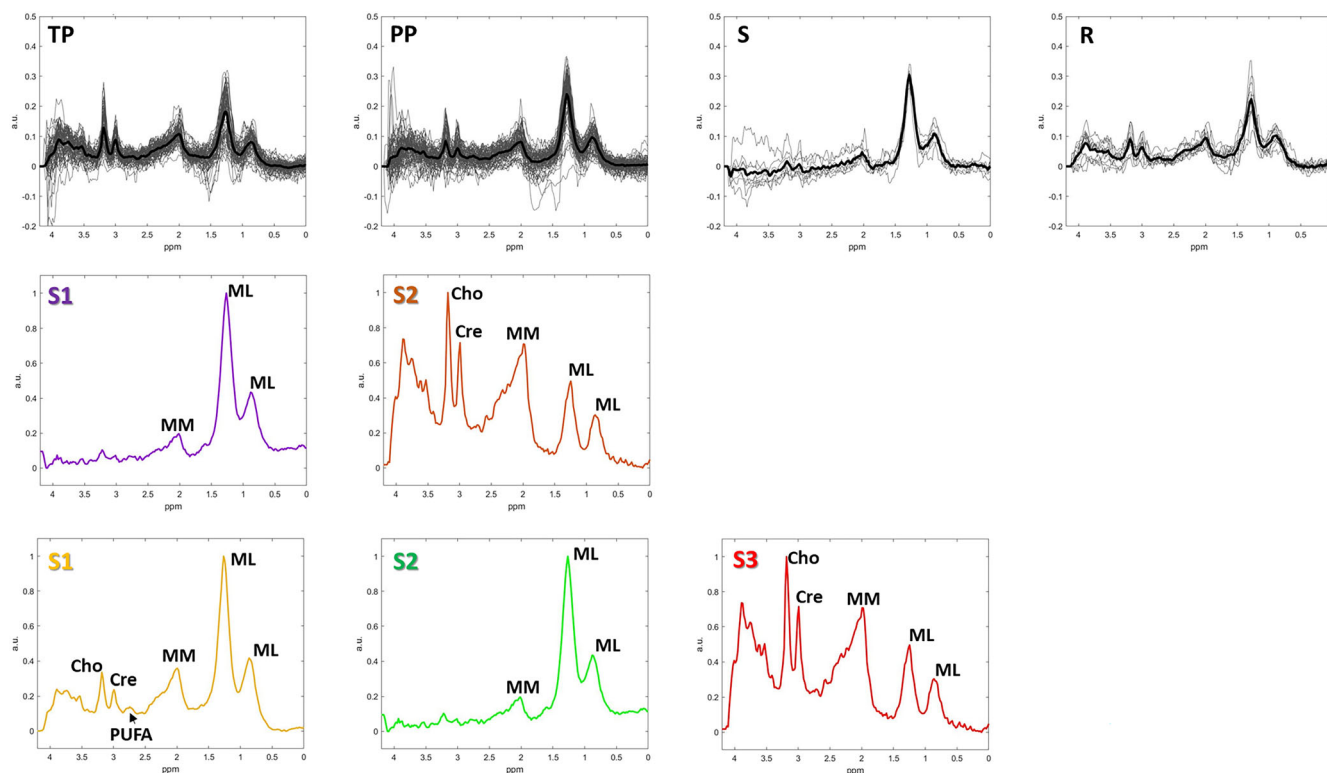**FIGURE 1** Flow diagram with patients and data included in the study.

FIGURE 2 Top row: mean (thick line) overlaid onto individual spectra with SNR > 20. Middle row: result of two-source extraction. S1, Source 1; S2, Source 2. Bottom row: result of three-source extraction. S1, Source 1; S2, Source 2; S3, Source 3.

TABLE 1 Contingency tables for two and three sources, depicting the winning source for each of the individual spectra, pooling all 31 patients.

	Type				Total
	TP	PP	R	S	
Two sources					
Source1	53	34	7	1	95
Source2	19	46	5	7	77
Total	72	80	12	8	172
Three sources					
Source1	29	48	9	1	87
Source2	1	12	0	6	19
Source3	42	20	3	1	66
Total	72	80	12	8	172

3.3 | Statistics

Regarding the association of sources with outcome, Cramér's V test between the winning source of each spectrum and the class resulted in moderate to high associations. The highest association was found for three sources and four classes ($V = 0.410$). For three sources and dichotomizing (TP versus the rest), association decreased ($V = 0.388$). Association also decreased when testing for two sources ($V = 0.350$ for four classes, and $V = 0.314$ for dichotomization).

Regarding heterogeneity, for three sources, voxel entropies (H) and their 95% confidence interval (CI) were $H(\text{TP}) = -0.986$ [-1.08, -0.891], $H(\text{PP}) = -1.141$ [-1.21, -1.07], $H(\text{R}) = -1.178$ [-1.30, -0.231], $H(\text{S}) = -0.553$ [-0.875, -0.231], with $p = 0.003$ on the one-way ANOVA. The post hoc Games–Howell test resulted in significant differences between S and PP ($p = 0.035$) and S and R ($p = 0.026$), meaning that only S voxels were less heterogeneous than the rest of the classes, and no differences in heterogeneity between TP, PP, and R were found. Voxel entropy was

not correlated with the number of voxels per patient (Spearman's rho = −0.019, p = 0.801), that is, the degree of heterogeneity was not related to the size of the SCER.

3.4 | Visualization

Given that the three-source extraction was the best associated with outcome, we chose this solution to visualize the source maps of each patient overlaid onto the MR reference image. As expected from the entropy and the source contribution results, the metabolic patterns of the SCER were almost always contributed by more than one source, and in some cases there was a remarkable metabolic heterogeneity for the spectra analyzed from the different individual voxels. Representative examples of the maps obtained can be seen in Figure 3, and a full report of the color maps obtained for all 31 cases included is available in Supporting Information (Figures S7–S37). Note that there are SCER areas in the maps that are not colored: this is because the spectra corresponding to these regions were of bad quality and therefore excluded from the analysis. Even so, we detect metabolic heterogeneity. As can be seen in Figure 3, there are patients for whom all voxels have a clear winning spectral pattern (Figure 3A,C, Winning source column), which is also the same source in all or practically all voxels. However, there are cases in which there is heterogeneity in the winning sources for each individual voxel and for the patient's set of SCER voxels (Figure 3B,D; TP cases can even show S1 source in particular areas of the SCER). For a detailed display of each case, please check Supporting Information S7 (Figures S7 to S37).

Figure 4 (left panel) shows the relationship between the winning source contribution for each voxel and each patient as well as the outcome class (TP, PP, S, R), for the three-source extraction. Again, homogeneous and heterogeneous metabolic patterns appear in each of the classes. For three sources, the S3 (proliferative pattern) contributes to TP and PP, while S1 (the lipid pattern with PUFA signal) is mainly found in PP, R, and TP. Finally, S2 (the necrotic pattern) contributes to S and PP.

Figure 4 (right panel) is a 3D scatterplot of the sources' contribution, and the message is the same: TP spectra are distributed in two groups, roughly those with a contribution from S3 lower than 0.5 or higher than 0.5. Most TP spectra have a contribution of more than 0.5 from S1. R spectra have an S1 contribution greater than 0.5 and S2 less than 0.5, whereas most S cases have a contribution from S2 of almost 1 and S1 almost 0. Most PP spectra have S1 contribution greater than 0.5, S2 greater than 0.5, and S3 less than 0.4, but another subset has similar factor contributions to the S spectra. A third subset of PP spectra has contributions of S1 greater than 0.5, S2 less than 0.5, and S3 less than 0.4.
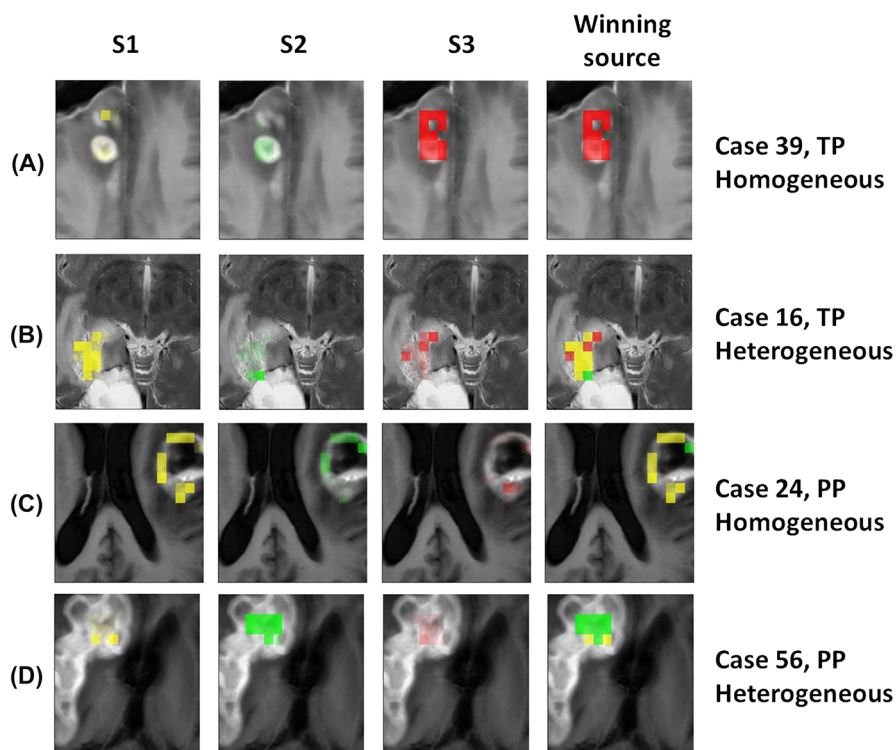

FIGURE 3 Four representative cases for the three-source extraction. Maps are overlaid onto contrast-enhanced T1-weighted image reference.
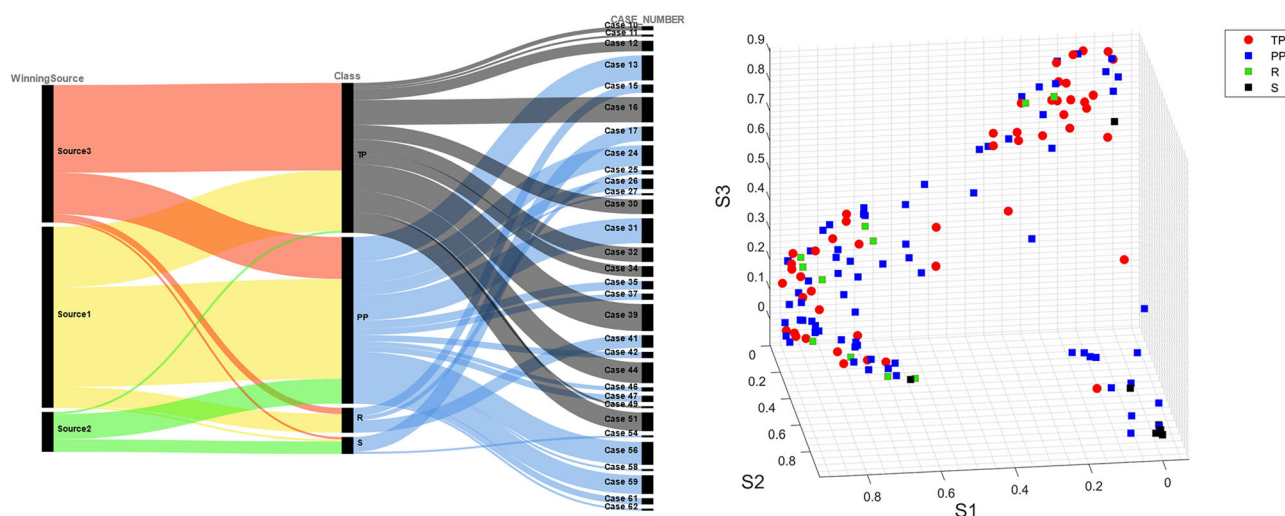
FIGURE 4 Left panel: Sankey diagrams displaying the relationship between the main contribution for each spectrum (winning source) and the class, per case. Right panel: source weight plots for each class in 3D perspective, for all pooled spectra. S1 axis, Source 1 weights; S2 axis, Source 2 weights; S3, Source 3 weights.

TABLE 2 Summary of classification results on the TP versus PP task.

Classifier	Train/test	Classification task	Balanced accuracy	Accuracy	Sensitivity	Specificity	Likelihood ratio (+)	AUC
LDA	Train	Voxel	65.2	65.1 ± 2.7	79.2 ± 2.4	61.2 ± 2.3	2, 0	65.2 ± 2.7
LDA	Test	Voxel	63.7	63.2 ± 2.4	79.1 ± 2.5	41.2 ± 2	1, 3	60.3 ± 2.4
LR	Train	Voxel	66.3	68.4 ± 2.5	66.4 ± 2.6	91.3 ± 2.6	7, 3	66.1 ± 2.5
LR	Test	Voxel	65.4	66.4 ± 2.5	35.3 ± 2.9	91.1 ± 1.8	3, 9	65.2 ± 2.4
RF	Train	Voxel	65.1	75.7 ± 3.0	91.6 ± 2.7	62.5 ± 2.7	2, 4	77.2 ± 3.0
RF	Test	Voxel	63.2	65.2 ± 3.2	91.6 ± 3.5	53.7 ± 2.9	1, 9	64.1 ± 3.2
SVM	Train	Voxel	77.2	75.3 ± 3.0	62.6 ± 3.2	91.2 ± 2.9	6, 9	77.2 ± 3.0
SVM	Test	Voxel	66.8	65.1 ± 3.2	53.7 ± 3.5	91.2 ± 2.7	5, 9	64.1 ± 3.2
LDA	Train	Patient	81.1	83.3 ± 3.9	50.3 ± 4.4	97.1 ± 3.0	16, 7	79.2 ± 3.4
LDA	Test	Patient	79.3	80.8 ± 7.3	50.5 ± 7	90.3 ± 4.0	5, 0	75.1 ± 3.9
LR	Train	Patient	79.1	83.1 ± 6.0	98.2 ± 2.0	50.4 ± 4.4	2, 0	79.4 ± 3.5
LR	Test	Patient	75.2	80.2 ± 7.3	95.3 ± 5.0	50.4 ± 8.0	1, 9	75.1 ± 6.4
RF	Train	Patient	79.2	83.3 ± 3.9	92.2 ± 8.0	66.2 ± 4.4	2, 7	82.1 ± 3.5
RF	Test	Patient	77.3	80.1 ± 7.3	92.3 ± 8.0	66.2 ± 8.0	2, 7	81.2 ± 6.3
SVM	Train	Patient	83.4	80.7 ± 1.2	92.2 ± 4.4	66.1 ± 3.6	2, 7	83.3 ± 7.0
SVM	Test	Patient	79.2	83.2 ± 3.9	92.5 ± 4.4	66.3 ± 3.6	2, 7	79.2 ± 3.8

Note: Balanced accuracy, accuracy, sensitivity, specificity, positive likelihood ratio, and AUC in % ± 95% CI.

3.5 | MLe classification analysis

Results of the TP versus PP classification task are shown in Table 2. Results of TP versus the rest are available in Table S2. For the voxel-based approach, RF had over 90% sensitivity (true positive detection) on both training and test sets at the cost of a moderate specificity (true negative detection), while LR and SVM had over 90% specificity on both training and test sets at the cost of a suboptimal sensitivity. The best overall results (AUC) were achieved by RF and SVM on the training set, but in the test set all models were in the 60% AUC range. Results were better when the patient-based approach was chosen, with balanced accuracies in the 80% range both for training and test set, irrespective of the model. Taking the patient into account, LDA was the best technique, with likelihood ratios of 16.7 in the training set and 5.0 in the test set. Sensitivity was over 92% in all models except LDA, for both the training and the test sets, while LDA worked in the opposite way (true negative detection

over 90% at the cost of almost random detection of true positives). In general, LDA had high specificity at the cost of a lower sensitivity, in contrast to LR, RF, and SVM.

4 | DISCUSSION

In this work, we retrospectively evaluated in 31 consecutive glioblastoma patients whether a conventional PRESS MV MRS exploration, coupled with an MLe-based classification analysis, would be predictive of early progression. The MRS pattern is not fully predictive, although there is an association with outcome. The reason is that most patients present metabolic heterogeneity. Heterogeneity is also the reason why the MLe approach does not have optimal results, although as we will see later our results are comparable to those in the literature.

We chose a blind-source separation approach, instead of directly trying supervised MLe, as we had previously found that the unsupervised approach was better than supervised MLe methods on their own in the prediction of follow-up, both in preclinical models[26,29] and in a preliminary study on a smaller set of these patients.[30]

The blind-source separation approach used, uncovered three robust metabolic patterns: one that is characteristic of a proliferative status, with choline as the most intense peak and lipid macromolecules at 0.9, 1.28, and 2–2.5 ppm respectively—the S3 source. Another pattern was typical of necrosis (the S2 source) and there was also a pattern very similar to S2 in terms of the 0.9 and 1.28 ppm peaks—the S1 source—with weaker signals from choline at 3.21 ppm and creatine at 3.03, and a 2.8 ppm peak tentatively corresponding to PUFAs. S1 was extremely similar to the spectral pattern that we and others had already found to be consistently associated with response to therapy in preclinical models of glioblastoma.[27–29,31–34]

The three patterns were associated with patient outcome but not in a univocal way. This was explained by the marked heterogeneity we detected when plotting the color maps, in the 3D scatterplots, and in the Sankey diagrams.

The observed intrapatient heterogeneity also explains the MLe results: we obtained either good (>90%) specificity or good sensitivity but not both, depending on the classification technique. LDA seemed to achieve more balanced results between sensitivity and specificity, albeit with a slightly lower performance. A similar result with univariate statistics using peak ratio values was also obtained in a recent study by Sidibe et al.,[35] with the multicenter prospective randomized Phase III NCT01507506 *SpectroGlio* trial, in which the Cho/Cr had sensitivity 83% and specificity 64%, and lactate/NAA had sensitivity 100% and specificity 63%. Our classification results were also comparable in terms of AUC to the study by Verma et al.[3] when considering the data on a patient basis. In general, distinguishing TP from PP remains difficult, and despite promising results no single or combined technique can unequivocally distinguish between these two entities. In recent reviews,[36–38] it is shown that there is a lack of a quantitative imaging biomarker for PP, despite promising results with different metabolic and physiological imaging techniques, one important pitfall being that cohorts of patients are small and sampled at different time points. In most studies, sensitivities and specificities fall in our range: 70%–90% (see tab. 1 in Reference 36), also sometimes when sensitivity is more than 90% specificity drops and vice versa. For spectroscopy in particular, Le Fèvre et al.[37] also report substantial variations in sensitivities (54%–100%), and specificities (40%–100%), similarly to a previous meta-analysis from 2014[39] focused on distinguishing recurrent glioma from radiation necrosis. All previous studies used metabolite ratios[35–37] in contrast to our approach and Reference 30, and as pointed out in Reference 39 there is an ample variety of techniques (SV, MV, 1.5 T, 3 T) and metabolite ratios, as well as combinations of them and varieties in the thresholds used by the different studies. In our study, we also found significant differences in the Cho/N-acetyl-aspartate-containing compound (NAC) ratio of the TP versus PP, but we deem this to be a limited picture of the real extent of the metabolic pattern differences in the spectral pattern.

With respect to the metabolic patterns identified, we can compare our results with those in the study by Verma et al.,[3] in which the authors acquired echo planar spectroscopic imaging super-resolution MRS data. They segmented the regions into contrast-enhancing, immediate peritumoral, and distal peritumoral regions, computing a single value of a ratio (Cho/Cr, Cho/NAA) for the entire segmented region. In this way, they found that there were significant differences in the Cho/Cr ratios between TP and PP. Their cohort of 27 patients covered a wider temporal range (6 months post concomitant chemoradiotherapy), with only eight patients having been explored at the same timepoint as in our study. In this sense, since glioblastoma evolves quickly, these authors might have worked with a dataset enriched in more "end-point" metabolic characteristics than ours. Despite these differences, when fig. 2 of their study is carefully evaluated, it can be appreciated that in the spectra of one PP patient, from the overall contrast-enhancing region, there is also a 2.8 ppm peak, whereas it is absent from the spectrum of the TP patient. Their work focused on the Cho/Cr and the Cho/NAC ratios, and they did not evaluate the lipid signals, although that could have been possible with their 17 ms T_E.

In another study, Sawlani et al.[40] performed a descriptive study on five patients (four PP and one TP) of the single voxel spectroscopy pattern at short (35 ms) and long (144 ms) T_E, with the same purpose. PP patients were explored 3 months after therapy. In their study, they acquired one single voxel from the whole suspicious area. Strikingly, at short T_E (close to our $T_E = 30$ ms), all four PP patients presented with the same pattern as our S1, including a very prominent 2.8 ppm PUFA peak. In addition, three of these four patients were alive after 1.5 years of the exploration. Their TP patient had the same pattern as our S3, with high choline and no PUFA signal. Our interpretation of their results is that they studied

patients more evolved into the disease than we had (our P1M versus their P3M), and sampled the global metabolic pattern of the SCER with a single voxel, while we captured the heterogeneity with an MV acquisition.

In addition, in previous studies, patient cohorts comprise a wide range of time points after the end of the radiotherapy[36,41]: for example, Elias et al. in 2011[42] monitored patients from 2 to 41 months after therapy in order to distinguish recurrence and radiation injury. Similarly, the Sawlani et al.[40] study used the P3M timepoint, and the Verma et al.[3] study focused on the 6 month period after concomitant treatment, just to name a few examples, making it difficult to establish a fair comparison in terms of the metabolic pattern through time, which has also been acknowledged recently in the previously mentioned *AVAglio* trial.[8]

Interestingly, there are also a few studies in which authors implicitly assume the existence of heterogeneity, and deal with it by considering a "mixed tumors" group—validating their results with biopsy specimen analysis. For example, Kim et al.[43] analyzed histograms of normalized perfusion cerebral blood volume values in a retrospective cohort of 39 patients with two or more explorations showing SCERs, in which there were 14 recurrences, 10 mixed tumors, and 15 patients with post-treatment changes, and they obtained sensitivities and specificities greater than 90%. In a previous study,[44] there were 21 TP, 8 PP, and 12 mixed tumors, which were analyzed by combining diffusion tensor imaging (DTI) and dynamic susceptibility contrast MRI (DSC-MRI). As can be seen, the percentages of heterogeneous tumors in these studies are 25% and 19% respectively. In our study, as per the winning source maps, we have heterogeneous metabolic patterns in 18 of the 31 cases (58%), and this is probably an underestimation as we had to exclude many voxels because of low quality in a proportion of cases, having only a partial representation of the metabolome in the entire SCERs. In addition, in the previous study,[44] the cohort was obtained over a longer time frame than was our patient set, as patients were included if they had an SCER detected 6 months after therapy. Patients were biopsied within 2 weeks after the discovery of the SCERs. The mixed tumor category was established when 25%–75% malignant features were identified in the histopathological specimens, and PP was when less than 25% malignant features were identified. Among histological features related to treatment response rather than malignancy, these authors considered lymphocyte and macrophage infiltrates.

Regarding the clinical validation of the progression status, we followed the RANO group guidelines,[9] which suggest discriminating the two entities in a follow-up examination done at month +3. Additionally, the RANO group accepts that we lack robust evidence about PET, perfusion-weighted imaging, and MRS for this differentiation. Histopathology is also known to be heterogeneous in these cases, and necrosis and viable tumor coexist on most surgical specimens. This is in full agreement with our results, in which we show this non-invasively.

Gliomas are infiltrated by glioma-associated microglia/macrophages (GAMs) during their progression,[45] and infiltration by $CD4^+$ and $CD8^+$ lymphocytes has also been reported in PP cases.[46] We have previously demonstrated, in a preclinical model, that the spectroscopic response pattern was associated with macrophage and lymphocyte infiltration,[28] particularly by GAMs with increased macrophage M1/M2 ratios and PD-L1 expression in responding mice.[33] Macrophage polarization to the M1 phenotype involves a full metabolic reprogramming, as has been recently demonstrated by one-dimensional nuclear Overhauser effect spectroscopy MRS and hyperpolarization.[47] In this sense, it could be hypothesized that we might be clinically detecting the same phenomenon. It has been reported that GAMs can contribute up to 30%–50% to the tumor microenvironment.[38]

Among the limitations of our study, we assume the limited sampling of the SCERs, due to the low spectral quality, which is to be expected since we were often sampling the peri-tumoral cavity, and in several other cases difficult locations near the ventricles, but even so, we detected heterogeneous patterns. In this sense, we may even be underestimating the true metabolic heterogeneity. Another limitation is the use of the standard PRESS package provided by the manufacturer in the public hospital where the data were acquired. Possible improvements should involve the harmonization of the semi-LASER sequence[48,49] in the standard vendors' packages, to allow for future multicenter studies and decrease the proportion of discarded spectra. Even with the above-mentioned limitations, the standard PRESS sequence, as performed in clinical routine, allowed us to detect meaningful metabolic information that was not otherwise available with other techniques.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORCID

Gülnur Ungan https://orcid.org/0000-0002-5436-4665

Albert Pons-Escoda https://orcid.org/0000-0003-4167-8291

Daniel Ulinic https://orcid.org/0009-0006-3318-0839

Carles Arús https://orcid.org/0000-0003-2510-2671

Sandra Ortega-Martorell https://orcid.org/0000-0001-9927-3209

Ivan Olier https://orcid.org/0000-0002-5679-7501

Alfredo Vellido https://orcid.org/0000-0002-9843-1911

Carles Majós https://orcid.org/0000-0003-0468-5150

Margarida Julià-Sapé https://orcid.org/0000-0002-3316-9027

REFERENCES

1. Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med*. 2005;352: 987-996. doi:10.1056/NEJMoa043330
2. D'Amore F, Grinberg F, Mauler J, et al. Combined ^{18}F-FET PET and diffusion kurtosis MRI in posttreatment glioblastoma: differentiation of true progression from treatment-related changes. *Neuro-Oncol Adv*. 2021;3(1):vdab044. doi:10.1093/noajnl/vdab044
3. Verma G, Chawla S, Mohan S, et al. Three-dimensional echo planar spectroscopic imaging for differentiation of true progression from pseudoprogression in patients with glioblastoma. *NMR Biomed*. 2019;32(2):e4042. doi:10.1002/nbm.4042
4. Weybright P, Maly P, Gomez-Hassan D, Blaesing C, Sundgren PC. MR spectroscopy in the evaluation of recurrent contrast-enhancing lesions in the posterior fossa after tumor treatment. *Neuroradiology*. 2004;46(7):541-549. doi:10.1007/s00234-004-1195-1
5. Zeng QS, Li CF, Zhang K, Liu H, Kang XS, Zhen JH. Multivoxel 3D proton MR spectroscopy in the distinction of recurrent glioma from radiation injury. *J Neurooncol*. 2007;84(1):63-69. doi:10.1007/s11060-007-9341-3
6. Matsusue E, Fink JR, Rockhill JK, Ogawa T, Maravilla KR. Distinction between glioma progression and post-radiation change by combined physiologic MR imaging. *Neuroradiology*. 2010;52(4):297-306. doi:10.1007/s00234-009-0613-9
7. Elshafeey N, Kotrotsou A, Hassan A, et al. Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudoprogression in glioblastoma. *Nat Commun*. 2019;10(1):3170. doi:10.1038/s41467-019-11007-0
8. Hagiwara A, Schlossman J, Shabani S, et al. Incidence, molecular characteristics, and imaging features of "clinically-defined pseudoprogression" in newly diagnosed glioblastoma treated with chemoradiation. *J Neurooncol*. 2022;159(3):509-518. doi:10.1007/s11060-022-04088-3
9. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: Response Assessment in Neuro-Oncology working group. *J Clin Oncol*. 2010;28(11):1963-1972. doi:10.1200/JCO.2009.26.3541
10. Wen PY, Chang SM, Van den Bent MJ, Vogelbaum MA, Macdonald DR, Lee EQ. Response assessment in neuro-oncology clinical trials. *J Clin Oncol*. 2017;35(21):2439-2449. doi:10.1200/JCO.2017.72.7511
11. Mocioiu V, Ortega-Martorell S, Olier I, et al. From raw data to data-analysis for magnetic resonance spectroscopy—the missing link: jMRUI2XML. *BMC Bioinformatics*. 2015;16(1):378. doi:10.1186/s12859-015-0796-5
12. Stefan D, Cesare FD, Andrasescu A, et al. Quantitation of magnetic resonance spectroscopy signals: the jMRUI software package. *Meas Sci Technol*. 2009;20(10):104035. doi:10.1088/0957-0233/20/10/104035
13. Tate AR, Underwood J, Acosta DM, et al. 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(4):411-434. doi:10.1002/nbm.1016
14. Pérez-Ruiz A, Julià-Sapé M, Mercadal G, Olier I, Majós C, Arús C. The INTERPRET Decision-Support System version 3.0 for evaluation of magnetic resonance spectroscopy data from human brain tumours and other abnormal brain masses. *BMC Bioinformatics*. 2010;11:581. doi:10.1186/1471-2105-11-581
15. Ding C, Li T, Jordan M. Convex and semi-nonnegative matrix factorizations. *IEEE Trans Pattern Anal Mach Intell*. 2010;32:45-55. doi:10.1109/TPAMI.2008.277
16. Ortega-Martorell S, Lisboa PJG, Vellido A, Julià-Sapé M, Arús C. Non-negative matrix factorisation methods for the spectral decomposition of MRS data from human brain tumours. *BMC Bioinformatics*. 2012;13:38. doi:10.1186/1471-2105-13-38
17. Ortega-Martorell S, Lisboa P, Vellido A, et al. Convex non-negative matrix factorization for brain tumor delimitation from MRSI data. *PLoS ONE*. 2012; 7(10):e47824. doi:10.1371/journal.pone.0047824
18. Hernández-Villegas Y, Ortega-Martorell S, Arús C, Vellido A, Julià-Sapé M. Extraction of artefactual MRS patterns from a large database using non-negative matrix factorization. *NMR Biomed*. 2022;35(4):e4193. doi:10.1002/nbm.4193
19. Edden RAE, Puts NAJ, Harris AD, Barker PB, Evans CJ. Gannet: a batch-processing tool for the quantitative analysis of gamma-aminobutyric acid-edited MR spectroscopy spectra. *J Magn Reson Imaging*. 2014;40(6):1445-1452. doi:10.1002/jmri.24478
20. Raudys S, Pikelis V. On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans Pattern Anal Mach Intell*. 1980;2(3):242-252. doi:10.1109/TPAMI.1980.4767011
21. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans Pattern Anal Mach Intell*. 1991;13(3):252-264. doi:10.1109/34.75512
22. García-Gómez JM, Luts J, Julià-Sapé M, et al. Multiproject–multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. *Magn Reson Mater Phys Biol Med*. 2009;22(1):5-18. doi:10.1007/s10334-008-0146-y

23. Cramér H. *Mathematical Methods of Statistics*. Princeton University Press; 1946.

24. jamovi. Accessed May 12, 2023. https://cloud.jamovi.org/

25. Pedregosa F, Varoquaux G, Gramfort A, et al. scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

26. Delgado-Goñi T, Julià-Sapé M, Candiota AP, Pumarola M, Arús C. Molecular imaging coupled to pattern recognition distinguishes response to temozolomide in preclinical glioblastoma. *NMR Biomed*. 2014;27(11):1333-1345. doi:10.1002/nbm.3194

27. Arias-Ramos N, Ferrer-Font L, Lope-Piedrafita S, et al. Metabolomics of therapy response in preclinical glioblastoma: a multi-slice MRSI-based volumetric analysis for noninvasive assessment of temozolomide treatment. *Metabolites*. 2017;7(2):20. doi:10.3390/metabo7020020

28. Wu S, Calero-Pérez P, Villamañan L, et al. Anti-tumour immune response in GL261 glioblastoma generated by temozolomide immune-enhancing metronomic schedule monitored with MRSI-based nosological images. *NMR Biomed*. 2020;33(4):e4229. doi:10.1002/nbm.4229

29. Delgado-Goñi T, Ortega-Martorell S, Ciezka M, et al. MRSI-based molecular imaging of therapy response to temozolomide in preclinical glioblastoma using source analysis. *NMR Biomed*. 2016;29(6):732-743. doi:10.1002/nbm.3521

30. Julià-Sapé M, Pons A, Arús C, Gil-Gil M, Majós C. MRSI and a supervised classifier predict treatment response in glioblastoma: preliminary results. *Magn Reson Mater Phys Biol Med*. 2017;30:S480. doi:10.1007/s10334-017-0632-1

31. Hakumäki JM, Poptani H, Sandmair AM, Ylä-Herttuala S, Kauppinen RA. ^{1}H MRS detects polyunsaturated fatty acid accumulation during gene therapy of glioma: implications for the in vivo detection of apoptosis. *Nat Med*. 1999;5(11):1323-1327. doi:10.1038/15279

32. Ferrer-Font L, Arias-Ramos N, Lope-Piedrafita S, et al. Metronomic treatment in immunocompetent preclinical GL261 glioblastoma: effects of cyclophosphamide and temozolomide. *NMR Biomed*. 2017;30(9):e3748. doi:10.1002/nbm.3748

33. Calero-Pérez P, Wu S, Arús C, Candiota AP. Immune system-related changes in preclinical GL261 glioblastoma under TMZ treatment: explaining MRSI-based nosological imaging findings with RT-PCR analyses. *Cancer*. 2021;13(11):2663. doi:10.3390/cancers13112663

34. Wu S, Calero-Pérez P, Arús C, Candiota AP. Anti-PD-1 immunotherapy in preclinical GL261 glioblastoma: influence of therapeutic parameters and non-invasive response biomarker assessment with MRSI-based approaches. *Int J Mol Sci*. 2020;21(22):8775. doi:10.3390/ijms21228775

35. Sidibe I, Tensaouti F, Gilhodes J, et al. Pseudoprogression in GBM versus true progression in patients with glioblastoma: a multiapproach analysis. *Radiother Oncol*. 2023;181:109486. doi:10.1016/j.radonc.2023.109486

36. Chawla S, Bukhari S, Afridi OM, et al. Metabolic and physiologic magnetic resonance imaging in distinguishing true progression from pseudoprogression in patients with glioblastoma. *NMR Biomed*. 2022;35(7):e4719. doi:10.1002/nbm.4719

37. Le Fèvre C, Constans JM, Chambrelant I, et al. Pseudoprogression versus true progression in glioblastoma patients: a multiapproach literature review. Part 2—radiological features and metric markers. *Crit Rev Oncol Hematol*. 2021;159:103230. doi:10.1016/j.critrevonc.2021.103230

38. Hambardzumyan D, Gutmann DH, Kettenmann H. The role of microglia and macrophages in glioma maintenance and progression. *Nat Neurosci*. 2016;19(1):20-27. doi:10.1038/nn.4185

39. Zhang H, Ma L, Wang Q, Zheng X, Wu C, Xu B-n. Role of magnetic resonance spectroscopy for the differentiation of recurrent glioma from radiation necrosis: a systematic review and meta-analysis. *Eur J Radiol*. 2014;83(12):2181-2189. doi:10.1016/j.ejrad.2014.09.018

40. Sawlani V, Taylor R, Rowley K, Redfern R, Martin J, Poptani H. Magnetic resonance spectroscopy for differentiating pseudo-progression from true progression in GBM on concurrent chemoradiotherapy. *Neuroradiol J*. 2012;25(5):575-586. doi:10.1177/197140091202500511

41. Sidibe I, Tensaouti F, Roques M, Cohen-Jonathan-Moyal E, Laprie A. Pseudoprogression in glioblastoma: role of metabolic and functional MRI—systematic review. *Biomedicine*. 2022;10(2):285. doi:10.3390/biomedicines10020285

42. Elias AE, Carlos RC, Smith EA, et al. MR spectroscopy using normalized and non-normalized metabolite ratios for differentiating recurrent brain tumor from radiation injury. *Acad Radiol*. 2011;18(9):1101-1108. doi:10.1016/j.acra.2011.05.006

43. Kim HS, Kim JH, Kim SH, Cho KG, Kim SY. Posttreatment high-grade glioma: usefulness of peak height position with semiquantitative MR perfusion histogram analysis in an entire contrast-enhanced lesion for predicting volume fraction of recurrence. *Radiology*. 2010;256(3):906-915. doi:10.1148/radiol.10091461

44. Wang S, Martinez-Lage M, Sakai Y, et al. Differentiating tumor progression from pseudoprogression in patients with glioblastomas using diffusion tensor imaging and dynamic susceptibility contrast MRI. *Am J Neuroradiol*. 2016;37(1):28-36. doi:10.3174/ajnr.A4474

45. Gieryng A, Pszczolkowska D, Walentynowicz KA, Rajan WD, Kaminska B. Immune microenvironment of gliomas. *Lab Invest*. 2017;97(5):498-518. doi:10.1038/labinvest.2017.19

46. Han S, Zhang C, Li Q, et al. Tumour-infiltrating CD4^{+} and CD8^{+} lymphocytes as predictors of clinical outcome in glioma. *Br J Cancer*. 2014;110(10):2560-2568. doi:10.1038/bjc.2014.162

47. Qiao K, Le Page LM, Chaumeil MM. Non-invasive differentiation of M1 and M2 activation in macrophages using hyperpolarized ^{13}C MRS of pyruvate and DHA at 1.47 Tesla. *Metabolites*. 2021;11(7):410. doi:10.3390/metabo11070410

48. Wilson M, Andronesi O, Barker PB, et al. Methodological consensus on clinical proton MRS of the brain: review and recommendations. *Magn Reson Med*. 2019;82(2):527-550. doi:10.1002/mrm.27742

49. Deelchand DK, Berrington A, Noeske R, et al. Across-vendor standardization of semi-LASER for single-voxel MRS at 3T. *NMR Biomed*. 2021;34(5):e4218. doi:10.1002/nbm.4218

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.