HAL
open science

Computation of reliable textural indices from multimodal brain MRI: suggestions based on a study of patients with diffuse intrinsic pontine glioma

Jessica Goya-Outi, Fanny Orlhac, Raphael Calmon, Agusti Alentorn, Christophe Nioche, Cathy Philippe, Stéphanie Puget, Nathalie Boddaert, Irène Buvat, Jacques Grill, et al.

To cite this version:

Jessica Goya-Outi, Fanny Orlhac, Raphael Calmon, Agusti Alentorn, Christophe Nioche, et al.. Computation of reliable textural indices from multimodal brain MRI: suggestions based on a study of patients with diffuse intrinsic pontine glioma. Physics in Medicine and Biology, 2018, 63 (10), pp.105003. 10.1088/1361-6560/aabd21 . hal-01801219

HAL Id: hal-01801219

https://hal.sorbonne-universite.fr/hal-01801219v1

Submitted on 28 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Physics in Medicine & Biology

PAPER

Computation of reliable textural indices from multimodal brain MRI: suggestions based on a study of patients with diffuse intrinsic pontine glioma

To cite this article: Jessica Goya-Outi *et al* 2018 *Phys. Med. Biol.* **63** 105003

View the article online for updates and enhancements.

Related content

- Applications and limitations of radiomics
  Stephen S F Yip and Hugo J W L Aerts
- Enhancement of multimodality texture-based prediction models via optimization of PET and MR image acquisition protocols: a proof of concept
  Martin Vallières, Sébastien Laberge, André Diamant et al.
- Significance of the impact of motion compensation on the variability of PET image features
  M Carles, T Bach, I Torres-Espallardo et al.

This content was downloaded from IP address 134.157.80.157 on 28/05/2018 at 10:51

IOP Publishing Phys. Med. Biol. 63 (2018) 105003 (12pp) https://doi.org/10.1088/1361-6560/aabd21

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

CrossMark

RECEIVED
15 November 2017

REVISED
26 March 2018

ACCEPTED FOR PUBLICATION
10 April 2018

PUBLISHED
10 May 2018

PAPER

Computation of reliable textural indices from multimodal brain MRI: suggestions based on a study of patients with diffuse intrinsic pontine glioma

Jessica Goya-Outi[1], Fanny Orlhac[1], Raphael Calmon[2], Agusti Alentorn[3,4], Christophe Nioche[1], Cathy Philippe[4], Stéphanie Puget[5], Nathalie Boddaert[2], Irène Buvat[1], Jacques Grill[6], Vincent Frouin[4] and Frederique Frouin[1]

[1] IMIV, Inserm, CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France
[2] Pediatric Radiology Department, Hôpital Necker Enfants Malades, AP-HP; Inserm U1000; Imagine Institut des Maladies Génétiques, UMR 1163; Université Paris Descartes, ComUE Sorbonne Paris Cité, Paris, France
[3] Neurology Department, Groupe Hospitalier Pitié-Salpêtrière, APHP; ICM, Inserm, CNRS, Sorbonne Universités, UPMC Univ. Paris 6, Paris, France
[4] UNATI, Neurospin, CEA, Université Paris-Saclay, Gif-sur-Yvette, France
[5] Pediatric Neurosurgery Department, Hôpital Necker Enfants Malades, AP-HP, Paris, France
[6] Département de Cancérologie de l'Enfant et de l'Adolescent, Gustave Roussy; CNRS UMR 8203, Université Paris Saclay, Villejuif, France

E-mail: jessicaouti@gmail.com

Keywords: radiomics, texture analysis, diffuse intrinsic pontine glioma, 3D MRI

Abstract

Few methodological studies regarding widely used textural indices robustness in MRI have been reported. In this context, this study aims to propose some rules to compute reliable textural indices from multimodal 3D brain MRI. Diagnosis and post-biopsy MR scans including T1, post-contrast T1, T2 and FLAIR images from thirty children with diffuse intrinsic pontine glioma (DIPG) were considered. The hybrid white stripe method was adapted to standardize MR intensities. Sixty textural indices were then computed for each modality in different regions of interest (ROI), including tumor and white matter (WM). Three types of intensity binning were compared d_i: constant bin width and relative bounds; d_{ii} : constant number of bins and relative bounds; d_{iii} : constant number of bins and absolute bounds. The impact of the volume of the region was also tested within the WM. First, the mean Hellinger distance between patient-based intensity distributions decreased by a factor greater than 10 in WM and greater than 2.5 in gray matter after standardization. Regarding the binning strategy, the ranking of patients was highly correlated for 188/240 features when comparing d_i with d_{iii}, but for only 20 when comparing d_i with d_{ii}, and nine when comparing d_{ii} with d_{iii}. Furthermore, when using d_i or d_{iii}, texture indices reflected tumor heterogeneity as assessed visually by experts. Last, 41 features presented statistically significant differences between contralateral WM regions when ROI size slightly varies across patients, and none when using ROI of the same size. For regions with similar size, 224 features were significantly different between WM and tumor. Valuable information from texture indices can be biased by methodological choices. Recommendations are to standardize intensities in MR brain volumes, to use intensity binning with constant bin width, and to define regions with the same volumes to get reliable textural indices.

Introduction

The advent of radiomics generates a lot of enthusiasm and development, especially for oncological studies. Radiomics helps in revealing tumor characteristics or predicting prognosis through the extraction of a great number of imaging indices inside the tumor area (Lambin *et al* 2012, Gevaert *et al* 2014, Aerts 2016, Gillies *et al* 2016, Skogen *et al* 2016). These indices are based on tumor shape, on first-order statistics, e.g. mean intensity, standard deviation, histogram-derived indices and on high-order statistics including textural indices derived

© 2018 Institute of Physics and Engineering in Medicine

from the gray level co-occurrence matrix (GLCM) (Haralick *et al* 1973), the gray level run-length matrix (GLRLM) (Tang 1998) or the gray level size-zone matrix (GLSZM) (Thibault *et al* 2009). Although these textural indices have been widely used in medical imaging, their variability as a function of the range and binning of image intensity and the importance of the volume of the region of interest in which they are computed has not been extensively studied, especially in MRI.

In PET imaging, several factors influencing textural indices computation have been recently studied. If the conclusions regarding some parameters including acquisition parameters and reconstruction methods (Yan *et al* 2015) are likely to be specific to PET studies, others may be relevant whatever the imaging modality. Among the possible sources of variation, it was previously shown that the volume of the region of interest (Orlhac *et al* 2014, Hatt *et al* 2015) and the intensity binning method (Leijenaar *et al* 2015, Orlhac *et al* 2015) play an important role.

Only a few methodological studies regarding the robustness of texture analysis in MRI have been reported so far. In Larroza *et al* (2016), the authors recommended the use of ROIs of the same size when the range of lesion sizes among samples is broad. In Molina *et al* (2017), it was showed that textural indices computed with different dynamic ranges of intensities and spatial resolutions are not comparable. In Mayerhoefer *et al* (2009), using phantoms with nodular pattern, the authors found that textural indices were sensitive to variations in acquisition parameters of T2 scans, especially with the spatial resolution.

Unlike PET images, MR images are acquired in arbitrary units and intensities values vary between visits and patients, even when using the same machine and the same protocol. Standardizing intensities for each modality among all patients and visits is a necessary step for a robust inter-patient comparison. Among different methods of standardization proposed for cerebral MRI studies (Nyul *et al* 2000, Christensen 2003, Madabhushi and Udupa 2006, Shinohara *et al* 2014), the hybrid white stripe (hWS) method (Shinohara *et al* 2014) proved to be successful for neurodegenerative diseases (Shinohara *et al* 2014) and oncological studies (Kickingereder *et al* 2016a).

In this context, the present study aimed at defining some rules to compute reliable textural indices from structural MRI. The experimental study was carried out on diffuse intrinsic pontine glioma (DIPG), a rare pediatric tumor with extremely bad prognostic (Cohen *et al* 2017), for which MR images remain our best mean of diagnosis and follow-up during treatment. Using these data, four questions were addressed: (i) the standardization of the MR intensities adapted to the DIPG images before textural index calculation; (ii) the impact of the intensity binning method; (iii) the dependency of textural indices with the volume of the region in which they are calculated; (iv) the ability of the different textural indices to distinguish between tumor and white matter (WM).

Subjects and methods

Patients and magnetic resonance imaging (MRI)

Images from thirty children (17 girls and 13 boys) from 3 to 15 years old (mean age 8 $\pm$ 3 years) diagnosed with DIPG were included in this retrospective study. Patients' families authorized the use of clinical and radiological data recorded during treatment for research purposes. Informed consent was obtained according to the IRB of the reference institution. All diagnoses were confirmed histologically and molecularly according to current criteria (Castel *et al* 2015). All patients underwent diagnostic scans including four structural MR modalities: T1 weighted, post-contrast T1 weighted (T1c), T2 weighted and FLAIR images, and post-biopsy scans that included at least one T1 or one T2 weighted scan. All MR acquisitions were performed using a 1.5 Tesla magnet (Signa HDxt/Discovery MR450, GE Medical Systems, Milwaukee, WI, USA) with almost identical acquisition parameters between studies (see details in table 1).

MR intensity standardization

The workflow of MR Intensity standardization is detailed in figure 1. Field inhomogeneity was corrected by means of the N4 algorithm (Tustison *et al* 2010). For each patient, the different images (pre and post-biopsy) were then realigned to the T2 diagnostic scan (subject's space) using rigid registration in FSL-FLIRT (Jenkinson *et al* 2002). Registrations were systematically checked visually and were always found to be satisfactory.

Intensities of the different modalities were standardized among all patients and visits using the hWS (Shinohara *et al* 2014), which is a segmentation-free standardization method based on the WM distribution. Starting from the R package 'WhiteStripe: WM Normalization for Magnetic Resonance Images using WhiteStripe, Version 1.1.1', some modifications were introduced to better take into account the specificities of our data. In the hWS method, the smoothed histogram corresponding to slices 80–120 in the MNI reference image is computed to estimate a WM mode noted $\mu_{i,j}^{m*}$, for patient i, visit j, modality m as the last peak of the histogram for T1 weighted scans and as the highest peak for T2 weighted scans. Selecting these slices as the reference region for standardization is convenient for patients with DIPG since they generally do not contain tumor. In this study however, in order to find the corresponding slices in each subject's space, diffeomorphic transformations (Avants *et al* 2007) were chosen to realign scans with the 1 mm^3 MNI template, as patients are children. Moreover,

Table 1. Acquisition parameters for T1, post injection T1 (T1c), T2 and FLAIR images.

	T1[a] (51 scans)	T1c[b] (47 scans)	T2 (57 scans)	FLAIR[c] (43 scans)
TE (ms)	3.6–5.4	3.6–6.8	93.0–115.1	129.9–172.8
TR (ms)	12.0–12.7	12–15.2	2500	8400–9002
TI (ms)	550	550	na	2223–2283
Flip angle	13	13	90	90
Plane	Axial	Axial	Sagittal	Sagittal
Pixel resolution (mm)	0.43 × 0.43–0.47 × 0.47	0.43 × 0.43–0.47 × 0.47	0.43 × 0.43–0.49 × 0.49	0.43 × 0.43–0.53 × 0.53
Slice thickness (mm)	1.4–1.6	1.2–1.6	1	1.2–1.4

[a] Except for one patient visit, where TE = 10.2 ms, TR = 500 ms, acquisition plane is sagittal and slice thickness is 1 mm.
[b] Except for one patient visit, where TE = 10.1 ms, TR = 600 ms, and acquisition plane is sagittal.
[c] Except for one patient visit, where acquisition plane is sagittal and slice thickness is 4 mm.

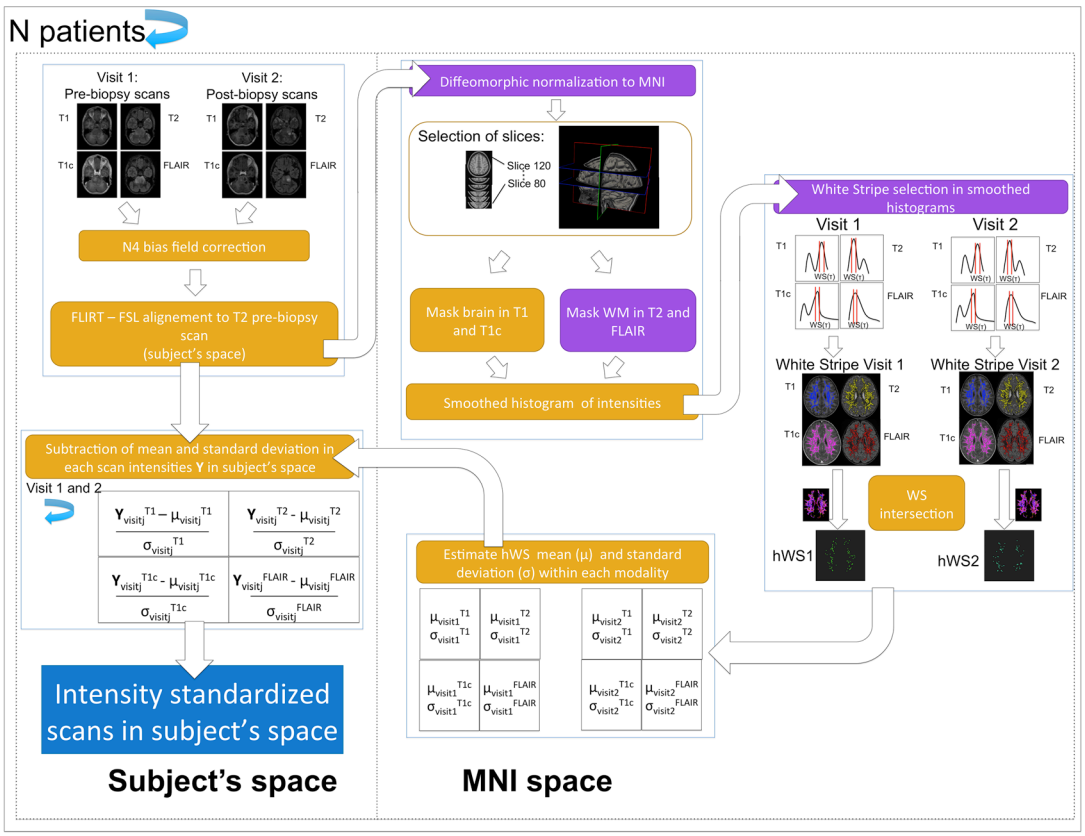

Figure 1. Workflow of MR Intensity standardization. Compared to the hWS method (Shinohara *et al* 2014), the adaptations needed to analyze DIPG scans are shown in purple.

for some T2 weighted scans, as shown in figure 2, the highest histogram peak could correspond to gray matter (GM). Hence a mask in the MNI reference image was defined with a probability of WM greater than 0.3, the WM being defined using tissue probability maps from SPM12 (Ashburner and Friston 2005). Using this constraint ensured that the highest histogram peak corresponds to WM (figure 2). Note that no precise segmentation was required and the method is still segmentation-free.

For each scan, the white stripe (WS) is defined as the voxels in the MNI space having intensity values close enough to the $\mu_{i,j}^{m*}$ (Shinohara *et al* 2014). A parameter called τ defines the proximity between voxels intensities and the WM mode $\mu_{i,j}^{m*}$. Parameter τ is a quantile tolerance around the WM mode in the cumulative distribution function. As the T2 (and FLAIR) scans were masked with WM, quantiles were increased to select the same number of WS voxels in the T2 (and FLAIR) scans as in the T1 weighted scans. The default value of 0.05 for τ was therefore used for T1 and T1c scans, while a value of 0.08 was used for T2 and FLAIR scans. As a result, the number of voxels contained in each WS was similar for each modality (about 100 000 voxels). The hybrid WS (hWS) of patient i, visit j was defined as the intersection of the WS of all modalities acquired (resulting in about 2000 voxels when 4 modalities are acquired), to get the so-called non-affected WM (Shinohara *et al* 2014). Each scan

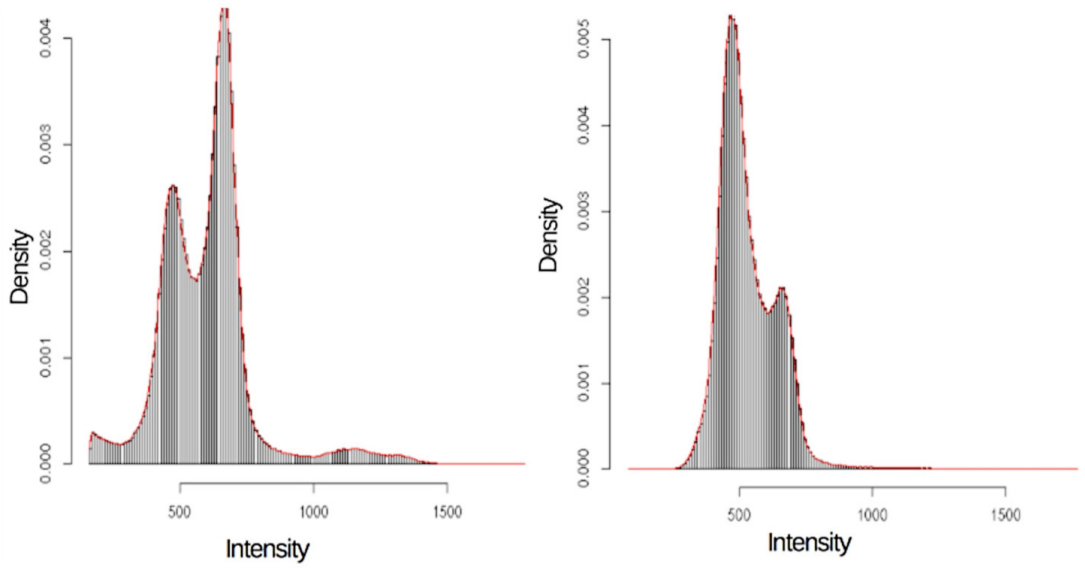

Figure 2. Histogram of slices 80–120 (in the MNI space) of T2 scan intensities for one DIPG patient without (left) and with (right) the application of the WM mask.

of visit j was finally standardized in the subject's space by the subtraction of the mean intensity of the hWS voxels and the division by their standard deviation.

Impact of intensity binning on textural indices in a tumor region

A tumor region was defined in diagnostic scans to evaluate the impact of three intensity binning methods on textural indices. After standardization of intensities and registration to the subject's diagnostic T2 scan, post-biopsy T2 (or T1) scans were subtracted from corresponding diagnostic scans to facilitate biopsy location. Each tumor region was chosen in order to correspond to the biopsy region (tissue extraction of about 1 cm^3), and was approximated by spheres of 4.5 mm radius (ROI_tumor).

Using PyRadiomics software (version 1.1.1) (Van Griethuysen *et al* 2017), 60 textural indices were computed (28 from the GLCM, 16 from the GLSZM and 16 from the GLRLM) inside the ROI_tumor copied on the four realigned diagnostic scans, resampled to isotropic 1 mm^3 voxels. PyRadiomics code was slightly modified to evaluate 3 types of intensity binning, that were proposed for PET or MR studies: (i) d_i, relative bounds, bin width fixed (Leijenaar *et al* 2015). For each ROI, the minimum and maximum intensity values inside the ROI were the binning bounds. The same bin width (bW = 2) was used for all ROIs; (ii) d_{ii}, relative bounds, number of bins fixed. For each ROI, the minimum and maximum intensity values inside the ROI were the binning bounds (Michoux *et al* 2015, Yang *et al* 2015, Molina *et al* 2016, Kickingereder *et al* 2016b). Here, the same number of discretization levels (nBins = 64) was used for all ROIs; (iii) d_{iii}: Absolute bounds, number of bins fixed (Orlhac *et al* 2015). The minimum of the 10 percentile values of the 30 individuals and the maximum of the 90 percentile values of the 30 individuals included in this study gave the binning bounds. The same number of discretization levels (nBins = 64) was used for all ROIs, which implies that the same bin width (variable according to the modality) was also used.

As textural indices are computed to quantify tumor heterogeneity (Orlhac *et al* 2015), three experts labeled the contents of ROI_tumor of each modality independently as homogeneous or heterogeneous, based on its three central slices (axial, sagittal and coronal) (figure 3). Experts were blinded to each other and to the values of textural indices. For each expert, the label was used to calculate the sensitivity and specificity of each texture feature to identify heterogeneous tumor tissue when using the three binning methods.

Impact of the volume of regions of interest (ROI) on textural indices

To evaluate the impact of the volume of the region in which the textural indices are computed, we chose to analyze healthy WM, as it is supposed to be similar in brain contra-lateral analysis and across patients. To avoid ROI positioning in each subject's space and analyze the same proportion of tissue in every patient, two concentric spherical regions (3 and 5 mm radius) were drawn in the WM of the 1 mm^3 MNI brain (figure 4). Employing the diffeomorphic transformation estimated in the intensity standardization process, the spherical regions were pasted to the subject's space scans (ROI_MNI3, ROI_MNI5, figure 4). The resulting ROI volumes varied from 39 to 424 mm^3 according to the sphere radius, the age and the brain size of subjects. Furthermore two spherical ROIs, ROI_4.5R and ROI_4.5L (radius 4.5 mm) were drawn directly in the subject's space such as ROI_4.5R

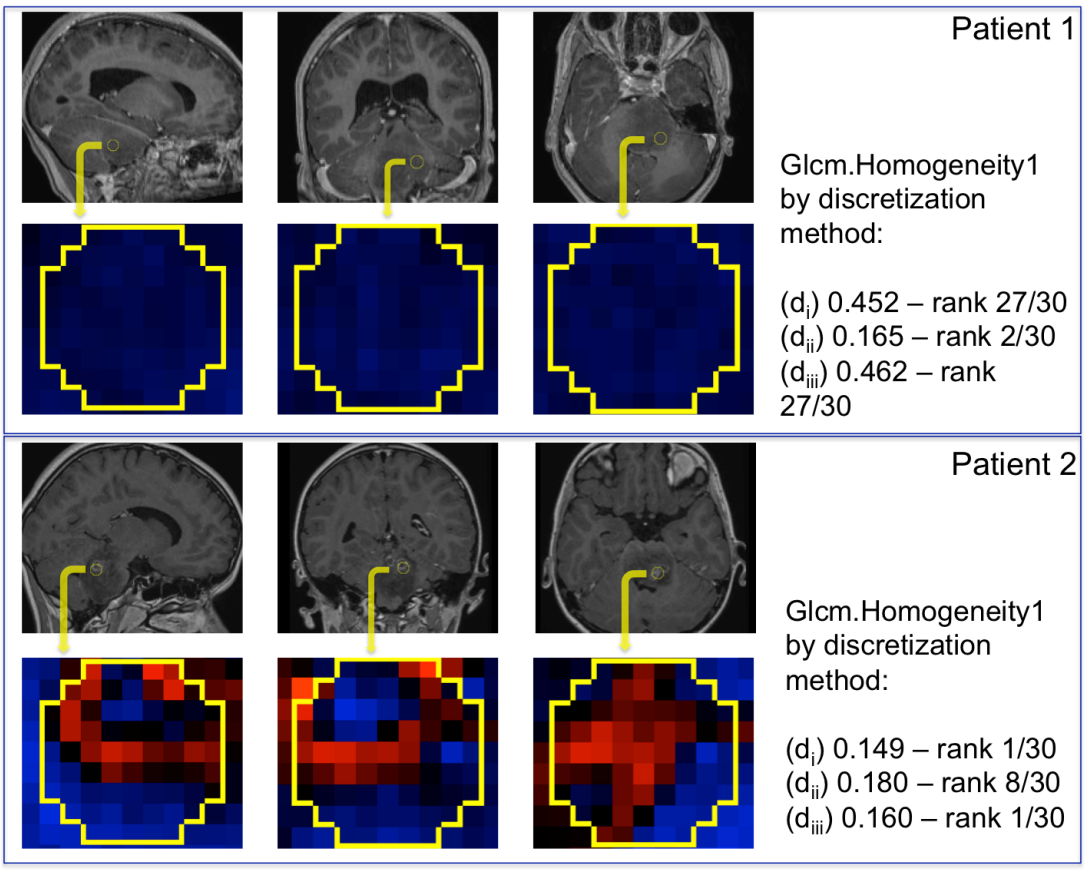

Figure 3. ROI_tumor of two patients in sagittal, coronal and axial views (diagnostic visit) with corresponding zoomed regions in heat colors. The same color scale was used for each image. Glcm Homogeneity1 values (range = [0.0, 1.0]) for three methods of binning (d_i d_{ii}, d_{iii}) are shown on the left together with the corresponding patient rank.

was concentric with ROI_MNI3 and ROI_MNI5 in the right hemisphere of the brain, and ROI_4.5L mirrored ROI_4.5R in the left hemisphere (figure 4). The 60 textural indices previously described were computed in each ROI pasted on the four realigned diagnostic scans modalities using binning d_i. All computations were done in subject's space resampled to isotropic 1 mm^3 voxels and after intensity standardization.

Tissue differentiation using textural indices

The ability of textural features to distinguish between tumor and WM tissue was analyzed by computing their values in ROI_4.5R and ROI_tumor using settings identical to those defined in the previous paragraph.

Statistical analysis

MR intensity standardization

For the sake of evaluation, masks of WM and GM were defined using the MNI atlas, with a probability greater than 0.9 for WM and 0.75 for GM, and excluding slices 31–51 to avoid the tumor. To compare intensity distributions before and after standardization, the Hellinger distance (Nikulin 2001) was computed between each pair of intensity histograms $H^m_{i_1,j_1}$ and $H^m_{i_2,j_2}$ with $(i_1, j_1) \neq (i_2, j_2)$. The smaller the distance, the closer are the two histograms.

Impact of intensity binning

For each textural index extracted from ROI_tumor, a ranking of the n = 30 patients was defined according to the increasing values of the index. Spearman's correlation coefficients (r) were calculated between all three pairwise combinations of rankings computed with binning d_i, d_{ii}, and d_{iii}. The higher the correlation, the more similar the ranking of patients resulted from the two methods.

To assess the ability of each texture feature to characterize tumor heterogeneity, the Area Under the receiver-operating characteristic Curve (AUC) was computed for each expert and each binning method. These AUC were defined for the 240 textural indices, the positive label (homogeneous or heterogeneous) for each index being defined as the one that gives the AUC greater than 0.5 for at least two of the binning methods.

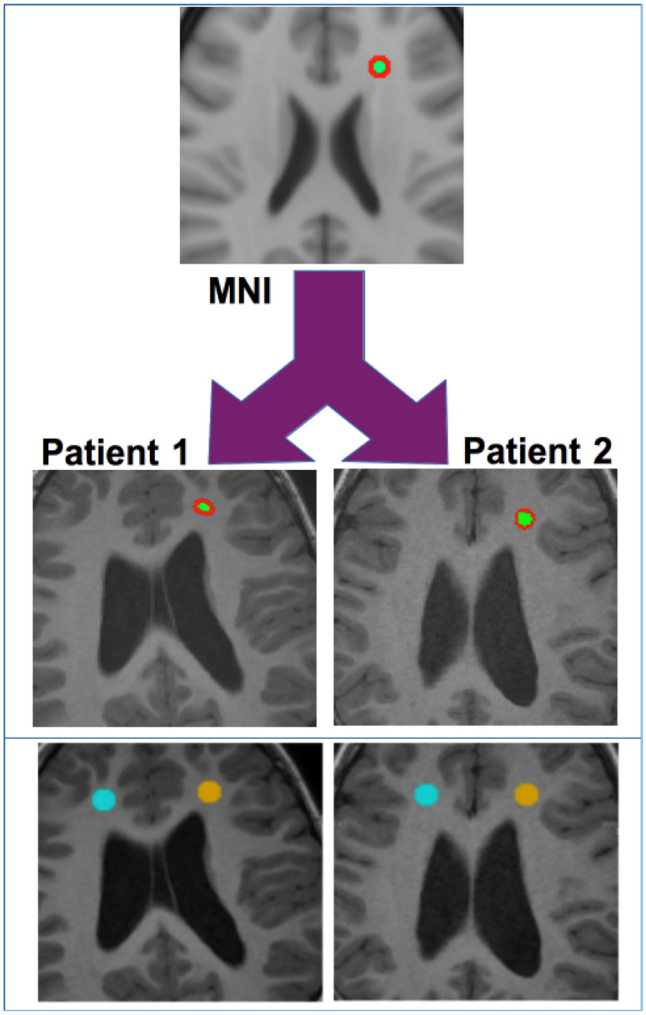

Figure 4. Concentric spheres (3 mm radius in green and 5 mm radius in red) drawn in the WM in the MNI space (top) and transferred to two patient scans (middle) using the inverse diffeomorphic transformation (ROI_MNI3 in green, ROI_MNI5 in red). Spheres (4.5 mm radius) drawn in the patient space (ROI_4.5R in golden, concentric with ROI_MNI3 and ROI_MNI5, contralateral sphere ROI_4.5L in cyan) are shown in the bottom.

Impact of volume of ROI

Spearman's correlation coefficient (r) was calculated between each index computed in the two concentric regions of different volumes: ROI_MNI3 and ROI_MNI5 and the corresponding volumes for the 30 subjects. In addition, Wilcoxon rank sum test followed by Benjamini–Hochberg correction for multiple tests was used to first compare indices computed in ROI_MNI5 and ROI_4.5L, indices computed in ROI_MNI5 and ROI_4.5R, then indices computed in ROI_4.5R and ROI_4.5L. As all three ROIs are in equivalent WM regions, textural indices are likely to be similar. If statistically significant differences are found between ROI_MNI5 and ROI_4.5L and between ROI_MNI5 and ROI_4.5R and not between ROI_4.5R and ROI_4.5L, they are probably an effect of the different volumes of ROI_MNI5.

Tissue differentiation

To determine whether the various features could distinguish between tumor region and WM, Wilcoxon rank sum test followed by Benjamini–Hochberg correction for multiple tests was used to compare indices computed in ROI_tumor and ROI_4.5L.

Results

MR intensity standardization

The WM and GM histograms before and after intensity standardization are shown in figure 5 for the four MR modalities. The mean Hellinger distance and its standard deviation between every histogram were also computed, showing a large decrease of all values after standardization (figure 5). For each modality, the mean Hellinger distance was decreased by a factor greater than 10 in WM, and by a factor greater than 2.5 in GM.

Impact of intensity binning

For 20 out of 240 indices, patient rankings obtained with binning d_i and d_{ii} were highly correlated ($|r| > 0.7$). This number increased to 188 when comparing ranking obtained with binning d_i and d_{iii} and reduced to 9 when comparing d_{ii} and d_{iii}. Figure 6 shows the $|r|$ values that were computed for each index. The two methods of binning d_i and d_{iii} provide quite similar results.

When analyzing the visual assessment of tumor heterogeneity, the binning methods d_i and d_{iii} provided a consistent ranking of patients for the three experts, which was not the case for d_{ii} (table 2). For instance, for the index T1c GLCM Homogeneity1, the mean AUC for the three experts was 0.90 for d_i, 0.43 for d_{ii} and 0.90 for d_{iii}. Figure 3 illustrates this result: Patient 1 (ranked 27 for d_i and d_{iii}; 2 for d_{ii}) shows a region classified as homogeneous by the 3 experts, while Patient 2 (ranked 1 for d_i and d_{iii}; 8 for d_{ii}) shows a region classified as heterogeneous by the 3 experts, rank 1 corresponding to the most homogeneous region.

Impact of volume of ROI

When studying textural indices measured within ROI_MNI3 and ROI_MNI5, 46/240 indices showed high correlation ($|r| > 0.7$) with the volume of the regions. Among the 60 indices, nine indices (GLCM IMC1, GLCM IMC2, GLRLM Gray Level Non Uniformity, GLRLM Run Length Non Uniformity, GLSZM High Intensity Large Area Emphasis, GLSZM Intensity Variability, GLSZM Large Area Emphasis, GLSZM Size Zone Variability, GLSZM Zone Variance) were systematically correlated with volume, whatever the MR modality. Figure 7(a) illustrates three types of indices, which show different relationships with the volumes: FLAIR GLCM Run Length Non Uniformity ($r = 0.99$), T2 GLCM Entropy ($r = 0.82$ with decreasing correlation as volume increases), and T1 Homogeneity1 ($r = 0.03$).

No statistically significant difference was found between indices computed within ROI_4.5R and ROI_4.5L (smallest p-value of 0.84), while 81 indices were significantly different (adjusted p-value < 0.05) in ROI_MNI5 and ROI_4.5R and 41 indices were significantly different (adjusted p-value < 0.05) in ROI_MNI5 and ROI_4.5L. Among these 81 (41) indices, 34 (29) were also highly correlated with volume (table 3). Figure 7(b) shows plots of the three indices illustrated in figure 7(a) within ROI_4.5R and ROI_4.5L.

Tissue differentiation

A total of 224/240 indices were significantly different (adjusted p-value < 0.05) in WM (ROI_4.5L) and in tumor (ROI_tumor) with binning d_i. Figure 7(c) shows plots of the three above selected indices within ROI_tumor and ROI_4.5L.

Discussion

Unlike in PET or CT for which the intensity range between patients can be compared relatively easily using SUV and Hounsfield units, clinical MRI data, even from mono-centric studies and using similar acquisition protocols, can yield quite different gray level range between patients and even in a given patient scanned at different times. The proposed standardization scheme is designed to deal with this problem. Most standardization techniques developed for MR images propose a matching of image histograms. In Nyul *et al* (2000), the transformation parameters to standardize histogram landmarks are estimated by a learning process, in Christensen (2003) even ordered derivatives are used to find the histogram peak corresponding to the WM, which is used to standardize all image intensities. Without the need of precise tissue segmentation and making use of the information across imaging modalities jointly, the modified hWS method proposed in this paper was successfully applied to 198 DIPG MR scans. When compared with the original method (Shinohara *et al* 2014) only slight and reproducible modifications were introduced for a robust detection of WM peak in T2 weighted images. Before intensity standardization, histograms in WM and GM varied from patient-to-patient and visit-to-visit, even when scans were acquired using similar parameters. After hWS standardization, histograms in WM were much more similar, with the Hellinger distance being drastically reduced. Although the reduction of the Hellinger distance was smaller in GM, the standardization was also effective. Furthermore, when comparing textural features before and after standardization on T1 weighted images, tumor and WM tissues were better separated for a large majority of features (results not shown). The correction that we proposed is simple, but it proved to be efficient for the data studied in this paper. This approach is thus promising for meaningful longitudinal follow-up and robust inter-patient studies.

In addition, our study shows that the choice of the intensity binning method is also crucial. Differently of the work reported by Molina *et al* (2017), this study did not directly compare textural indices values when testing different options of binning (since the values are different) but the ranking of patients that can be deduced from the different settings, similarly to Leijenaar *et al* (2015). For both d_i and d_{iii}, bin width is constant and changing the bounds of intensity levels (d_i) or keeping it constant (d_{iii}) does not modify largely the ranking of patients based on each index. The ranking derived from binning d_{ii} notably differs from the other two methods even when the

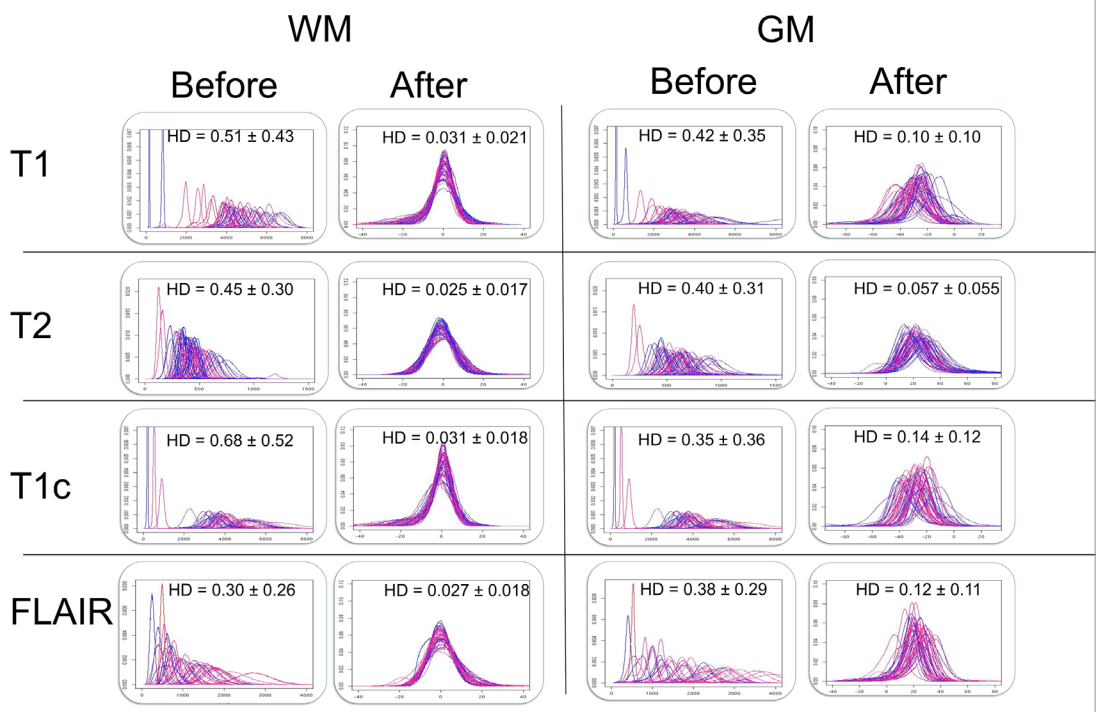

Figure 5. Histograms of intensities in WM and GM before and after standardization for T1, T2, post injection T1 (T1c) and FLAIR scans. Each blue line corresponds to a pre-biopsy scan and each pink line to a post-biopsy scan. For each type of MR scan, the mean Hellinger distance (HD) and its standard deviation between each pair of histograms are shown.

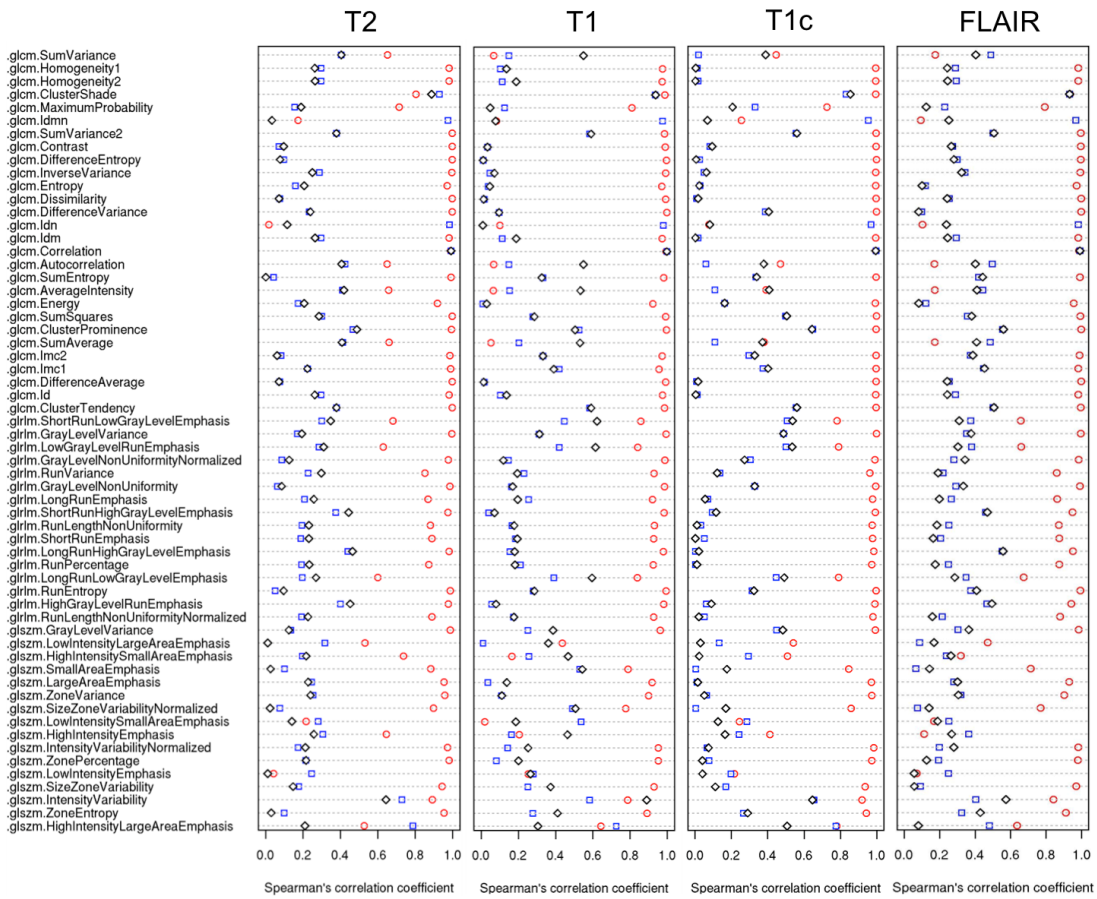

Figure 6. Spearman's correlation coefficient between rankings of patients with binning d_i and d_{ii} in blue, binning d_i and d_{iii} in red and binning d_{ii} and d_{iii} in black for each of the 240 textural indices.

Table 2. Mean of the AUCs obtained for the 240 textural indices according to the three experts and to the three binning indices.

	Binning d_i	Binning d_{ii}	Binning d_{iii}
Expert 1	0.87	0.63	0.83
Expert 2	0.83	0.61	0.79
Expert 3	0.76	0.59	0.74

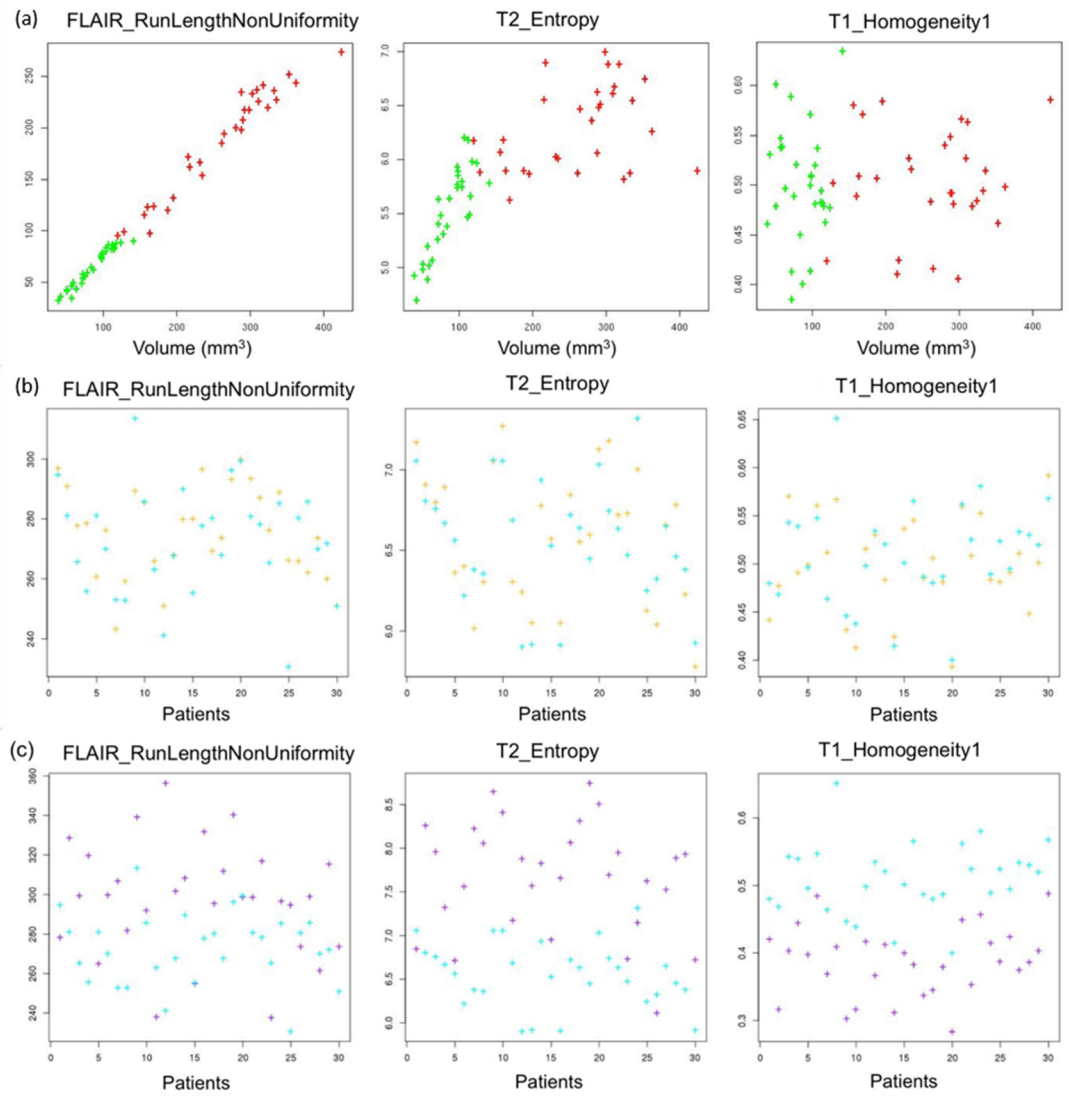

Figure 7. (a) Three different indices plotted according to region of interest (ROI) volumes (ROI_MNI3 in green, ROI_MNI5 in red): FLAIR GLRLM RunLengthNonUniformity (left) is highly correlated with volume; T2 GLCM Entropy (middle) is correlated with volume, with decreasing correlation when volume increases; T1 GLCM Homogeneity (right) is not correlated with ROI volume. (b) The same indices plotted by patients sorted by volumes of their ROI_MNI5, showing corresponding indices for ROI_4.5R (golden) and ROI_4.5L (cyan): FLAIR GLRLM RunLengthNonUniformity (left), T2 GLCM Entropy (middle), and T1 GLCM Homogeneity (right) have similar values for ROI_4.5R and ROI_4.5L. (c) The same indices plotted by patients sorted by volumes of their ROI_MNI5, showing values in ROI_4.5L (cyan) and ROI_tumor (purple): FLAIR GLRLM RunLengthNonUniformity (left), T2 GLCM Entropy (middle) and T1 GLCM Homogeneity (right) have different values for ROI_tumor and ROI_4.5L.

same number of intensity levels was used in d_{ii} and d_{iii}. Interestingly, these results are in agreement with those reported by Orlhac *et al* (2015) in PET who compared d_{ii} with d_{iii} and also with those of Leijenaar *et al* (2015) in PET who compared d_i with d_{ii}. Furthermore the ranking given by d_i and d_{iii} is consistent with the visual assessment, while the ranking provided by d_{ii} is not (figure 3), exactly as previously reported in PET (Orlhac *et al* 2015). The values for nBins and bW were set to specific values, but similar patient rankings were obtained for the large majority of the 240 textural features when using different values of nBins (8, 16, 32, 64, 128) or bW (0.75, 1, 2 3, 4). The choice of d_i binning (relative bounds and fixed bin width) after intensity standardization appears thus to be the simplest approach as it does not require the setting of some absolute lower and upper bounds. For that reason, d_i was chosen for the remaining analyses.

Table 3. Forty-one and eighty-one textural indices for which statistically significant differences were found between ROI_MNI5 and ROI_4.5L (left) and between ROI_MNI5 and ROI_4.5R (right). The 29 among these 41 and 34 among these 81 indices for which Spearman's correlation coefficient with volume (in absolute value) is greater than 0.7 are in bold.

	Texture matrix	Textural index name[a] (ROI_MNI5 versus ROI_4.5L)	Textural index name[a] (ROI_MNI5 versus ROI_4.5R)
T1	GLCM	Correlation, Idmn, Idn	Correlation, Idmn, Idn, ClusterShade, ClusterProminence, AverageIntensity, Autocorrelation, SumVariance, SumAverage
	GLRLM	**GrayLevelNonUniformity**, LongRunHighGrayLevelEmphasis, LowGrayLevelRunEmphasis, **RunLengthNonUniformity**, ShortRunLowGrayLevelEmphasis	**GrayLevelNonUniformity**, LongRunHighGrayLevelEmphasis, LowGrayLevelRunEmphasis, **RunLengthNonUniformity**, ShortRunLowGrayLevelEmphasis, HighGrayLevelRunEmphasis, LongRunLowGrayLevelEmphasis, RunEntropy, ShortRunHighGrayLevelEmphasis
	GLSZM	**HighIntensityLargeAreaEmphasis, LargeAreaEmphasis, ZoneVariance**	**HighIntensityLargeAreaEmphasis, LargeAreaEmphasis, ZoneVariance**, GrayLevelVariance, HighIntensityEmphasis, HighIntensitySmallAreaEmphasis, LowIntensityEmphasis, LowIntensitySmallAreaEmphasis, SizeZoneVariability, ZoneEntropy
T2	GLCM	**Entropy, Imc1, Imc2**	**Entropy, Imc1, Idn, Idnm, MaximumProbability**
	GLRLM	**GrayLevelNonUniformity, RunLengthNonUniformity**	**GrayLevelNonUniformity, RunLengthNonUniformity**, RunEntropy
	GLSZM	**HighIntensityLargeAreaEmphasis, IntensityVariability, LargeAreaEmphasis, SizeZoneVariability, ZoneEntropy, ZoneVariance**	**HighIntensityLargeAreaEmphasis, IntensityVariability, SizeZoneVariability, ZoneEntropy, ZoneVariance**, GrayLevelVariance, IntensityVariabilityNormalized
T1c	GLCM	Correlation, Idmn, **Idn**	Correlation, Idmn, **Idn**, ClusterShade, ClusterProminence, AverageIntensity,Autocorrelation, SumAverage, SumVaraince
	GLRLM	**GrayLevelNonUniformity**, LongRunHighGrayLevelEmphasis, LowGrayLevelRunEmphasis, **RunLengthNonUniformity**	**GrayLevelNonUniformity**, LongRunHighGrayLevelEmphasis, LowGrayLevelRunEmphasis, **RunLengthNonUniformity**, ShortRunLowGrayLevelEmphasis, ShortRunHighGrayLevelEmphasis, RunEntropy, LongRunLowGrayLevelEmphasis, HighGrayLevelRunEmphasis
	GLSZM	GrayLevelVariance, **HighIntensityLargeAreaEmphasis**, HighIntensitySmallAreaEmphasis, **LargeAreaEmphasis, SizeZoneVariability, ZoneVariance**	GrayLevelVariance, **HighIntensityLargeAreaEmphasis**, HighIntensitySmallAreaEmphasis, **LargeAreaEmphasis, SizeZoneVariability, ZoneVariance, ZoneEntropy**, LowIntensityEmphasis, IntensityVariabilityNormalized, HighIntensityEmphasis
FLAIR	GLCM	**Imc1**	**Idn, Idmn**
	GLRLM	**GrayLevelNonUniformity, RunLengthNonUniformity**	**GrayLevelNonUniformity, RunLengthNonUniformity**, RunEntropy
	GLSZM	**HighIntensityLargeAreaEmphasis, IntensityVariability, SizeZoneVariability**	**HighIntensityLargeAreaEmphasis, IntensityVariability, SizeZoneVariability, ZoneEntropy**, GrayLevelVariance

[a] Naming of indices is similar to PyRadiomics naming.

This study also shows that the volume of the ROI in which texture analysis is performed can bias many textural indices. Similar results were obtained in Sikio *et al* (2015) for 2D regions in MRI. This volume dependency may lead to erroneous conclusions in comparative studies, as it is showed in our investigations: when comparing textural indices computed from two WM regions of constant volume whatever the patient (ROI_4.5R and ROI_4.5L), no statistically significant difference between the two regions was found whatever the textural index. Meanwhile the comparison of textural indices computed from a WM region of variable volume (ROI_MNI5) with those computed on a WM region of constant volume (ROI_4.5L) showed that 41 textural indices were significantly different in the two regions, and among these 41 indices, 70% were previously identified as highly correlated with volume ($|r| > 0.7$). When comparing ROI_MNI5 and ROI_4.5R, which are concentric, even more indices were significantly different, which may be explained by the great intersection of the two ROIs, making any small difference discriminant. This study also demonstrates that the comparison of textural indices must be performed carefully when regions have different volumes. For instance, GLRLM Gray Level Non Uniformity is highly correlated with volume as found in the present study and in Orlhac *et al* (2014). This implies that radiomic models proposing this index (e.g. Michoux *et al* 2015) should be revisited to establish its real role when compared with the ROI volumes.

Finally using regions with identical volume and shape, the comparison of textural indices in the WM and in the tumor within the same scan was powerful to separate the two kinds of tissue even for textural indices that are

highly correlated with volume. Of note, first-order features should also be investigated to determine whether they could also contribute to the identification of tissue types.

Although this work was based on a rare pediatric cancer, some results can be translated to other studies. For instance, the standardization method has already been used in other brain diseases, and is exploitable independently of textural analysis. The correlation of some textural indices with the volume of the ROI was analyzed in healthy WM and thus does not depend on the pathology and this result was already reported in PET and MR studies. The present study shows that this dependence can bias further comparisons. Finally the intensity binning focused on heterogeneity comparisons and ranking of patients (not on specificities of DIPG tumors), and the choice of a constant bin width seem to be primordial to quantify heterogeneity.

This study relies on datasets from 30 patients scanned with the same scanner and our results should thus be confirmed by multi-scanner data. The modified hWS method could also be improved to better take the GM into account. However, a finer segmentation would be needed and the main advantage of the current method is that it does not depend on any sophisticated segmentation technique. Furthermore similar investigation should be performed in studies outside the brain. Our present study also focused on a series of textural indices, but others could be used, or the same could be calculated from pre-processed images (e.g. filtered images using wavelets or Laplacian of Gaussian as proposed in PyRadomics). Many of the textural indices are also correlated and depending on the intended approach (e.g. tumor segmentation, tumor classification), proper index selection will be needed. Finally the volume dependency that was highlighted in this study could be reduced when considering regions with larger volumes as suggested by some results shown in figure 7. To further investigate the robustness of textural indices, and make results transferable across centers, other aspects need to be studied, including the signal to noise ratio in the images, the spatial resolution and some differences between MR acquisition sequences.

Conclusion

In conclusion, this study suggests that structural MRI radiomic quantification through textural indices can provide valuable information in neuro-oncological studies but shows that the processing pipeline can strongly affect some intra or inter-patient comparison. Therefore, we recommend: (i) to standardize MR intensities by an appropriate method such as the modified hWS method we propose; (ii) to use a constant bin width with relative bounds to quantize the standardized images before textural index calculation; (iii) to use standard regions with the same volume to minimize its impact in the values of textural indices.

Acknowledgments

This work was partly supported by the 'Lidex-PIM' project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

ORCID iDs

Jessica Goya-Outi https://orcid.org/0000-0003-2031-5774
Frederique Frouin https://orcid.org/0000-0002-7288-1543

References

Aerts H J W L 2016 The potential of radiomic-based phenotyping in precision medicine *JAMA Oncol.* **2** 1636–42
Ashburner J and Friston K J 2005 Unified segmentation *Neuroimage* **26** 839–51
Avants B, Grossman M and Gee J C 2007 Symmetric diffeomorphic image registration: evaluating automated labeling of elderly and neurodegenerative cortex *Med. Image Anal.* **12** 26–41
Castel D *et al* 2015 Histone H3F3A and HIST1H3B K27M mutations define two subgroups of diffuse intrinsic pontine gliomas with different prognosis and phenotypes *Acta Neuropathol.* **130** 815–27
Christensen J D 2003 Normalization of brain magnetic resonance images using histogram even-order derivative analysis *Magn. Reson. Imaging* **21** 817–20
Cohen K J, Jabado N and Grill J 2017 Diffuse intrinsic pontine gliomas—current management and new biologic insights. Is there a glimmer of hope? *Neuro-Oncology* **19** 1025–34
Gevaert O, Mitchell L A, Achrol A S, Xu J, Echegaray S, Steinberg G K, Cheshier S H, Napel S, Zaharchuk G and Plevritis S K 2014 Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features *Radiology* **273** 168–74
Gillies R J, Kinahan P E and Hricak H 2016 Radiomics: images are more than pictures, they are data *Radiology* **278** 563–77
Haralick R, Shanmugan K and Dinstein I 1973 Textural features for image classification *IEEE Trans. Syst. Man Cybern.* **3** 610–21
Hatt M *et al* 2015 ^{18}F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort *J. Nucl. Med.* **56** 38–44
Jenkinson M, Bannister P, Brady M and Smith S 2002 Improved optimization for the robust and accurate linear registration and motion correction of brain images *Neuroimage* **17** 825–41

Kickingereder P *et al* 2016a Radiogenomics of glioblastoma: machine learning—based classification of molecular characteristics by using multiparametric and mutiregional MR imaging features *Radiology* **281** 907–18

Kickingereder P *et al* 2016b Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models *Radiology* **280** 880–9

Lambin P *et al* 2012 Radiomics: extracting more information from medical images using advanced feature analysis *Eur. J. Cancer* **48** 441–6

Larroza A, Bodí V and Moratal D 2016 Texture analysis in magnetic resonance imaging: review and considerations for future applications *Assessment of Cellular and Organ Function and Dysfunction using Direct and Derived MRI Methodologies* ed C Constantinides (London: IntechOpen) pp 75–106

Leijenaar R T H, Nalbantov G, Carvalho S, van Elmpt W J C, Troost E G C, Boellaard R, Aerts H J W L, Gillies R J and Lambin P 2015 The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis *Sci. Rep.* **5** 11075

Madabhushi A and Udupa J K 2006 New methods of MR image intensity standardization via generalized scale *Med. Phys.* **33** 3426–34

Mayerhoefer M E, Szomolanyi P, Jirak D, Materka A and Trattnig S 2009 Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study *Med. Phys.* **36** 1236–43

Michoux N, Van den Broeck S, Lacoste L, Fellah L, Galant C, Berlière M and Leconte I 2015 Texture analysis on MR images helps predicting non-response to NAC in breast cancer *BMC Cancer* **15** 574

Molina D, Pérez-Beteta J, Martínez-González A, Martino J, Velasquez C, Arana E and Pérez-García V M 2017 Lack of robustness of textural measures obtained from 3D brain tumor MRIs impose a need for standardization *PLoS One* **12** e0178843

Molina D, Pérez-Beteta J, Luque B, Arregui E, Calvo M and Borrás J M 2016 Tumour heterogeneity in glioblastoma assessed by MRI texture analysis: a potential marker of survival *Br. J. Radiol.* **89** 1064

Nikulin M S 2001 Hellinger distance *Encyclopedia of Mathematics* (Dordrecht: Springer)

Nyul L G, Udupa J K and Zhang X 2000 New variants of a method of MRI scale standardization *IEEE Trans. Med. Imaging* **19** 143–50

Orlhac F, Soussan M, Chouahnia K, Martinod E and Buvat I 2015 ^{18}F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer *PLoS One* **10** e0145063

Orlhac F, Soussan M, Maisonobe J-A, Garcia C A, Vanderlinden B and Buvat I 2014 Tumor texture analysis in ^{18}F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis *J. Nucl. Med.* **55** 414–22

Shinohara R T, Sweeney E M, Goldsmith J, Shiee N, Mateen F J, Calabresi P A, Jarso S, Pham D L, Reich D S and Crainiceanu C M 2014 Statistical normalization techniques for magnetic resonance imaging *Neuroimage Clin.* **6** 9–19

Sikio M, Holli-Helenius K K, Ryymin P, Dastidar P, Eskola H and Harrison L 2015 The effect of region of interest size on textural parameters *9th Int. Symp. on Image and Signal Processing and Analysis, ISPA 2015 (Zagreb)* pp 149–53

Skogen K, Schulz A, Dormagen J B, Ganeshan B, Helseth E and Server A 2016 Diagnostic performance of texture analysis on MRI in grading cerebral gliomas *Eur. J. Radiol.* **85** 824–9

Tang X 1998 Texture information in run-length matrices *IEEE Trans. Image Process.* **7** 1602–9

Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, Sequeira J and Mari J 2009 Texture indexes and gray level size zone matrix application to cell nuclei classification *Pattern Recognition and Information Processing: Proc. of the Tenth Int. Conf. PRIP'2009 (Minsk)* 140–5

Tustison N J, Avants B B, Cook P A, Zheng Y, Egan A, Yushkevich P A and Gee J C 2010 N4ITK: improved N3 bias correction *IEEE Trans. Med. Imaging* **29** 1310–20

Van Griethuysen J J M, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan R G H, Fillion-Robin J C, Pieper S and Aerts H J W L 2017 Computational radiomics system to decode the radiographic phenotype *Cancer Res.* **77** e104–7

Yan J, Chu-Shern J L, Loi H Y, Khor L K, Sinha A K, Quek S T, Tham I W K and Townsend D 2015 Impact of image reconstruction settings on texture features in ^{18}F-FDG PET *J. Nucl. Med.* **56** 1667–73

Yang D, Rao G, Martinez J, Veeraraghavan A and Rao A 2015 Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma *Med. Phys.* **42** 6725–35