MAGNETIC RESONANCE

How can we combat multicenter variability in MR radiomics? Validation of a correction procedure

Fanny Orlhac[1,2] · Augustin Lecler[3] · Julien Savatovski[3,4] · Jessica Goya-Outi[1] · Christophe Nioche[1] ·
Frédérique Charbonneau[3] · Nicholas Ayache[2] · Frédérique Frouin[1] · Loïc Duron[3] · Irène Buvat[1]

Abstract

Objective Test a practical realignment approach to compensate the technical variability of MR radiomic features.

Methods T1 phantom images acquired on 2 scanners, FLAIR and contrast-enhanced T1-weighted (CE-T1w) images of 18 brain tumor patients scanned on both 1.5-T and 3-T scanners, and 36 T2-weighted (T2w) images of prostate cancer patients scanned in one of two centers were investigated. The ComBat procedure was used for harmonizing radiomic features. Differences in statistical distributions in feature values between 1.5- and 3-T images were tested before and after harmonization. The prostate studies were used to determine the impact of harmonization to distinguish between Gleason grades (GGs).

Results In the phantom data, 40 out of 42 radiomic feature values were significantly different between the 2 scanners before harmonization and none after. In white matter regions, the statistical distributions of features were significantly different ($p < 0.05$) between the 1.5- and 3-T images for 37 out of 42 features in both FLAIR and CE-T1w images. After harmonization, no statistically significant differences were observed. In brain tumors, 41 (FLAIR) or 36 (CE-T1w) out of 42 features were significantly different between the 1.5- and 3-T images without harmonization, against 1 (FLAIR) or none (CE-T1w) with harmonization. In prostate studies, 636 radiomic features were significantly different between GGs after harmonization against 461 before. The ability to distinguish between GGs using radiomic features was increased after harmonization.

Conclusion ComBat harmonization efficiently removes inter-center technical inconsistencies in radiomic feature values and increases the sensitivity of studies using data from several scanners.

Key Points

• *Radiomic feature values obtained using different MR scanners or imaging protocols can be harmonized by combining off-the-shelf image standardization and feature realignment procedures.*

• *Harmonized radiomic features enable one to pool data from different scanners and centers without a substantial loss of statistical power caused by intra- and inter-center variability.*

• *The proposed realignment method is applicable to radiomic features from different MR sequences and tumor types and does not rely on any phantom acquisition.*

Keywords Magnetic resonance imaging · Neoplasms · Diagnostic imaging · Image processing · Computer-assisted methods

---

✉ Fanny Orlhac
orlhacf@gmail.com

[1] Laboratoire d'Imagerie Translationnelle en Oncologie (LITO), Inserm, Institut Curie Centre de Recherche, Université Paris-Saclay, Bât 101B, rue Henri Becquerel, 91401 Orsay, France

[2] Université Côte d'Azur, Inria Sophia Antipolis - Méditerranée, Epione Project Team, 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France

[3] Department of Neuroradiology, Fondation Ophtalmologique A. Rothschild, 29 rue Manin, 75019 Paris, France

[4] Centre Imagerie Médicale Paris 13, 17 avenue d'Italie, 75013 Paris, France

Abbreviations

CE-T1w	Contrast-enhanced T1-weighted
CT	Computed tomography
D1/D2	Prostate cancer patient database 1/2
hWS	Hybrid white stripe
LDA	Linear discriminant analysis
MRI	Magnetic resonance imaging
PET	Positron emission tomography
ROI	Region of interest
T2w	T2-weighted
VOI	Volume of interest
WM	White matter

Introduction

Radiomics faces the critical issue of a lack of reproducibility that still hampers the successful translation of radiomic model discovery into better diagnosis, patient classification, or monitoring radiomics-based tools. Indeed, radiomic feature values are significantly affected by the technical settings of the imaging devices and protocols, as demonstrated in positron emission tomography (PET), computed tomography (CT), and magnetic resonance imaging (MRI) [1–3]. To tackle the variability of radiomic features induced by different technical settings, radiomic models can be designed using a wide variety of images encompassing most technical settings, or images or radiomic feature values have to be harmonized before designing models. Ignoring the center effect, as is often observed in many papers, results in the lack of generalization of the radiomic models [4].

In prospective studies, imaging protocols could be harmonized upstream between centers to minimize the impact of imaging protocols on feature values [5, 6], although harmonizing between machines of different generations often comes with degrading the image quality achieved by the most recent scanners [5]. In retrospective studies, this approach is not an option. Several groups have proposed to reduce the variability by resampling the images to a common voxel size or by filtering the images to match spatial resolution [7, 8]. However, this requires accessing the images retrospectively and the filtering procedure reduces the quality of images acquired using the most recent devices. Others apply a z-score transformation [9] to each feature value based on mean and standard deviation measured in each center for that feature, but this assumes that images produced by the different centers have been obtained in similar patient samples (e.g., same proportion of advanced and early-stage tumors), which is sometimes difficult to achieve.

In genomics, researchers face a similar problem called batch effect and caused by the handling of samples by different laboratories, different technicians, or different days that can obscure individual variations. To deal with that problem in genomics, Johnson et al [10] introduced the ComBat

realignment method. The method realigns all data in a single space in which the batch effect is discarded without altering the biological information. This approach has already been successfully validated for radiomic features measured from PET [11] and CT [12, 13] images of patient or phantom data in studies supporting the relevance of harmonization.

In MR, the challenge is even more difficult as, unlike in PET and CT where images are expressed in kBq/mL and Hounsfield units, respectively, there is no standard MR intensity grayscale, implying the lack of a tissue-specific absolute-intensity numeric meaning, even within the same MR imaging protocol, for the same body region, for images obtained on the same scanner, and for the same patient. The standardization of image intensities among patients is therefore absolutely needed for comparing values of intensity-based features. In brain MR, standardization approaches have been proposed to correct for the intensity variability [3, 14–16]. In particular, the hybrid white stripe (hWS) method proved to be successful in the context of neurodegenerative diseases and cancer [17, 18]. ComBat has been validated in MRI for the harmonization of cortical thickness measurements across scanners [19]. Although it has been used in MR radiomic studies [20–23], it has never been validated in that highly challenging context.

Here, we extend the ComBat approach to provide a harmonization procedure applicable to any radiomic feature. We demonstrate that by combining the image standardization (such as hWS) with ComBat realignment, MR radiomic features can be pooled without being adversely impacted by multiple sources of variability, ensuring higher sensitivity and specificity of multicenter MR radiomic studies.

Materials and methods

The study was carried out in accordance with the World Medical Association's Declaration of Helsinki. For experiment 2 involving MRI brain studies, the institutional review board of the Fondation Ophtalmologique A. Rothschild approved the study (IRB No. 1512-016-726), and the requirement to obtain written informed consent was waived because of the retrospective nature of the study. For experiment 3 using MRI prostate studies, all patient data are publicly available in https://datadryad.org/resource/10.5061/dryad.b3d257g and were initially studied in [24]. All patient data were anonymized. All authors had control of the data and information submitted for publication.

Experiment 1: phantom studies

The phantom data used in this study have been extracted from the RIDER Phantom MRI study [25] and are publicly available in the TCIA platform [26]. The phantom consists of 19 doped gel-filled tubes containing a gadolinium-based contrast

agent. We used the T1 acquisitions obtained with a 1.5-T scanner and a 3-T scanner (scanners B and D, respectively, in [25], details in Supplemental data 1). For each image, 19 spherical volumes of interest (VOIs) of 3.5 mL centered on each tube were drawn. We computed 42 radiomic features (Supplemental data 2) using LIFEx freeware [27] (www.lifexsoft.org), including an open-source radiomic protocol compliant with the Image Biomarker Standardisation Initiative guidelines [28]. Radiomic features were calculated using a fixed bin size [3, 29] set to the average standard deviation of the signal intensity, between the minimum and the maximum intensity measured in all VOIs. This discretization step is required to set voxels with similar intensity to the same value, hence to reduce the impact of noise.

Experiment 2: MRI brain studies

For experiment 2, we retrospectively selected 18 patients (13 men; mean age, 50 ± 18 years; age range, 26–85 years; Table 1) with grade III and IV glial tumors from January 2017 to May 2018 from an institutional database. All patients underwent two MRI scans using the same protocol: one on a 1.5-T scanner (Philips Achieva, Philips Medical Systems) and the other one on a 3-T scanner (Philips Ingenia). The median delay between the two scans was 30 days (range 4–93 days) without chemotherapy, surgery, radiotherapy, and any visual evolution of the tumor and tumor heterogeneity between the scans. Two MR sequences (details in Supplemental data 1) were acquired: a 3D FLAIR scan (17 patients) and a 3D contrast-enhanced T1-weighted (CE-T1w) scan (14 patients).

For each patient and each sequence, the 3-T images were co-registered to the 1.5-T images using rigid transformations in FSL-FLIRT [30]. Field inhomogeneity was corrected using the N4 algorithm [31] owing to the publicly available ANTs software (http://stnava.github.io/ANTs) with the standard setting of hyper-parameters.

For each sequence, the tumor lesions were manually segmented based on a consensus of two radiologists (A.L. and L.D. with 9 years and 2 years of experience, respectively) on the 1.5-T images and the resulting regions were copied on the 3-T images. Three slices (top, middle, bottom) were selected in each tumor to obtain three 2D regions of interest (ROIs) per tumor, yielding a total of 54 tumor ROIs for FLAIR images (= 3 × 18 tumors; one patient had two distinct lesions) and 51 tumor ROIs for CE-T1w images (= 3 × 17 tumors; one patient had two distinct lesions and another had three). In addition, in each patient, 6 regions of 0.5 mL each were drawn in the white matter (WM), yielding 102 WM-VOIs for FLAIR images and 84 WM-VOIs for CE-T1w images that were copied onto the 3-T images.

Each patient's image volume was standardized irrespective of the other patients using the hWS method [17] as previously described [3]. The hWS method applies a z-score transformation to the brain voxel values based on the normally appearing distribution of WM intensities.

For each ROI and VOI based on native and hWS-standardized images resampled at $1 \times 1 \times 1 \text{ mm}^3$, we computed 42 radiomic features using LIFEx. Radiomic features were calculated using a fixed bin size [3, 29] set to the average standard deviation of the WM signal intensity, between the minimum and the maximum intensity measured in all WM and tumor VOIs for each sequence separately (details in Supplemental data 1).

Experiment 3: MRI prostate studies

Two prostate cancer patient databases (D1 and D2; Table 1) with publicly available radiomic features were used [24].

Table 1 Patient characteristics

Parameter	MRI brain studies	MRI prostate studies
Sex		
Men	13	D1 = 23; D2 = 13
Women	5	0
Mean age (years)	50 ± 18	Not reported
No. of tumor regions		
FLAIR images	18 3D regions (54 2D regions)	–
CE-T1w images	17 3D regions (51 2D regions)	–
T2w images	–	D1 = 65; D2 = 40
Gleason grade		
Low risk	–	D1 = 21(32%); D2 = 26 (65%)
Intermediate/high risk	–	D1 = 44 (68%); D2 = 14 (35%)

The mean age of patients is not reported in the study of Penzias et al [24]

CE-T1w contrast-enhanced T1-weighted, *D1* prostate dataset 1, *D2* prostate dataset 2, *T2w* T2-weighted

These two databases have been initially entirely studied by an independent team to investigate the relationship between features computed from MR images and from digitized tissue images in order to discriminate between prostate cancer grades, without taking into account that MRI scans were acquired in two different centers [24]. Here, we precisely investigate how accounting for the center effect actually changes the ability of each MRI feature to distinguish between tumor grades. The experimental protocols were approved under the IRB protocol #02-13-42C by the University Hospitals of Cleveland Institutional Review Board. Patients underwent T2-weighted (T2w) MRI before a radical prostatectomy. In D1, 23 patients from University of Pennsylvania were scanned between 2009 and 2011 (3 T Verios, Siemens Healthcare; echo time 107–127 ms; repetition time 3690–7090 ms). In D2, 13 patients from St. Vincent's Hospital were scanned between 2012 and 2014 (11 patients [3 T, Philips Medical Systems; echo time 67–100 ms; repetition time 2525–3567 ms] and 2 patients [1.5 T, Siemens Healthcare; echo time 119 ms; repetition time 3760 ms]). After surgery, the resected prostate gland was analyzed by pathologists to determine the Gleason scores, categorizing in low (score of 3 + 3) or intermediate/high (score of 3 + 4, 4 + 3, 4 + 4, or higher) risk. D1 consisted of 21 low-risk regions and 44 intermediate/high-risk regions, while D2 included 26 low-risk regions and 14 intermediate/high-risk regions (Table 1). Based on a co-registration with histology images, the corresponding tumor regions were manually segmented by a radiologist. MR images were standardized to a template distribution based on the per-patient median of intra-prostatic pixel intensities of D1 [32]. For each region, 2379 radiomic features were computed using a homemade software (details of feature calculation described in [24]) and we selected the 2326 features available for all patients for our analysis.

Realignment method

To correct for the scanner effect, the ComBat realignment method was used [10]. In the context of radiomics, ComBat has already been validated for PET [11] and CT features [12, 13]. The method directly applies to the radiomic feature values and estimates the scanner effect by matching the statistical distributions of the feature values measured in VOI j for each scanner i

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i \varepsilon_{ij}$$

where α is the average value for feature y_{ij}, X is the design matrix for the covariates of interest, β is the vector of regression coefficients corresponding to each covariate, γ_i is an additive scanner effect, and δ_i is a multiplicative protocol effect affected by an error term (ε_{ij}). The model parameters α, β, γ_i, and δ_i are estimated using a maximum likelihood

approach based on the set of available observations from the two scanners in experiments 1 and 2 and based on the two patient databases for experiment 3. The corrected values are obtained using

$$y_{ij}^{ComBat} = \frac{y_{ij} - \widehat{\alpha} - X_{ij}\widehat{\beta} - \widehat{\gamma}_i}{\widehat{\delta}_i} + \widehat{\alpha}$$

where $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\gamma}_i$, and $\widehat{\delta}_i$ are estimators of α, β, γ_i, and δ_i, respectively.

The non-parametric form of the model was used, with no assumption regarding the statistical laws followed by the features and a transformation determined for each feature separately. For experiments 1 and 2, no biologic covariate was used (i.e., $X = 0$) since the data came from the same patients or phantom scanned on 1.5-T and 3-T machines, and we realigned feature values computed from WM and tumor regions in patient data separately. For experiment 3, we introduced the Gleason grade as a binary covariate since the proportion of low versus intermediate/high-risk regions was very different between the 2 databases (32% low-risk VOI in D1 and 68% in D2; Table 1).

To facilitate the access to the ComBat method for medical imaging professionals, we provide a free online application (available at https://forlhac.shinyapps.io/Shiny_ComBat/), named ComBaTool, with example input files (Supplemental data 3 and 4) and a step-by-step tutorial (Supplemental data 5). This application embeds a free function called ComBat [19] (https://github.com/Jfortin1/ComBatHarmonization) based on the R software, but running the application does neither require R or any third-party software to be installed nor require having any programming skills.

Statistical analysis

Statistical analysis was performed with the R software (version 3.6.1).

In experiment 1, we performed univariate two-sided Friedman tests before and after ComBat realignment between the two phantom scans. In experiment 2, we used two-sided Friedman tests for each radiomic feature to test whether the values derived from the 1.5-T and 3-T scans were significantly different both in the WM and in the tumor regions in three configurations: (C1) native images without ComBat realignment, (C2) hWS-standardized images without realignment, and (C3) hWS-standardized images with realignment. The Benjamini-Hochberg procedure was used to control the false discovery rate [33]. p values less than 0.05 were interpreted as statistically significant. Bland-Altman graphs were plotted to demonstrate the differences in feature values calculated from the 1.5-T and 3-T scans.

In experiment 3, we performed Wilcoxon tests with the Benjamini-Hochberg procedure for all radiomic features to

distinguish between low-risk and intermediate/high-risk groups when pooling patients from D1 and D2, without ComBat realignment, with realignment, and with realignment including the Gleason grade as a covariate of interest. To show that ComBat does not create false positive results, we repeated these tests after randomly assigning a label to each VOI to get 53 sham low-risk VOIs and 52 sham intermediate/high-risk VOIs. To identify the risk group, we built a multivariate signature by means of a linear discriminant analysis (LDA) using D1 dataset as a training set and including only the features with a p value of univariate Wilcoxon test less than 5%. We tested the classification performance on D2 data by calculating the Youden Index (= sensitivity + specificity − 1). We repeated this procedure in three configurations: without ComBat realignment, with realignment, and with realignment including the Gleason grade as a covariate of interest.

Results

Patient characteristics are shown in Table 1.

Experiment 1

In the phantom data, 40 out of 42 p values of the Friedman test were lower than 5% without realignment. Only two p values (coarseness and gray-level non-uniformity) were greater than 0.05 between the two acquisitions. After ComBat, all p values of Friedman tests were greater than 0.05, showing that the protocol effect was no longer detectable.

Experiment 2

A total of 37 out of 42 radiomic features (88%) computed from WM-VOI and 41 out of 42 (98%) from tumor lesions yielded Friedman tests' p values less than 0.05 between 1.5-T and 3-T native FLAIR brain images without hWS standardization nor ComBat realignment (Table 2; Supplemental data 6). Using the hWS standardization of MR images, 29/42 (69%) of p values for WM regions and 25/42 (60%) of p values for tumor lesions were less than 0.05. Combining the hWS standardization with the ComBat feature distribution realignment, only one p value (long-zone emphasis) was less than 0.05 for tumor lesions ($p = 0.017$), demonstrating that the scanner effect was no longer detectable for the vast majority of radiomic features. Figure 1 shows the evolution of the distribution of the correlation radiomic feature calculated from the gray-level co-occurrence matrix. On native FLAIR images, the plot shows a shift in distribution with greater values for WM-VOI and tumor lesions for 3-T scans compared to 1.5-T scans. After hWS standardization and realignment, the distributions between the two scanners better overlap. To clarify the respective role of hWS and ComBat, Fig. 2 shows the Bland-Altman plots of the mean value measured in WM-VOI for FLAIR images based on 3-T scans and 1.5-T scans. The hWS standardization within each patient rescaled the values to make them similar between the two scans. The realignment reduced the systematic difference between the two.

The same trends were observed for CE-T1w images (Table 2; Supplemental data 7).

Experiment 3

On T2w prostate images after standardization performed by [24], 461 out of 2326 radiomic features had p values of Wilcoxon tests less than 0.05 for distinguishing between low and intermediate/high risks when pooling the two patient cohorts (D1 + D2). After ComBat without any co-variate, 460 out of 2326 p values were less than 0.05. Using the Gleason grade co-variate in ComBat, 636 out of 2326 p values were less than 0.05. Figure 3 demonstrates a better alignment of radiomic feature values extracted from low-risk VOI and intermediate/high-risk VOI separately between the two patient groups after using ComBat with a co-variate accounting for the recruitment specificity of each center.

Table 2 Significant Friedman tests ($p < 0.05$) without and with hybrid white stripe (hWS) standardization and/or ComBat realignment for brain scans

	w/o hWS standardization	With hWS standardization	
	w/o ComBat	w/o ComBat	With ComBat
FLAIR images			
WM	37/42 (88)	29/42 (69)	0/42 (0)
Tumor lesions	41/42 (98)	25/42 (60)	1/42 (2)
CE-T1w images			
WM	37/42 (88)	27/42 (64)	0/42 (0)
Tumor lesions	36/42 (86)	2/42 (5)	0/42 (0)

Data are numerator/denominator; data in parentheses are percentage

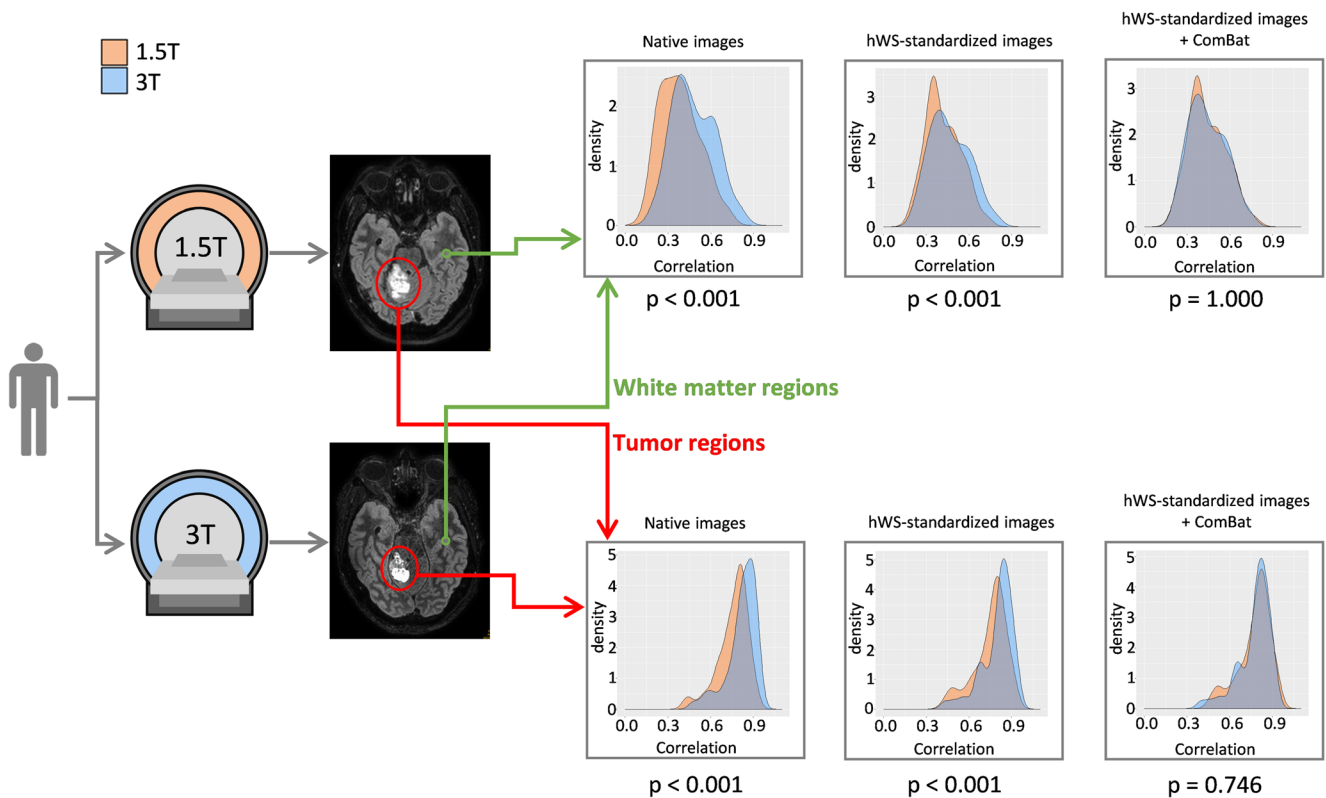WM white matter, CE-T1w contrast-enhanced T1-weighted, w/o without

Fig. 1 Experiment 2. 18 patients with brain lesions were scanned on both 1.5-T and 3-T scanners. Based on native or for hybrid white stripe (hWS)-standardized images, 42 radiomic features were computed in a tumor region and in a white matter region. As an example, the probability density function (%) of the correlation radiomic feature calculated from the gray-level co-occurrence matrix on FLAIR images is plotted here without and with ComBat realignment (ComBaTool was applied separately on the two tissue types: white matter and tumor) for 1.5-T MRI (in orange) and 3-T MRI (in blue). *p* values are for Friedman tests of each tissue between the two MRI devices

When a risk (low or moderate/high) was randomly assigned to each VOI, no *p* value was less than 0.05 before and after ComBat without and with a co-variate representing the Gleason grade.

The multivariate radiomic model identified using LDA on the D1 data to distinguish low versus intermediate/high risk was applied to D2 patients, yielding a Youden Index of 0.12 (sensitivity = 19%, specificity = 93%) before ComBat. After ComBat, the Youden Index increased to 0.20 (sensitivity = 27%, specificity = 93%) and to 0.43 (sensitivity = 58%, specificity = 86%) using the Gleason grade as co-variate in ComBat.

Discussion

The scanner effect affects the radiomic feature values extracted from MR images, introducing major confounding factors in multicentric or multiprotocol studies. Here, we validated a harmonization procedure combining ComBat realignment with MR image standardization to co-analyze MR radiomic features extracted from different scanners. Using phantom data and brain scans acquired for the same patients (without any tumor evolution detected visually between the two scans) with 1.5-T and 3-T scanners, we showed that this harmonization
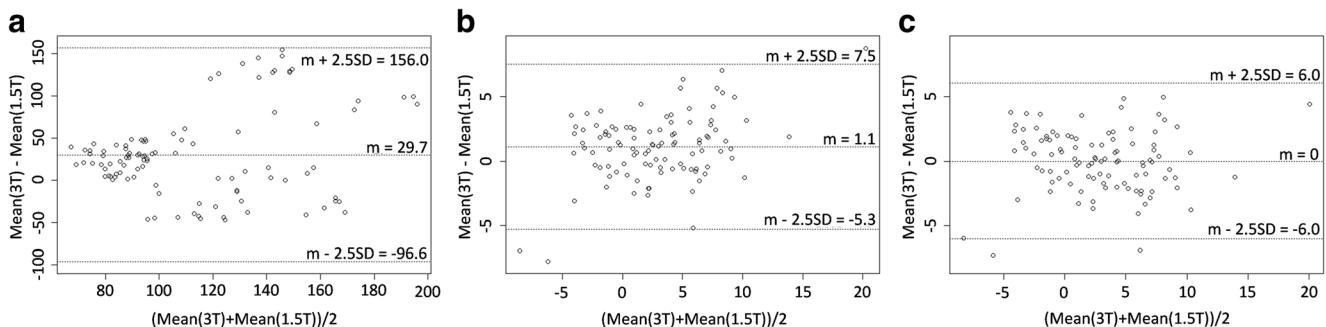

Fig. 2 Experiment 2. Bland-Altman plots of the mean value computed in white matter regions based on 1.5-T and 3-T scans for FLAIR native images (**a**), for hybrid white stripe (hWS)-standardized images (**b**), and for hWS-standardized images with ComBat realignment (**c**)
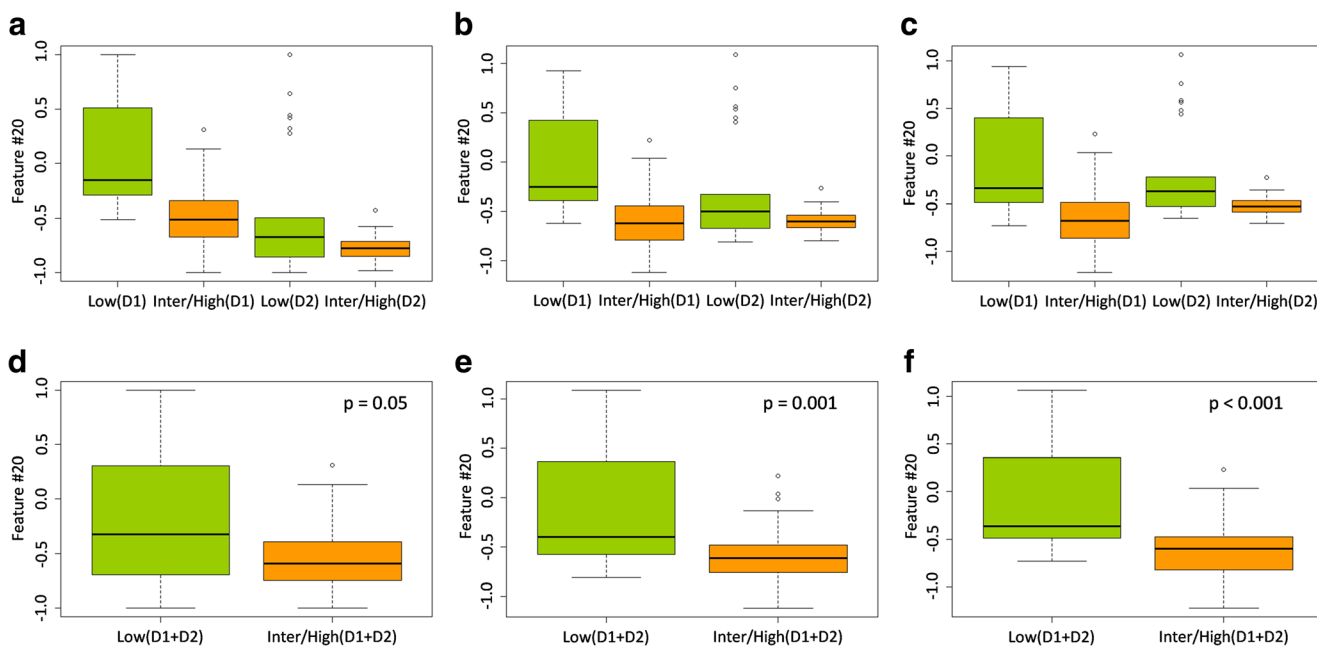
Fig. 3 Experiment 3. Boxplots of feature #20 (called Gabor:cos:theta=0:lambda=2:Standard Deviation in [24]) for low-risk VOI and intermediate/high-risk VOI, before ComBat realignment (**a**, **d**), after ComBat realignment without covariate (**b**, **e**), and after ComBat realignment with covariate (**c**, **f**) for the prostate patient cohorts D1 and D2 separately (**a**–**c**) or together (**d**–**f**). *p* values are from Wilcoxon tests

procedure realigns radiomic feature distributions and removes the scanner effect for T1, FLAIR, and CE-T1w images. The goal was not to test our ability to reproduce feature values measured in 3-T MR images from 1.5-T images, since we expect different signals from the two devices with more details in the 3-T images (cf Fig. 1). Yet, in the context of radiomics, pooling images acquired using different devices and different acquisition and reconstruction protocols is often needed to increase the size of cohort. In that context, we demonstrated that ComBat could realign feature values so that all data could be analyzed together, even if images had been acquired with different magnetic fields. It is important to underline that a different ComBat transformation is estimated for each sequence and each tissue type independently because imaging protocols do not have the same effect on each tissue. Using the prostate scans acquired in different patients from two centers, we confirmed the effectiveness of the harmonization for T2w images and demonstrated that harmonization did not alter the discriminant information conveyed by the features. This experiment also shows that pooling data corrected for the scanner effect could increase the statistical power, identify more radiomic features able to distinguish between the low-risk and intermediate/high-risk regions in prostate lesions, and yield a more discriminant multivariate model. Importantly, we showed that when no difference between groups was expected, here between the sham low-risk and intermediate/high-risk VOIs, ComBat did not introduce any false positive differences.

The ComBat realignment method is fast and easy to use and operates directly on radiomic feature values (no training set needed, no phantom acquisition, no need to access images). It is applicable to radiomic features extracted from different MR sequences after a first step of image standardization, as previously described [3]. We also demonstrated the added value of the covariate in the realignment process when patient characteristics are different between centers (here Gleason grade) for univariate and multivariate analyses. To deal with the center effect, other authors reported the potential of generative adversarial networks to transform images from one imaging protocol (or a domain) to another [34]. Although promising results have been reported in the literature [35, 36], these techniques require access to the images, unlike ComBat. The ComBat realignment method has been previously used in MR radiomic studies [20–23] without any explicit validation or investigation of the respective role of the image standardization and of the scanner/protocol effect compensation as studied here (Figs. 1 and 2). In [20], authors reported an increased accuracy of entropy extracted from apparent diffusion coefficient MR images to predict the locoregional control in cervical cancer after ComBat, fully consistent with our findings.

Our study has some limitations. We could only include 18 patients in experiment 2 because it is very uncommon for patients to undergo MR on both 1.5-T and 3-T scanners within a time lapse during which the tumor has not visually evolved. Still, this small sample allowed us to confirm results obtained using the phantom data. In addition, such a small number allowed us to demonstrate that ComBat performed well even with a limited number of cases, confirming results published

in genomic applications [10]. Another limitation is that our findings should still be validated for other cancer types, MR sequences, and devices.

In conclusion, we demonstrated that the ComBat realignment method in combination with intra-patient image standardization could efficiently remove the scanner/protocol effect while preserving the individual variations in phantom, brain, and prostate MR scans. This approach enables large MR multicentric studies to investigate the added value of radiomic analysis in patient management. To facilitate large multicenter/multiprotocol radiomic studies, we provide the ComBat method as an online ComBaTool application.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Fanny Orlhac.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise (Fanny Orlhac).

Informed consent Written informed consent was waived by the institutional review board.

Ethical approval Institutional review board approval was obtained.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in the study of Penzias et al [24].

Methodology
• retrospective
• experimental
• multicenter study

References

1. Yan J, Chu-Shern JL, Loi HY et al (2015) Impact of image reconstruction settings on texture features in 18F-FDG PET. J Nucl Med 56:1667–1673
2. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J et al (2018) Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 288:407–415
3. Goya-Outi J, Orlhac F, Calmon R et al (2018) Computation of reliable textural indices from multimodal brain MRI: suggestions based on a study of patients with diffuse intrinsic pontine glioma. Phys Med Biol 63:105003
4. Reuzé S, Orlhac F, Chargari C et al (2017) Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. Oncotarget 8:43169–43179
5. Boellaard R, Delgado-Bolton R, Oyen WJG et al (2015) FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging 42:328–354
6. Clarke LP, Nordstrom RJ, Zhang H et al (2014) The quantitative imaging network: NCI's historical perspective and planned goals. Transl Oncol 7:1–4
7. Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E (2018) Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep 8:10545
8. Mackin D, Fave X, Zhang L et al (2017) Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One 12:e0178524
9. Chatterjee A, Vallières M, Dohan A et al (2019) Creating robust predictive radiomic models for data from independent institutions using normalization. IEEE TRPMS 3:210–215
10. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8:118–127
11. Orlhac F, Boughdad S, Philippe C et al (2018) A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med 59:1321–1328
12. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I (2019) Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology 291:53–59
13. Mahon RN, Ghita M, Hugo GD, Weiss E (2020) ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. Phys Med Biol 65:015010
14. Zhuge Y, Udupa JK (2009) Intensity standardization simplifies brain MR image segmentation. Comput Vis Image Underst 113:1095–1103
15. Ge Y, Udupa JK, Nyúl LG, Wei L, Grossman RI (2000) Numerical tissue characterization in MS via standardization of the MR image intensity scale. J Magn Reson Imaging 12:715–721
16. Nyúl LG, Udupa JK (1999) On standardizing the MR image intensity scale. Magn Reson Med 42:1072–1081
17. Shinohara RT, Sweeney EM, Goldsmith J et al (2014) Statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin 6:9–19
18. Kickingereder P, Bonekamp D, Nowosielski M et al (2016) Radiogenomics of glioblastoma: machine learning-based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. Radiology 281:907–918
19. Fortin J-P, Cullen N, Sheline YI et al (2018) Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120
20. Lucia F, Visvikis D, Vallières M et al (2018) External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging 46:864–877
21. Whitney HM, Li H, Ji Y, Liu P, Giger ML (2020) Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. J Med Imaging (Bellingham) 7:012707
22. Wang H, Zhang J, Bao S et al (2020) Preoperative MRI-based radiomic machine-learning nomogram may accurately distinguish between benign and malignant soft-tissue lesions: a two-center study. J Magn Reson Imaging. https://doi.org/10.1002/jmri.27111
23. Zhang L-L, Huang M-Y, Li Y et al (2019) Pretreatment MRI radiomics analysis allows for reliable prediction of local recurrence in non-metastatic T4 nasopharyngeal carcinoma. EBioMedicine 42:270–280
24. Penzias G, Singanamalli A, Elliott R et al (2018) Identifying the morphologic basis for radiomic features in distinguishing different

Gleason grades of prostate cancer on MRI: preliminary findings. PLoS One 13:e0200730

25. Jackson EF, Barboriak DP, Bidaut LM, Meyer CR (2009) Magnetic resonance assessment of response to therapy: tumor change measurement, truth data and error sources. Transl Oncol 2:211–215

26. Clark K, Vendt B, Smith K et al (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26:1045–1057

27. Nioche C, Orlhac F, Boughdad S et al (2018) LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res 78:4786–4789

28. Zwanenburg A, Vallières M, Abdalah MA et al (2020) The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high throughput image-based phenotyping. Radiology 295:328–338

29. Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I (2015) 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. PLoS One 10:e0145063

30. Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841

31. Tustison NJ, Avants BB, Cook PA et al (2010) N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29:1310–1320

32. Nyúl LG, Udupa JK, Zhang X (2000) New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 19:143–150

33. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodology 57:289–300

34. Qu L, Wang S, Yap P-T, Shen D (2019) Wavelet-based semi-supervised adversarial learning for synthesizing realistic 7T from 3T MRI. Med Image Comput Comput Assist Interv 11767:786–794

35. Zhong J, Wang Y, Li J et al (2020) Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development. Biomed Eng Online 19:4

36. Modanwal G, Vellal A, Buda M, Mazurowski MA (2020) MRI image harmonization using cycle-consistent generative adversarial network. Medical Imaging 2020: Computer-Aided Diagnosis. https://doi.org/10.1117/12.2551301.