T2w-MRI signal normalization affects radiomics features reproducibility

Elisa Scalco[a)]
CNR, Institute of Biomedical Technologies (ITB), Segrate, Italy
CNR, Institute of Molecular Bioimaging and Physiology (IBFM), Segrate, Italy

Antonella Belfatto
CNR, Institute of Molecular Bioimaging and Physiology (IBFM), Segrate, Italy

Alfonso Mastropietro
CNR, Institute of Biomedical Technologies (ITB), Segrate, Italy
CNR, Institute of Molecular Bioimaging and Physiology (IBFM), Segrate, Italy

Tiziana Rancati
Fondazione IRCCS Istituto Nazionale dei Tumori, Prostate Cancer Program, Milano, Italy

Barbara Avuzzi
Radiation Oncology 1, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy

Antonella Messina
Radiology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy

Riccardo Valdagni
Fondazione IRCCS Istituto Nazionale dei Tumori, Prostate Cancer Program, Milano, Italy
Radiation Oncology 1, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy
Department of Oncology and Hemato-oncology, Università degli Studi di Milano, Milano, Italy

Giovanna Rizzo
CNR, Institute of Biomedical Technologies (ITB), Segrate, Italy
CNR, Institute of Molecular Bioimaging and Physiology (IBFM), Segrate, Italy

(Received 8 August 2019; revised 13 January 2020; accepted for publication 14 January 2020; published xx xxxx xxxx)

Purpose: Despite its increasing application, radiomics has not yet demonstrated a solid reliability, due to the difficulty in replicating analyses. The extraction of radiomic features from clinical MRI (T1w/T2w) presents even more challenges because of the absence of well-defined units (e.g. HU). Some preprocessing steps are required before the estimation of radiomic features and one of this is the intensity normalization, that can be performed using different methods. The aim of this work was to evaluate the effect of three different normalization techniques, applied on T2w-MRI images of the pelvic region, on radiomic features reproducibility.

Methods: T2w-MRI acquired before (MRI1) and 12 months after radiotherapy (MRI2) from 14 patients treated for prostate cancer were considered. Four different conditions were analyzed: (a) the original MRI (No_Norm); (b) MRI normalized by the mean image value (Norm_Mean); (c) MRI normalized by the mean value of the urine in the bladder (Norm_ROI); (d) MRI normalized by the histogram-matching method (Norm_HM). Ninety-one radiomic features were extracted from three organs of interest (prostate, internal obturator muscles and bulb) at both time-points and on each image discretized using a fixed bin-width approach and the difference between the two time-points was calculated (*Δfeature*). To estimate the effect of normalization methods on the reproducibility of radiomic features, ICC was calculated in three analyses: (a) considering the features extracted on MRI2 in the four conditions together and considering the influence of each method separately, with respect to No_Norm; (b) considering the features extracted on MRI2 in the four conditions with respect to the inter-observer variability in region of interest (ROI) contouring, considering also the effect of the discretization approach; (c) considering *Δfeature* to evaluate if some indices can recover some consistency when differences are calculated.

Results: Nearly 60% of the features have shown poor reproducibility (ICC < 0.5) on MRI2 and the method that most affected features reliability was Norm_ROI (average ICC of 0.45). The other two methods were similar, except for first-order features, where Norm_HM outperformed Norm_Mean (average ICC = 0.33 and 0.76 for Norm_Mean and Norm_HM, respectively). In the inter-observer setting, the number of reproducible features varied in the three structures, being higher in the prostate than in the penile bulb and in the obturators. The analysis on *Δfeature* highlighted that more than 60% of the features were not consistent with respect to the normalization method and confirmed the high reproducibility of the features between Norm_Mean and Norm_HM, whereas Norm_ROI was the less reproducible method.

Conclusions: The normalization process impacts the reproducibility of radiomic features, both in terms of changes in the image information content and in the inter-observer setting. Among the considered methods, Norm_Mean and Norm_HM seem to provide the most reproducible features with respect to the original image and also between themselves, whereas Norm_ROI generates less reproducible features. Only a very small subset of feature remained reproducible and independent in any tested condition, regardless the ROI and the adopted algorithm: skewness or kurtosis, correlation and one among Imc2, Idmn and Idn from GLCM group. © *2020 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.14038]

Key words: MRI intensity normalization, prostate cancer, radiomics, reproducibility assessment

1. INTRODUCTION

Radiomic analysis was introduced in the field of oncology only recently[1] but it is increasingly adopted in numerous studies, becoming one of the most relevant techniques to extract quantitative biomarkers from medical images. In fact, radiomics has revealed its potential in identifying[2] and classifying tumors,[3] and in predicting treatment response both for tumor[4] and normal tissues.[5] However, results of radiomic analyses are often difficult to replicate, due to the lack of standardized procedure in image acquisition, reconstruction, processing and analysis. For this reason, radiomic features that present high classification/prediction power should also present high reliability, since both these properties are necessary to build a reliable radiomic signature.[6,7]

In the previous years, several works have dealt with the assessment of radiomic features reliability, both considering the evaluation of repeatability, i.e features that remain the same when calculated multiple times in the same subject with the same conditions, and reproducibility, i.e. features that remain the same when calculated using different acquisition or processing conditions.[8] The focus of these works was principally on the evaluation of the impact of image acquisition and reconstruction,[9,10] image discretization[11] and region of interest (ROI) delineation,[12] especially considering positron emission tomography (PET) and computed tomography (CT) images. Other factors influencing the robustness of radiomic feature computation have not been exhaustively explored yet, particularly regarding the image processing aspects (noise filtering, artifacts correction, algorithms used for features computation, etc.).[8]

While CT and PET imaging have established their role in radiomics,[1,13] magnetic resonance imaging (MRI) showed some initial difficulties in imposing itself as a robust imaging modality for the extraction of reliable features, despite its great potential in assessing several tissue properties. This fact can be explained by the nonquantitative image intensity signal of the standard clinical acquisitions (especially T2-weighted (T2w) and T1-weighted (T1w) MRI), which makes the comparison of radiomic features within a study population nonfeasible, even if the same acquisition protocol is adopted; in addition, even sequences providing quantitative parameters [such as apparent diffusion coefficient (ADC) maps] are subject to reproducibility limitations due to the large spectrum of acquisition parameters and possible artifacts. Nonetheless, the great availability of T2w-MRI in clinical practice and its ability in offering excellent anatomical details and contrast, together with the adoption of some necessary preprocessing steps, have allowed the routine use of radiomic analyses even on these images. In fact, the adoption of some image processing pipelines, aimed at harmonizing image resolution, correcting artifacts, such as the magnetic field inhomogeneities, and adjusting the nonquantitative image intensity values, was also suggested by the Image Biomarker Standardization Initiative (IBSI).[14]

Image intensity normalization is a necessary step if nonquantitative MRI images are considered for different subjects or for longitudinal studies, as the aim of this procedure was to remove the variability between patients/longitudinal studies and increase the MRI repeatability.[15] Different normalization methods are described in the literature and adopted by different authors, but the effects of this processing on radiomic features extraction have been studied only by very few groups,[7,16] of which only one was based on prostate MRI for the evaluation of radiomics repeatability. It is important to understand if the normalization step affects radiomics reproducibility and if features obtained using different methods are reliable; this can help the optimization of the study design or the comparison between results coming from different studies. Intensity standardization is most frequently carried out following one of these approaches: (a) by normalizing gray level values with respect to a ROI with fixed and stable value,[17] (b) by centering the image at its mean value,[18] or (c) by adjusting the histogram to a reference one.[19]

In the context of prostate cancer, radiomics is currently routinely performed on T2w-MRI, as the most commonly available MRI acquisition,[15,20] to detect the tumor[21] or to explore the association with biochemical recurrence after radiotherapy (RT)[22] and to assess the effect of irradiation on organs at risk, such as the internal obturator muscles.[23] It is known that the performance of radiomic features are dependent on the type of analyzed tissue, for example tumor or normal tissue. Therefore, the reproducibility evaluation should be performed taking into consideration both pathological and healthy tissues in order to improve knowledge about how the structural properties of the different organs can impact the radiomics estimation.[20] In this context, recent studies have already evaluated the cross-site reproducibility and discriminability characteristics of different radiomic features and feature families in a multisite setting.[20,24]

In this work we aimed at evaluating in a cohort of prostate cancer patients if T2w-MRI signal normalization has an impact on the image information content provided by textural features, by means of the evaluation of their reproducibility. Features reproducibility was also evaluated taking into consideration other relevant conditions in the radiomics procedure, i.e. the ROI delineation by multiple observers[25] and the image discretization approach.[26] In addition, the impact of different normalization techniques was also assessed on delta features extracted from longitudinal images, as a typical condition that occurs in radiomics, especially for RT evaluation.

2. MATERIALS AND METHODS

2.A. Study population and image acquisition

Fourteen patients treated for prostate cancer with exclusive radical external beam RT at the National Cancer Institute in Milan were considered. The study protocol was approved by the local Ethics Committee (INT 73/13) and written informed consent was obtained from the patients involved in this study.

T2w-MRI was performed before RT (MRI1) and 12 months after RT completion (MRI2) using a 1.5 T scanner (Achieva, Philips Medical Systems, Best, the Netherlands) equipped with a SENSE-XL-Torso coil with 16 channels. Images were acquired using a Turbo Spin Echo sequence with axial slicing (TR = 4000 ms and TE = 120 ms; resolution = 0.456 × 0.456 mm; matrix = 268 × 768; slice thickness = 3 mm; NSA = 4).

2.B. ROI identification

For this study, three different ROIs were selected: the central zone of the prostate, the penile bulb and the obturator muscles (both right and left). This choice was made in order to take into consideration a target organ, such as the prostate, and normal tissues receiving irradiation during RT, such as the penile bulb and the obturators. In addition, these structures have different range of gray-level intensities, covering different regions of the whole image histogram.

ROI contours were manually delineated on the MRI1 (see Fig. 1) independently by two operators, a medical imaging researcher (C1.1) and a senior medical physicist (C2) with 6 and 15 years of experience in MRI pelvic images, respectively, and converted into a binary label mask using 3DSlicer.[27] The first operator recontoured the ROIs to assess the intra-observer reproducibility (C1.2). Contours were then automatically propagated on MRI2 by applying the deformation field estimated by the elastic registration between the two images. A more detailed description of the image registration and contour propagation procedure can be found in previous works.[23,28]

2.C. Image processing and intensity normalization

The image processing workflow is schematically represented in Fig. 2. All T2w-MR images were first corrected for magnetic field inhomogeneities by using the nonparametric nonuniform intensity normalization (N4) algorithm.[29] Regarding the normalization step, four conditions were considered:

1. No normalization (No_Norm): no intensity homogenization was performed
2. Normalization by the mean image value (Norm_Mean): both images (MRI1 and MRI2) were normalized by centering them at their respective mean value with standard deviation of the whole original image, as suggested in the user-guide of PyRadiomics.[30]

$$f(x) = \frac{(x - \mu_x)}{\sigma_x} + 3\sigma_x$$

where x and $f(x)$ are the original and normalized intensity, respectively; μ_x and σ_x are the mean and standard deviation of the image.

3. Normalization by the mean and standard deviation of the urine in the bladder (Norm_ROI): both images (MRI1 and MRI2) were normalized by centering them at the mean value and standard deviation of the bladder in their respective original images.[17] The signal intensity of the urine in the bladder was chosen as a signal not influenced by dose- or time-dependent factors.

$$f(x) = \frac{(x - \mu_{ROI})}{\sigma_{ROI}} + 3\sigma_{ROI}$$

where x and $f(x)$ are the original and normalized intensity, respectively; μ_{ROI} and σ_{ROI} are the mean and standard deviation of the urine within the bladder.

4. Normalization using the histogram-matching method (Norm_HM): this algorithm, proposed by Nyul et al.,[31] seeks the global correspondence between MRI1 and MRI2 in a specific number of reference points of the histogram. The intensity value of the reference points of the MRI2 histogram are linearly mapped on the intensity value of the corresponding reference point of the MRI1 histogram. The following configuration was adopted[31]: number of histogram bins = 256; reference points = the 9 deciles and the maximum and minimum percentiles. MRI2 was corrected so that its histogram matched that of MRI1. Conversely, no normalization was applied to MRI1.

2.D. Features computation

Ninety-one radiomic features were computed in each ROI and for every normalization condition on MRI1 and MRI2 using PyRadiomics open-source software[30] (version 2.0.1), implemented in Python. Specifically, the following indices were extracted: 18 first-order (FO), 22 from Gray-Level Co-occurrence Matrix (GLCM), 16 from Gray-Level Run-Length Matrix (GLRLM), 16 from Gray-Level Size-Zone Matrix (GLSZM), 14 from Gray-Level Dependence

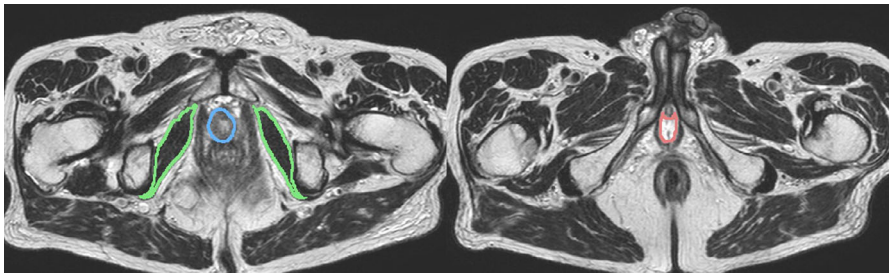

FIG. 1. Contours of internal obturator muscles (green), prostate (blue) and bulb (red) manually delineated on T2w-MRI acquired before RT.

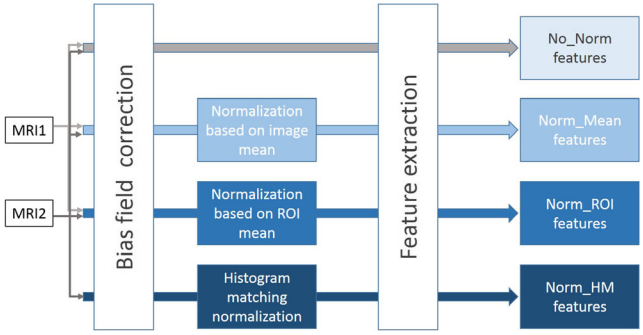

FIG. 2. Schematic representation of the image processing workflow for the extraction of radiomic features.

Matrix (GLDM) and five from Neighbouring Gray Tone Difference Matrix (NGTDM). Resegmentation (i.e. the exclusion of outliers farther from the mean than 3 standard deviations) was performed from the original ROI. Discretization was performed by considering a fixed bin-width (FBW), rather than considering a fixed bin count (FBC), since FBW has highlighted higher reproducibility in MRI inter-observer contours variability.[26] A bin-width of 25 was used for No_Norm, Norm_ROI and Norm_HM and a bin-width of 5 was used for Norm_Mean, in order to guarantee a similar number of bins (median [range]: 42 [34–60] for No_norm; 37 [29–54] for Norm_Mean; 36 [31–52] for Norm_ROI; 40 [35–58] for Norm_HM) among the images normalized using the different methods. In addition, the original image was also discretized using the FBC approach with 64 bins, as it can introduce an intrinsic normalizing effect.[14] Features extraction was performed in 2D, since the slice thickness was much larger than the in-plane dimension (3 mm vs. 0.456 mm).[23] For this reason, images were not resampled to isotropic voxels, in order to not introduce another step involving interpolation, as pixels were already isotropic in the in-plane resolution. No filtering was applied to the images.

Differences of each feature between MRI1 and MRI2 were calculated as follows:

$$\Delta feature = \frac{feature_2 - feature_1}{feature_1} \times 100.$$

where $feature_1$ and $feature_2$ corresponded to pre- and post-RT values, respectively.

2.E. Reproducibility estimation

The reproducibility of radiomic features was tested using the intraclass correlation coefficient (ICC) in three different conditions:

1. Features extracted on MRI2 and considering the No_Norm condition as the reference image with the original texture, in order to evaluate the effect of normalization on the image information content. The FBW discretization was adopted in this phase. The impact of using a normalization method was assessed a) by considering the effect of the four methods (No_Norm, Norm_Mean, Norm_ROI, Norm_HM) together, in order to identify the features that are less affected by the normalization procedure, regardless the algorithm chosen to perform it (ICC global), and b) by considering the influence of each method separately, with respect to the reference image (No_Norm vs. Norm_Mean; No_Norm vs. Norm_ROI; No_Norm vs. Norm_HM), in order to identify the normalization algorithm that less affects the extraction of radiomic features (ICC couples).
2. Features extracted in the ROIs delineated by the different operators, to evaluate the normalization approach that could better preserve reproducibility between observer delineations. The spatial overlap between delineations was estimated using the Dice coefficient.[32] ICC was computed for each pair of inter-observer (C1.1 vs. C2) and intra-observer (C1.1 vs. C1.2) delineations, considering five conditions: No_Norm, Norm_Mean, Norm_ROI and Norm_HM, where images were discretized using the FBW approach, and No_Norm discretized with FBC approach. A feature was considered reproducible if it reached at least a significant ($P < 0.05$ after Bonferroni correction for multiple comparisons) ICC of 0.75 in both the experiments (intra- and inter-observer ICC).[26]
3. *Δfeatures* calculated using Norm_mean, Norm_ROI and Norm_HM with FBW discretization, to evaluate if the calculation of delta can mitigate or increase the alteration in the information content induced by the normalization. The aim of this analysis was to assess whether different normalization methods can lead to comparable and highly reproducible delta features. In

this analysis, No_Norm condition was not considered, since the *Δfeatures* computed without performing any normalization are meaningless. For this reason, the ICC was calculated only considering the three normalization methods together (ICC_global) and considering all the possible couple combinations, namely Norm_Mean vs. Norm_ROI (n1-n2), Norm_Mean vs. Norm_HM (n1-n3) and Norm_ROI vs. Norm_HM (n2-n3).

ICC (two-way mixed effect model, single rater type)[33] was computed for consistency estimation in conditions 1) and 3):

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$$

and for absolute agreement estimation in condition 2):

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$$

where MS_R is the mean square for rows (observations), MS_C is the mean square for columns, MS_E is the mean square for error and k is the number of raters (normalization methods or observers). It was previously reported that ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability and values greater than 0.9 indicate excellent reliability.[34] ICC values were considered significant for P-values < 0.05, after Bonferroni correction for multiple comparisons; features with nonsignificant ICC were considered as poorly reproducible, even if ICC was greater than 0.5. The Spearman correlation between radiomic features and the ROI volume was assessed, since it has been reported that many features intrinsically embed volume information.[35] In this way, it is possible to discard the highly correlated ones (significant P-value < 0.05, after Bonferroni correction for multiple comparisons), in order to consider only features that embed information related to the texture. In addition, the assessment of inter-correlations between features was also performed using Spearman correlation (significant P-value < 0.05 after Bonferroni correction for multiple comparisons). The reproducibility results were reported considering the whole set of features, and results of this correlation analysis were used to be sure that the final set of most reproducible features did not contain indices highly correlated within themselves.

3. RESULTS

3.A. Reproducibility on features extracted from MRI2

3.A.1. ICC global

The reproducibility evaluated on MRI2 taking into consideration the influence of any normalization technique on radiomic features revealed that most parameters are very sensitive to this image processing step. In fact, 67% of the features in the prostate, 38% in the obturators and 63% in the bulb have shown poor reproducibility (ICC < 0.5 or non-significant ICC; see Fig. 3 for more details). Only a small part of the features presented an ICC value greater than 0.9 (14% in the prostate, 12% in the obturators and 14% in the bulb). Some of these features were the same in any ROI: kurtosis and skewness for FO features, correlation, Inverse Difference Moment Normalized (Idmn) and Inverse Difference Normalized (Idn) for GLCM features, Gray-Level Non-Uniformity for GLSZM features and Coarseness for NGTDM features. However, the two lasts presented a high correlation with volume and thus they were discarded; moreover, skewness and kurtosis presented high correlation within themselves as well as Idmn and Idn. Looking at the features class, the average ICC value within every group was between 0.44 and 0.68 (see Table I and Table S1 for more details).

3.A.2. ICC couples

When ICC was assessed for each normalization technique separately, with respect to the reference image without any normalization, it was highlighted that the method that most affected the feature estimation was Norm_ROI. In fact, as can

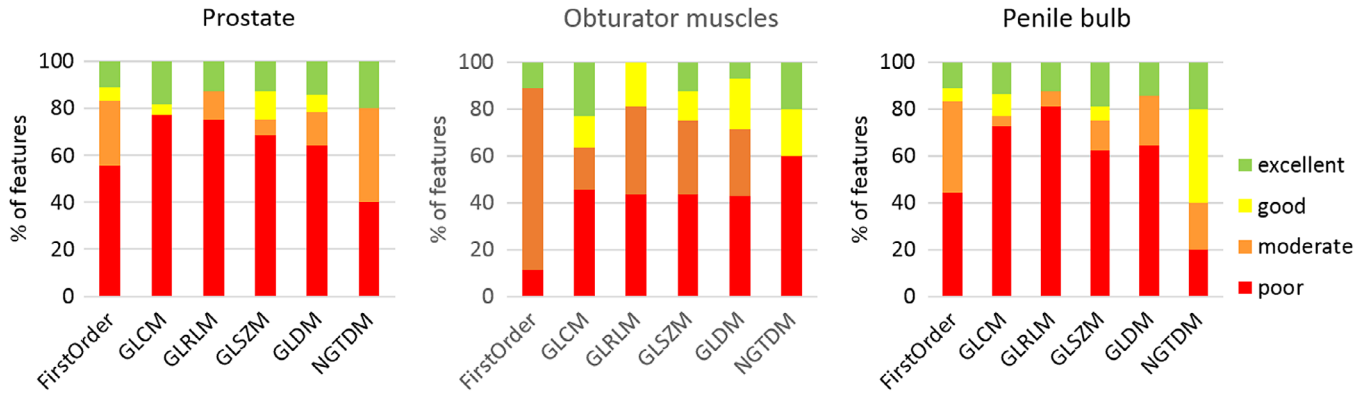

FIG. 3. Percentage of poor, moderate, good and excellent features for each class and in each structure, based on ICC values calculated considering the four normalization techniques on MRI2. ICC, intraclass correlation coefficient.

TABLE I. Average ICC value for each feature class, calculated on MRI2 among the features belonging to the considered class, for the three structures.

	ICC global			No_Norm vs. Norm_Mean			No_Norm vs. Norm_ROI			No_Norm vs. Norm_HM		
	Prost	Obt.	Bulb	Prost	Obt.	Bulb	Prost	Obt.	Bulb	Prost	Obt.	Bulb
FO	0.55	0.61	0.55	0.33	0.36	0.31	0.50	0.46	0.53	0.72	0.83	0.71
GLCM	0.49	0.51	0.50	0.69	0.65	0.69	0.44	0.45	0.46	0.67	0.72	0.68
GLRLM	0.44	0.48	0.46	0.65	0.57	0.65	0.35	0.35	0.43	0.65	0.69	0.63
GLSZM	0.47	0.50	0.50	0.66	0.60	0.67	0.40	0.41	0.47	0.66	0.69	0.65
GLDM	0.50	0.51	0.51	0.68	0.60	0.68	0.43	0.40	0.48	0.68	0.70	0.65
NGTDM	0.60	0.53	0.68	0.76	0.77	0.82	0.57	0.59	0.63	0.73	0.79	0.78
Mean	0.50	0.52	0.51	0.61	0.57	0.61	0.43	0.43	0.48	0.68	0.73	0.67

Values are reported as ICC global and ICC couples for each normalization technique, with respect to the reference image.
ICC, intraclass correlation coefficient.

be seen in Fig. 4 and in Table I, the Norm_ROI presented the lowest ICC values, regardless of the feature class and ROI. As for the other two normalization methods, the reported ICC values were similar for textural features classes, with a clear tendency of Norm_HM to maintain a larger number of reproducible features than Norm_Mean in the obturators (56 vs. 40 features with ICC > 0.75, for Norm_HM and Norm_-Mean, respectively). In the FO group, the Norm_HM method was the only one able to better preserve the original features (ICC of 0.72, 0.83 and 0.71 in the prostate, obturators and bulb, respectively). The ICC values, confidence intervals and *P*-values of each feature in the three ROIs can be found in Table S1.

3.B. Reproducibility with respect to inter-observer delineations

The intra- and inter-observer variability in ROI delineation, in terms of the Dice coefficient, was reported in Table II. The reproducibility analysis has highlighted very different results, depending on the considered ROI. In fact, the number of reproducible features varied in the three structures, showing high reproducibility in the prostate (except for the Norm_FBC normalization) and poor or very poor reproducibility in the penile bulb and in the obturators (see Table II). Norm_ROI condition presented the higher number of features in all the three organs, whereas Norm_HM was similar to the original condition. Discretization with FBC led to discordant results: in the prostate and in the obturator muscles it generated less reproducible features, whereas in the penile bulb it presented the highest number. The complete set of ICC values, together with confidence intervals and *P*-values can be found in Table S2.

3.C. Reproducibility on *Δfeatures*

3.C.1. ICC global

The reproducibility of *Δfeatures* was similar or even lower than that measured on MRI2, as reported in Table III and shown in Fig. 5. In fact, the number of poorly reproducible features remained stable for the prostate and the penile bulb (66% and 64%, respectively) and increased to 63% for the obturators, whereas the number of highly stable features decreased to 7% in the prostate, 11% in the obturators and 8% in the bulb. In particular, the most reproducible features were a subset of the stable features found in the previous analysis (correlation, Informational Measure of Correlation 2 (Imc2), Idmn and Idn from GLCM, coarseness from NGTDM, presenting ICC > 0.9 in each ROI; kurtosis and skewness from FO, Gray-Level Non-Uniformity from GLSZM presenting ICC > 0.9 in two out of three ROIs). Similar to the POST-RT analysis, some of these features presented high correlation with volume (Gray-Level Non-Uniformity from GLSZM) or within themselves (skewness and kurtosis, Idmn, Idn and Imc2), thus reducing the number of reproducible and independent features.

3.C.2. ICC couples

The analysis between the coupled normalization methods confirmed findings on MRI2. In particular, an excellent reproducibility between *Δfeatures* calculated using Norm_-Mean and Norm_HM was highlighted (see the second column in Fig. 6 and Table III). Only FO features presented low ICC values (78% of features with ICC < 0.5 in each ROI), while the average ICC in the other classes was in the range of 0.79–0.93. On the contrary, features were not stable when Norm_ROI was used (average ICC of the different classes in the range of 0.22-0.58), except for the features that showed high reproducibility with ICC global (see the first and third columns in Fig. 6). The ICC values of each feature in the three ROIs can be found in Table S3.

4. DISCUSSION

In this work, an extensive evaluation of the influence of the image intensity normalization procedure on the reproducibility of radiomic features in longitudinal T2w-MRI studies and with respect to inter-observer variability in ROI contouring was presented. Three different normalization algorithms, generally adopted in literature within the field of

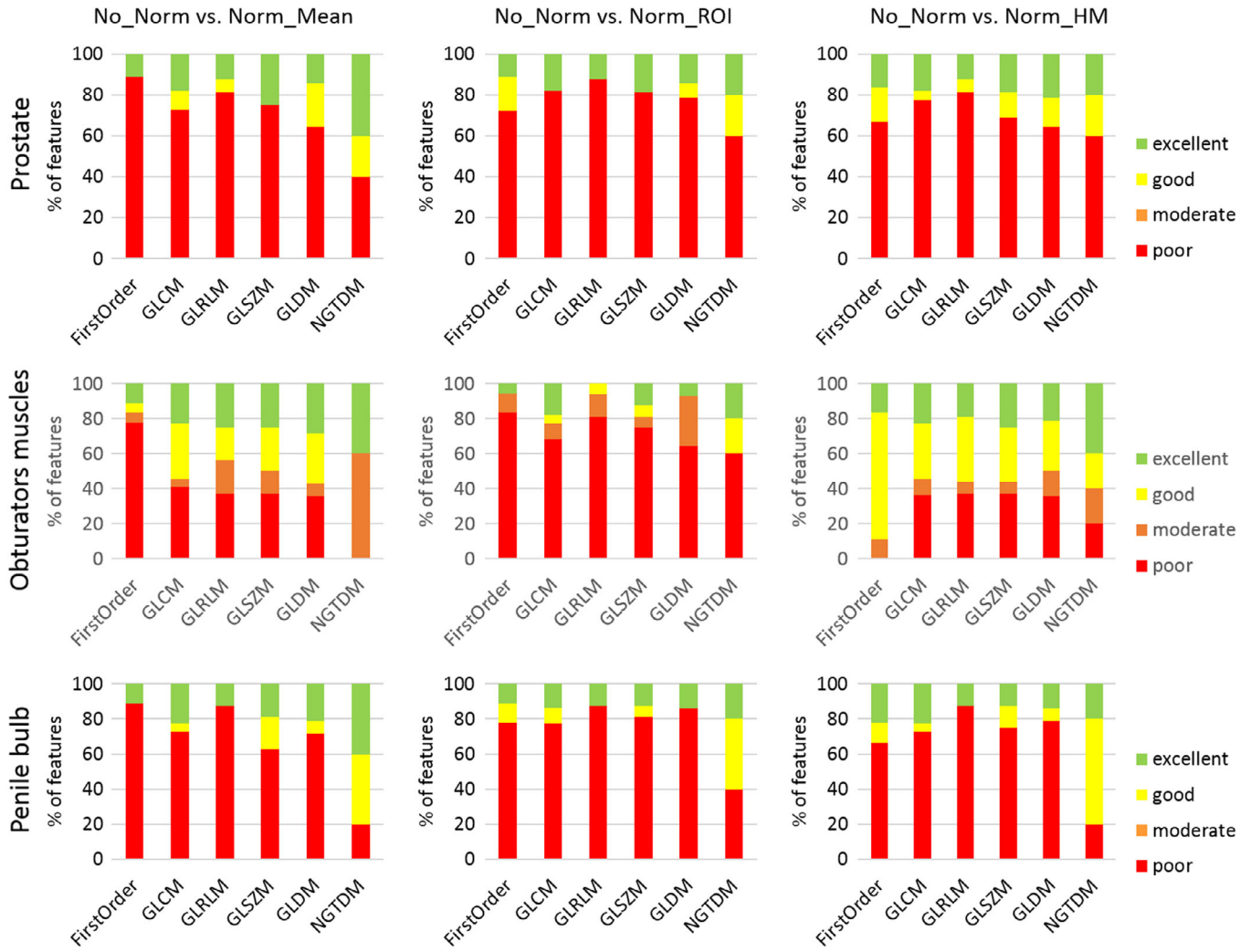

FIG. 4. Percentage of poor, moderate, good and excellent features for each class and in each structure, based on ICC values calculated considering the three normalization techniques separately with respect to the reference, on MRI2. ICC, intraclass correlation coefficient.

TABLE II. Reproducibility in the intra- and inter-observer setting in the three ROIs.

		Prostate	Obturator muscles	Penile bulb
Delineation agreement (Dice)	Intra-observer	0.83 ± 0.06	0.80 ± 0.05	0.78 ± 0.10
	Inter-observer	0.77 ± 0.09	0.75 ± 0.08	0.73 ± 0.13
Number of reproducible features	No_Norm	55	5	13
	Norm_Mean	48	3	11
	Norm_ROI	59	10	15
	Norm_HM	55	4	11
	Norm_FBC	8	2	21

Contour variability is measured with Dice coefficient (reported as mean ± SD). The number of reproducible features considering the different normalization is reported. Features are considered reproducible if a significant ICC > 0.75 is computed both in the intra- and inter-observer settings.
ICC, intraclass correlation coefficient.

radiomic MRI, were studied. The main finding of this study was that intensity normalization affects the image information content and thus the extraction of textural features, both on single acquisition and on delta, especially when the normalization through the mean value of a stable ROI was adopted. As for the inter-observer reproducibility, results were mostly dependent on the ROI and on the used discretization approach. Among the 91 features considered in this work, only a very small subset of features remained reproducible to those computed on the original image, excluding those correlated to volume and other features, in any condition (skewness or kurtosis from FO class, correlation and one among Imc2, Idmn and Idn from GLCM class), with an ICC > 0.75 regardless the ROI and the adopted algorithm.

From these findings we drew some main conclusions. First, the normalization techniques have a high influence on the radiomic computation. In particular, the choice of the normalization method may alter the textural features and not only the histogram, on which the normalization algorithms are

TABLE III. Average ICC value for each feature class, calculated among *Δfeatures* belonging to the considered class, for the three structures.

	ICC global			Norm_Mean vs. Norm_ROI			Norm_Mean vs. Norm_HM			Norm_ROI vs. Norm_HM		
	Prost	Obt.	Bulb	Prost	Obt.	Bulb	Prost	Obt.	Bulb	Prost	Obt.	Bulb
FO	0.36	0.35	0.30	0.24	0.24	0.22	0.40	0.42	0.40	0.43	0.45	0.34
GLCM	0.61	0.50	0.49	0.57	0.43	0.47	0.93	0.87	0.80	0.50	0.48	0.42
GLRLM	0.37	0.37	0.46	0.28	0.30	0.43	0.91	0.84	0.79	0.25	0.30	0.36
GLSZM	0.40	0.45	0.47	0.34	0.39	0.43	0.88	0.84	0.80	0.29	0.41	0.39
GLDM	0.45	0.43	0.49	0.37	0.35	0.46	0.92	0.83	0.76	0.32	0.39	0.40
NGTDM	0.57	0.56	0.59	0.49	0.45	0.56	0.93	0.86	0.82	0.44	0.58	0.50
Mean	0.45	0.43	0.45	0.38	0.35	0.41	0.81	0.77	0.71	0.38	0.42	0.39

Values are reported as ICC global and ICC couples considering the three normalization techniques.
ICC, intraclass correlation coefficient.

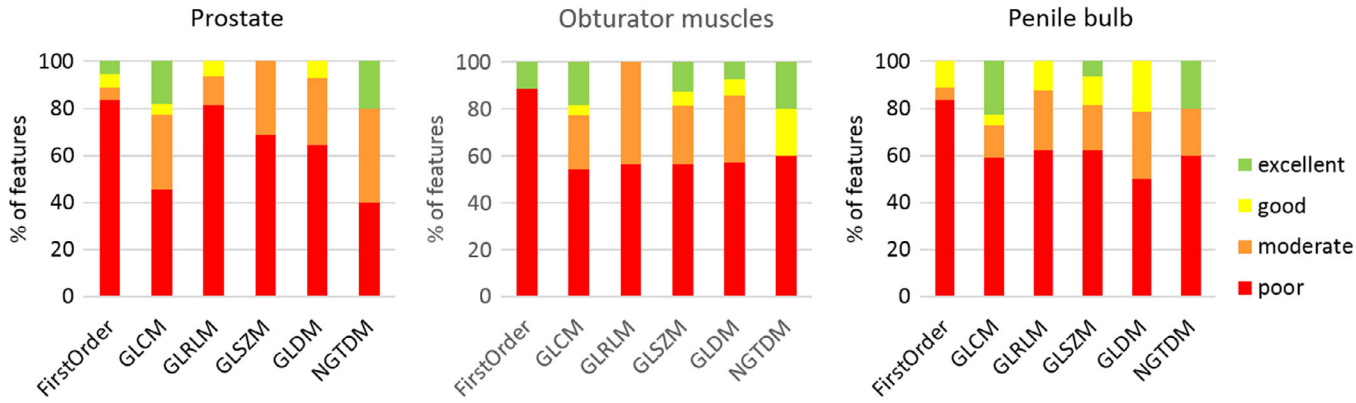

FIG. 5. Percentage of poor, moderate, good and excellent features for each class and in each structure, based on ICC values calculated considering the three normalization techniques on *Δfeatures*. ICC, intraclass correlation coefficient.

predominantly based. Considering the analysis on MRI2, where the reference were the features extracted from the original image, it can be seen that the description of the texture was modified by the application of a normalization technique and it happened particularly when Norm_ROI was applied.

Second, among the three analyzed methods, Norm_HM and Norm_Mean seem to better preserve the original pattern of the MRI images, with a tendency of features computed on Norm_HM images of being more similar to the original ones in the obturator muscles. Nonetheless, in the estimation of FO features, the histogram-matching method is the only one that allows high reproducibility. On the contrary, Norm_ROI was highlighted as the method that mostly affects radiomics reproducibility. For this reason, it should be discarded in favor of one of the other two. This may be explained by a different and flatter gray-level distribution after normalization with respect to the other techniques, which resulted in a different bin association when an FBW discretization was adopted.

When the influence of normalization methods was evaluated on *Δfeatures*, the calculation of difference does not improve the reproducibility with respect to the features calculated on MRI2 only; on the contrary, an increase in poorly reproducible features was found in the obturators. Looking at the coupled analyses, Norm_Mean and Norm_HM can estimate features highly reproducible between themselves, except for FO features, as also found on the features computed on MRI2. This may suggest that results found using these two methods might be considered interchangeably and equally valid (excluding FO features), especially in longitudinal studies. As for the Norm_ROI method, its poor reproducibility found also on *Δfeatures,* may be explained by the possible intrinsic modifications occurred between the first and second MRI acquisition on the ROI chosen as reference (in this case the urinary bladder). Changes in urinary composition might have happened and, thus, correcting for these modifications could alter the normalization process. This problem would have been risen even if another ROI was chosen as reference, since it is quite impossible to find a reference region that never changes its properties.

Our findings were similar among the different types of radiomic features, suggesting that there are no textural matrices (GLCM, GLRLM, GLSZM, GLDM nor NGTDM) more stable than others. The only group that presented a slightly higher ICC was the first-order group between No_Norm and Norm_HM. This fact can be explained by the similar histogram distributions between MRI1 and MRI2 of the same patient, being the MRI acquisitions performed on the same

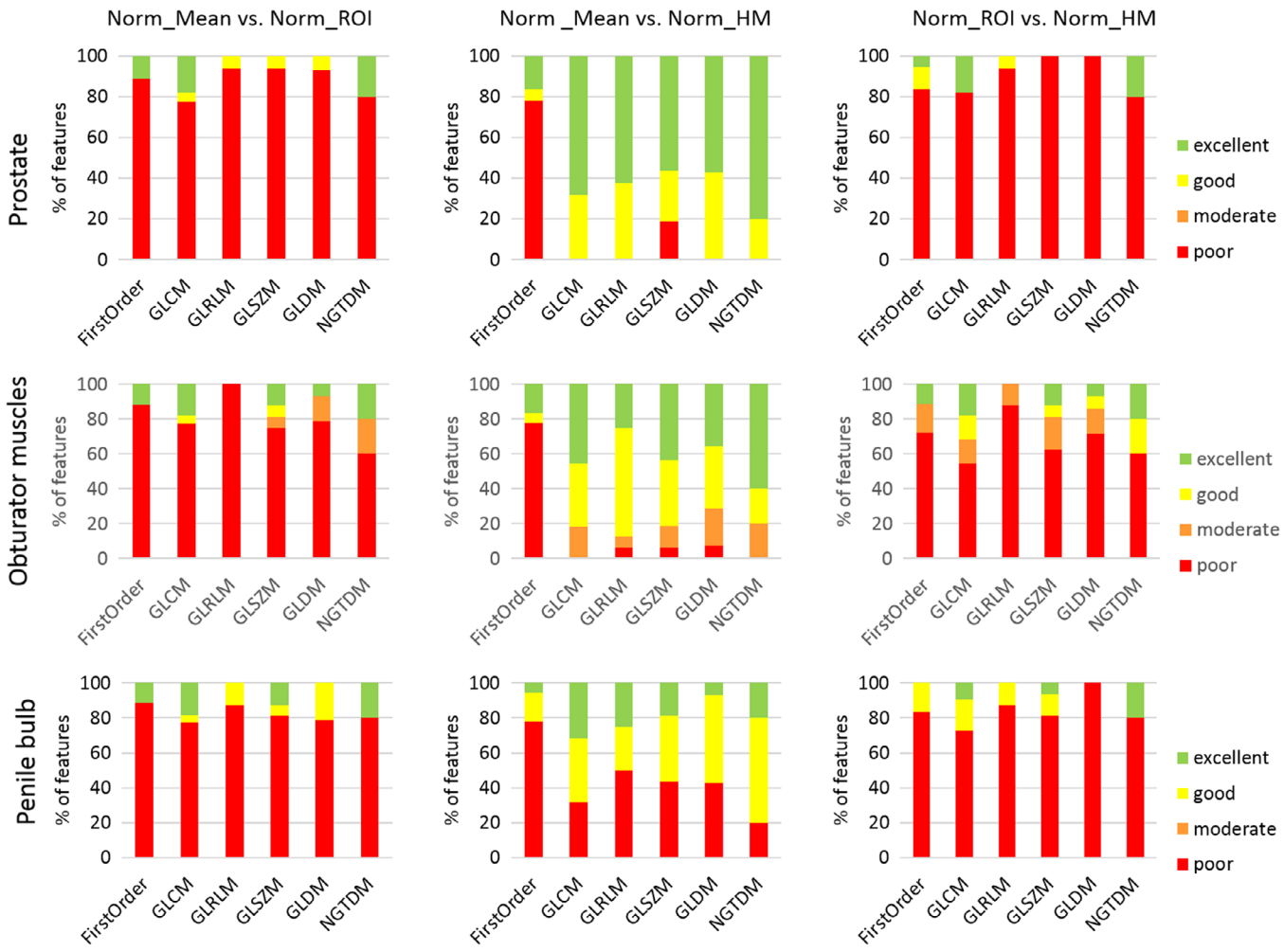

FIG. 6. Percentage of poor, moderate, good and excellent features for each class and in each structure, based on ICC values calculated considering the couples between the three normalization techniques, on *Δfeatures*. ICC, intraclass correlation coefficient.

scanner and with the same protocol. At the same time, the results were confirmed in the three organs considered in this work (i.e. prostate, internal obturator muscle and penile bulb), with a tendency of maintaining higher reproducible features on MRI2 in the obturators for Norm_Mean and Norm_HM conditions, with respect to Norm_ROI. This suggests that the conclusions drawn from this work regarding the influence of normalization on the image information content might be extended to other body districts scanned by a T2w-MRI protocol, as the considered organs represent three different gray-level distribution in the images (the obturators, as muscular tissue, present hypo-intense T2w-MRI values, the penile bulb is mainly represented by hyper-intense values and the prostate is in an intermediate range).

On the contrary, results of intra- and inter-observer reproducibility evaluation have highlighted that contour variability plays a different role with respect to image normalization depending on the considered ROI. In fact, the central zone of the prostate, which is a quite homogeneous region, has shown a high and stable number of reproducible features, regardless of the normalization method (except for the FBC condition), despite a certain variability in ROI definition (Dice of 0.83 and 0.77 for intra- and inter-observer variability). Conversely, the penile bulb, being a very heterogeneous and small structure, with lower delineation agreement, is more affected by the chosen normalization approach. The obturator muscles also showed poor reproducibility, probably due to the contouring variability in the upper slices, where ROIs become smaller. In this context, Norm_ROI seems to be the normalization method that better preserves reproducibility: this may be explained by the changes in the intensity ranges in the ROIs that, combined with the FBW approach, flatten the intensity distribution, which becomes less sensitive to the contour variability. However, one should take into consideration that Norm_ROI was the method with lower reproducibility in the other two conditions. The FBC discretization has provided discordant results, suggesting that an FBW approach, combined with a normalization algorithm, could better preserve both the original information content and the inter-observer reproducibility.

This work can contribute to the literature regarding the assessment of radiomics reliability on T2w-MRI. Regarding the evaluation of image intensity normalization, Collewet et al.[16] firstly evaluated the impact of three normalization

methods, different from those considered in this study, on the classification performance in T2w-MRI cheese images, but how the normalization impacts features reproducibility on human studies was not studied. A very recent work[7] has proposed an in depth evaluation of image intensity normalization on T2w-MRI and ADC of the prostate, performed by an intensity rescaling and shifting and by the ROI-based method (using a muscle ROI as reference). They found that repeatability is very sensitive to preprocessing steps (both normalization and discretization) and that it was lower when ROI-based normalization is performed. They did not test the effect of normalization on the original information content and on delta radiomics. Among the most repeatable features, they found kurtosis, skewness and GLCM Idm and correlation, which resulted reproducible also in our study.

Other works have reported reproducibility results on MRI radiomics. Fiset et al.[25] studied the repeatability and reproducibility of 1761 radiomic features calculated on T2w-MRI in patients affected by cervical cancers in three conditions: (a) repeatability via test–retest; (b) reproducibility between diagnostic MRI and simulation MRI; (c) reproducibility in inter-observer setting. It was reported that T2w-MRI features were highly unstable and only the inter-observer setting has highlighted high reproducibility. Interestingly, some of the features that were found reliable in Fiset et al., have also been highlighted here as highly reproducible with respect to the normalization step. In particular, the GLCM correlation, the Gray-Level Non-Uniformity calculated from GLSZM, GLDM and GLRLM, the GLDM Dependence Non-Uniformity, the GLRLM Run-Length Non-Uniformity and the NGTDM Coarseness are common in both studies. Some of these features have been also reported as predictive of treatment response in esophageal cancer,[36] able in differentiating prostate cancer aggressiveness[37] or able in characterizing structural modifications induced by RT in normal tissues.[23] Other studies focused on the reproducibility of T2w-MRI radiomic features in the prostate using different scanners,[20,24] and, among the considered features families, GLCM was the most reproducible and with higher ability in discriminating tumor and nontumor regions. Duron et al.[26] found that the method used for image discretization has a major impact on the stability of MRI features and that the adoption of a fixed bin-width is preferable to the fixed bin number. Again, among the most reproducible features, they found GLRLM Gray-Level Non-Uniformity, GLDM Gray-Level Non-Uniformity and GLDM Dependence Non-Uniformity.

Additional considerations should be given about features that are often reported as highly reproducible (e.g. Gray-Level-Non-Uniformity, GLDM Dependence-Non-Uniformity, GLRLM Run-Length-Non-Uniformity and NGTDM Coarseness). In fact, the correlation analysis with volume has highlighted that some of them intrinsically embedded an information related to the ROI size. For this reason, it was not surprising that they are found as highly reliable, especially when volume is not modified in the considered conditions (as in the present study) or when the inter-observer agreement is high (Dice > 0.9). This is in line with recent publications that highlighted some vulnerabilities in the radiomic signature development, related to the risk of including features that are mainly correlated to the volume in prediction models.[35,38]

Reproducibility of radiomic features was here assessed using the ICC metric, able to combine information about the degree of correlation and agreement between measurements.[34] This coefficient is one of the most adopted for the estimation of repeatability and reproducibility of radiomic indices, as reported in Traverso et al.[8] In fact, it has been used also in the latest work considering T2w-MRI.[25,26] However, a reference standard for reliability metric has not been established yet and, thus, other works adopted different statistical tests, such as the Concordance Correlation Coefficient (CCC),[39] t-test[40] or Spearman correlation[11] (see the review of Traverso et al.[8] for more references and statistical measures). For this reason, a quantitative comparison between studies is difficult.

This work presents some limitations. First, only the three normalization methods mainly adopted in RT studies were considered, but other normalization techniques may be used, thus enriching the stability analysis. Second, the effect of normalization among a population of different subjects was not evaluated, but only on the same subjects at different time-points. This point can affect in particular the Norm_HM method, since it is based on the reference image that, in this work was the MRI1 of the same subject, acquired with the same scanner and the same protocol. Nonetheless, the intra-subject normalization is generally adopted in longitudinal studies, which allow the estimation of tissue modifications induced by irradiation on tumor and normal tissues. Third, the normalization procedure may have a different impact on T1w images and this point should be explored deeper in dedicated analyses. Finally, the impact of bias correction and other preprocessing procedures, that may have higher effect on radiomic features calculation, was not assessed here. The focus of this work was limited to the evaluation of the intensity normalization as a necessary step in the general framework of the MRI radiomics analysis; a complete evaluation of the whole pipeline is beyond the scope of the work.

In conclusion, the normalization process impacts the reproducibility of radiomic features, both in terms of changes in the image information content (estimated on features calculated on the single image and on the *Δfeatures* calculated on longitudinal images) and in the inter-observer setting. Considering the impact of normalization both on MRI2 and on *Δfeatures*, Norm_Mean and Norm_HM provide the most reproducible features with respect to the original image and also between themselves; whereas, Norm_ROI generates the least reliable features. Conversely, Norm_ROI provides the most reproducible features in the inter-observer setting, but it may alter, at the same time, the original information. This suggests that radiomic models generated using different normalization methods may be directly compared only if the

most reproducible features are considered or if Norm_Mean and Norm_HM are adopted.

Starting from these results, future works could address a deeper evaluation of the impact of MRI normalization methods, also considering the effects on the prognostic power of the features and testing the complete pipeline of preprocessing steps, in order to give a stronger basis for the definition of a standardized protocol for the T2w-MRI radiomic analysis involving normalization.

ACKNOWLEDGMENTS

The work was partially funded by Italo Monzino Foundation and by Italian Ministry of Health (MoH) and MIUR (5 x 1000 Funds – 2016).

CONFLICT OF INTEREST

The authors have no relevant conflict of interest to disclose.

[a)]Author to whom correspondence should be addressed. Electronic mail: elisa.scalco@itb.cnr.it; Telephone: +39 02 26422216

REFERENCES

1. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
2. Al-Kadi OS, Watson D. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. *IEEE Trans Biomed Eng*. 2008;55:1822–1830.
3. Wibmer A, Hricak H, Gondo T, et al. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *Eur Radiol*. 2015;25:2840–2850.
4. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med*. 2011;52:369–378.
5. Scalco E, Fiorino C, Cattaneo GM, Sanguineti G, Rizzo G. Texture analysis for the assessment of structural changes in parotid glands induced by radiotherapy. *Radiother Oncol*. 2013;109:384–387.
6. Scalco E, Rizzo G. Texture analysis of medical images for radiotherapy applications. *Br J Radiol*. 2017;90:20160642.
7. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of multiparametric prostate MRI radiomics features. *Sci Rep*. 2019;9:9441.
8. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol*. 2018;102:1143–1158.
9. Yan J, Chu-Shern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. *J Nucl Med*. 2015;56:1667–73.
10. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [(18)F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol Imaging Biol*. 2016;18:788–795.
11. Hatt M, Majdoub M, Vallières M, et al. 18F-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*. 2015;56:38–44.
12. Parmar C, Velazquez ER, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS ONE*. 2014;9:1–8.
13. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? *Eur J Nucl Med Mol Imaging*. 2017;44:151–165.
14. Zwanenburg A, Leger S, Vallières M, Löck S.and for the I.B.S. Initiative, "Image biomarker standardisation initiative", (July), (2016).
15. Lemaitre G, Martì R, Freixenet J, Vilanova JC, Walker PM, Meriaudeau F. Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Comput Biol Med*. 2015;60:8–31.
16. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn Reson Imaging*. 2004;22:81–91.
17. Niaf E, Rouvière O, Mège-Lechevallier F, Bratan F, Lartizien C. Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. *Phys Med Biol*. 2012;57:3833–3851.
18. Xiao D-D, Yan P-F, Wang Y-X, Osman MS, Zhao H-Y. Glioblastoma and primary central nervous system lymphoma: preoperative differentiation by using MRI-based 3D texture analysis. *Clin Neurol Neurosurg*. 2018;173:84–90.
19. Scalco E, Marzi S, Sanguineti G, Farneti A, Rizzo G, Vidiri A. Evaluation of CT-based features and ADC values to assess tumor control on cervical lymph nodes treated with chemo-radiotherapy. *Phys Medica*. 2016;32:61.
20. Chirra P, Leo P, Yim M, et al. Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI. *J Med Imaging*. 2019;6:024502.
21. Ginsburg SB, Algohary A, Pahwa S, et al. Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: preliminary findings from a multi-institutional study. *J Magn Reson Imaging*. 2017;46:184–193.
22. Gnep K, Fargeas A, Gutiérrez-Carvajal RE, et al. Haralick textural features on T2 -weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer. *J Magn Reson Imaging*. 2017;45:103–117.
23. Scalco E, Rancati T, Pirovano I, et al. Texture analysis of T1-w and T2-w MR images allows a quantitative evaluation of radiation-induced changes of internal obturator muscles after radiotherapy for prostate cancer. *Med Phys*. 2018;45:1518–1528.
24. Chirra P, Bloch NB, Rastinehead A, et al. Empirical evaluation of cross-site reproducibility in radiomic features for characterizing prostate MRI. *Med Imaging 2018 Comput Diagnosis*. 2018;10575:10.
25. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiother Oncol*. 2019;135:107–114.
26. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS ONE*. 2019;14:e0213459.
27. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30:1323–1341.
28. Broggi S, Scalco E, Belli ML, et al. A comparative evaluation of 3 different free-form deformable image registration and contour propagation methods for head and neck MRI. *Technol Cancer Res Treat*. 2017;16:373–381.
29. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29:1310–1320.
30. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77:e104–e107.
31. Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging*. 2000;19:143–150.
32. Dice L. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297–302.
33. McGraw K, Wong S. *Forming inferences about some intraclass correlation coefficients*. American Psychological Association; 1996.
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
35. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother. Oncol*. 2019;130:2–9.

36. Hou Z, Li S, Ren W, Liu J, Yan J, Wan S. Radiomic analysis in T2W and SPAIR T2W MRI: predict treatment response to chemoradiotherapy in esophageal squamous cell carcinoma. *J Thorac Dis*. 2018;10:2256–2267.
37. Nketiah G, Elschot M, Kim E, et al. T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results. *Eur Radiol*. 2017;27:3050–3059.
38. Wang HYC, Donovan EM, Nisbet A, et al. The stability of imaging biomarkers in radiomics: a framework for evaluation. *Phys Med Biol*. 2019;64:165012.
39. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and prognosis of quantitative features extracted from CT images. *Transl Oncol*. 2014;7:72–87.
40. Buch K, Kuno H, Qureshi MM, Li B, Sakai O. Quantitative variations in texture analysis features dependent on MRI scanning parameters: a phantom model. *J Appl Clin Med Phys* (June). 2018;16:253–264.

SUPPORTING INFORMATION

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

Table S1. ICC values, 95% of confidence intervals and p_values (adjusted after Bonferroni correction for multiple comparisons) calculated for each radiomic feature estimated in the three ROIs on MRI2, considering the four normalization methods.

Table S2. ICC values, 95% of confidence intervals and p_values (adjusted after Bonferroni correction for multiple comparisons) calculated for each radiomic feature estimated in the three ROIs on MRI2, considering the intra- and inter-observers variability in contouring. Five normalization approaches were considered: N0_Norm, Norm_Mean, Norm_ROI, Norm_HM and Norm_FBC.

Table S3. ICC values, 95% of confidence intervals and p_values (adjusted after Bonferroni correction for multiple comparisons) calculated for each radiomic delta-feature estimated in the three ROIs, considering the three normalization methods.