

Research Article

Parameters Influencing PET Imaging Features: A Phantom Study with Irregular and Heterogeneous Synthetic Lesions

Francesca Gallivanone ¹, Matteo Interlenghi ¹, Daniela D'Ambrosio ²,
Giuseppe Trifirò,³ and Isabella Castiglioni ¹

¹Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Milan, Italy

²Medical Physics Unit, IRCCS Fondazione S. Maugeri, Pavia, Italy

³Nuclear Medicine Unit, IRCCS Fondazione S. Maugeri, Pavia, Italy

Correspondence should be addressed to Francesca Gallivanone; francesca.gallivanone@ibfm.cnr.it

Received 23 March 2018; Revised 25 July 2018; Accepted 2 August 2018; Published 10 September 2018

Academic Editor: Ali Azhdarinia

Copyright © 2018 Francesca Gallivanone et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aim. To evaluate reproducibility and stability of radiomic features as effects of the use of different volume segmentation methods and reconstruction settings. The potential of radiomics in really capturing the presence of heterogeneous tumor uptake and irregular shape was also investigated. *Materials and Methods.* An anthropomorphic phantom miming real clinical situations including synthetic lesions with irregular shape and nonuniform radiotracer uptake was used. ¹⁸F-FDG PET/CT measurements of the phantom were performed including 38 lesions of different shape, size, lesion-to-background ratio, and radiotracer uptake distribution. Different reconstruction parameters and segmentation methods were considered. COVs were calculated to quantify feature variations over the different reconstruction settings. Friedman test was applied to the values of the radiomic features obtained for the considered segmentation approaches. Two sets of test-retest measurement were acquired and the pairwise intraclass correlation coefficient was calculated. Fifty-eight morphological and statistical features were extracted from the segmented lesion volumes. A Mann–Whitney test was used to evaluate significant differences among each feature when calculated from heterogeneous versus homogeneous uptake. The significance of each radiomic feature in terms of capturing heterogeneity was evaluated also by testing correlation with gold standard indexes of heterogeneity and sphericity. *Results.* The choice of the segmentation method has a strong impact on the stability of radiomic features (less than 20% can be considered stable features). Reconstruction affects the estimate of radiomic features (only 26% are stable). Thirty-one radiomic features (53%) resulted to be reproducible, 11 of them are able to discriminate heterogeneity. Among these, we found a subset of 3 radiomic features strongly correlated with GS heterogeneity index that can be suggested as good features for retrospective evaluations.

1. Introduction

From its introduction in clinical practice, medical imaging has gained a central role in the management of a large variety of diseases. In particular, in oncology, medical imaging shows its unique property of characterizing, in vivo and noninvasively, the onset and progression of pathological processes at different stages of diseases [1].

In clinical practice, at a first level, medical images are qualitatively inspected by radiologists or nuclear medicine physicians [2]. However, such qualitative analysis presents several limitations: a certain level of subjectivity that can

cause a lack of standardization in the assessment as well as problems in some follow up evaluations.

To overcome such limitations, a great effort was focused in the recent years to develop quantitative approaches to medical image analysis. These approaches exploit the fact that digital medical images are inherently quantitative and that their quantitative values express several tissue functional characteristics, such as metabolism or proliferation [3], with a role for the onset and progression of cancer. Recent studies have been devoted to the development of automatic or semiautomatic methods for the extraction of quantitative indexes from

images to be used as imaging biomarkers of cancer disease [4, 5].

Thanks to the advancements in such image processing methods, macroscopic indexes such as the Standardized Uptake Value (SUV) for Positron Emission Tomography (PET) or the Apparent Diffusion Coefficient (ADC) for Magnetic Resonance Imaging (MRI), and measuring global functional properties of an oncological lesion, were proven effective biomarkers for diagnosis or treatment response in oncological clinical studies [6, 7]. However, novel quantitative features have been more recently explored to capture regional characteristics of a cancer lesion not always perceivable to the naked eye, such as inter- and intratumor heterogeneity, that may have an impact on the clinical outcome of different cancer phenotypes [8]. The rationale behind such advanced image features is the hypothesis that imaging *in vivo* heterogeneities of a cancer lesion is able to reflect the tumor phenotype with the advantage of a non-invasive technique. The recent literature shows the high potential of such quantitative heterogeneity features, defined “radiomics” [5], thanks to their proven abilities to be correlated with “omics” data. Radiomics refer to a large number of mathematical image descriptors extracted from the volume of an entire cancer lesion by the use of different image analytics methods, including morphological and statistical analyses [8].

Promising results published in several increasing papers proved that radiomic traits reflect tumor heterogeneity which is correlated to bad prognosis [9]. However, few studies were performed to evaluate how really those radiomic features are related to the actual shape or tissue heterogeneity of the tumor [10].

Furthermore, from a methodological point of view, one of the key problems emerging when defining image quantitative features is to assess their reproducibility, which is the closeness of the agreement between the results of successive measurements of the features carried out under the same conditions of measurement.

Moreover, different measurement conditions, such as different image reconstruction settings or lesion volume segmentation methods, can highly impact on the image feature stability, posing serious issues on the use of some image features as disease biomarkers [11–13].

The main purpose of this work is to evaluate reproducibility and stability of some radiomic features as effects of the use of different volume segmentation methods and reconstruction settings, which currently represent the more common variables in retrospective clinical oncological studies. We then assessed the significance of such radiomic features in effectively characterizing the lesion heterogeneity and shape.

These aims were pursued with the use of a realistic dataset of PET images obtaining from a thorax anthropomorphic phantom miming realistic oncological lesions with irregular shape and heterogeneous uptake of radiotracer whose GSs were known. Our work is helpful in determining the limits and the quantitative properties for clinical application of the radiomics approach with respect to the tested methods and parameters.

2. Materials and Methods

2.1. Phantom Setting and PET Data Acquisition. The anthropomorphic Alderson Thorax phantom (Radiology Support Devices, Inc.) was used to simulate man/woman thorax or breast body districts. Several synthetic lesions of irregular shape and both homogeneous and heterogeneous uptakes were realized and placed inside the thorax or the breasts of the anthropomorphic phantom within ^{18}F -FDG radioactive background. In order to simulate realistic patient PET studies, each phantom compartment was filled with a different background of ^{18}F -FDG radioactivity concentration: lungs with 0.004 MBq/cc, liver with 0.013 MBq/cc, myocardial wall with 0.023 MBq/cc, thorax with 0.006–0.007 MBq/cc, and breasts with 0.002–0.009 MBq/cc [14].

The whole procedure of preparing the phantom before PET acquisitions took about two hours. The ^{18}F -FDG radioactivity concentration used during the preparation of the phantom took into account this time frame and was recalculated based on the half life of the ^{18}F -FDG.

A strategy to produce realistic oncological lesions of irregular shape with a homogeneous or a heterogeneous uptake of ^{18}F -FDG [15] was adopted by using 3D-printed irregular shells filled with different concentrations of radioactive gels.

To obtain realistic oncological lesions with irregular shape, we defined 3D shells by segmenting the lesion volumes of different oncological lesions on ^{18}F -FDG PET/CT images of real patients. The segmented volumes were then processed in order to generate images of 3D surfaces of lesions, saved in digital files. These surfaces were then cut into two parts by image manipulation and 3D printed using a 3D printer (Renkforce RF1000 Single Extruder) equipped with plastic filaments of 3 mm diameter (Renkforce PLA300 Plastic PLA 3 mm), thus manufacturing plastic moulds of patient-derived oncological lesions.

The availability of the printed shells allowed obtaining the gold standard (GS) for the sphericity of the shells to be compared with geometrical characteristics of radiomic features as extracted from the PET images of the experimental studies performed with the phantom. In particular, for each printed mould, an index of sphericity was defined as the ratio between the surface of the sphere, with volume equivalent to actual mould volume (V_m) and the actual shell surface (S_m) of the mould.

$$S_{\text{GS}} = \frac{\pi^{1/3} (6V_m)^{2/3}}{S_m}. \quad (1)$$

This index ranges from 0 to 1, where $S_{\text{GS}} = 1$ expresses a full spherical shape.

For the PET experimental measurements, the shells were filled with a radioactive gel produced with a fast-setting, chromatic, dust-free alginate powder (phase plus, Zhermack Clinical SpA–Badia Polesine (RO), Italy) mixed with a water solution of ^{18}F -FDG [4]. Lesions with a uniform radioactive uptake were simulated using a gel preparation at a single radioactivity concentration, while gels obtained at different ^{18}F -FDG concentrations were used for lesions simulating heterogeneous uptake.

Seven experimental configurations were studied, with different radioactivity concentrations ($C_0 = 0$ MBq/cc, $C_2 = 5 * C_1$, C_1 ranging from 0.03 to 0.16 MBq/cc). We thus obtained realistic oncological lesions with heterogeneous uptake, including necrotic tissues or multifocal uptake (Figure 1).

GSs for the lesion volumes (V_{GS}) and the radioactivity concentrations were easily obtained for both homogeneous and heterogeneous lesions and the gel density and the net shell weight were found by using an analytical balance and a gamma counter (PerkinElmer 1480-011 Wizard 3[®]). In particular, in the case of heterogeneous lesions, for each filling with gels at different radioactivity concentrations, gel weights were obtained by the exact weight estimation of gel contributions at different radioactivity concentrations.

To obtain GSs for assessing the heterogeneity significance of radiomic features (as extracted from the PET studies of the phantom), two different indices of heterogeneity were considered.

The coefficient of variation of the different gels was measured as an index of heterogeneity in the radioactivity uptake, defined as the percentage ratio between the standard deviation and the mean of the radioactivity concentration within the lesion volume (COV_{GS}).

The Gini index [16] was used to quantify the impact of spatial distribution of the uptake within the shell volume (I_{G-GS}), has values from 0 in case of minimal heterogeneity to 1 in case of maximal heterogeneity, and was defined by processing digital files of shell surfaces used to print the shells. Generally, this index measures the heterogeneity of a statistical distribution in terms of the relative frequencies of the different modalities of a statistical variable. For each shell, each voxel of the corresponding shell image was considered as a statistical unit. For each voxel, the different modalities were defined as the different radioactivity concentrations used to fill the shell. Relative frequencies of each modality were calculated as the percentage of voxels occupied by each distinct radioactivity concentration.

The product of COV_{GS} and I_{G-GS} was considered as a GS index of total heterogeneity (H_{GS}), with values from 0 to 100 for lower to higher heterogeneity.

The GS for lesion-to-background ratio (L/B_{GS}) of each lesion was evaluated by measuring by the gamma counter the radioactive background of each phantom compartment where the lesions were arranged (breast containers and thorax).

^{18}F -FDG PET-CT phantom measurements were performed on a Discovery 690 PET/CT system (General Electric Medical Systems) [17]. Each PET study had an acquisition time of 180 sec for each bed position (two bed positions for each PET acquisition). Image noise was also evaluated as COV of uptake distribution.

Images were reconstructed with a standard protocol optimized for whole-body clinical oncological studies: ordered subset expectation maximization (OSEM) in 3D mode, including Point Spread Function (PSF) [18–21] and Time of Flight modelling (TOF) [19–22], 3 iterations and 18 subsets, 5 mm filter cut-off and standard z axis filter, reconstructed matrix size 256×256 , and transaxial field-of-view of 70 cm.

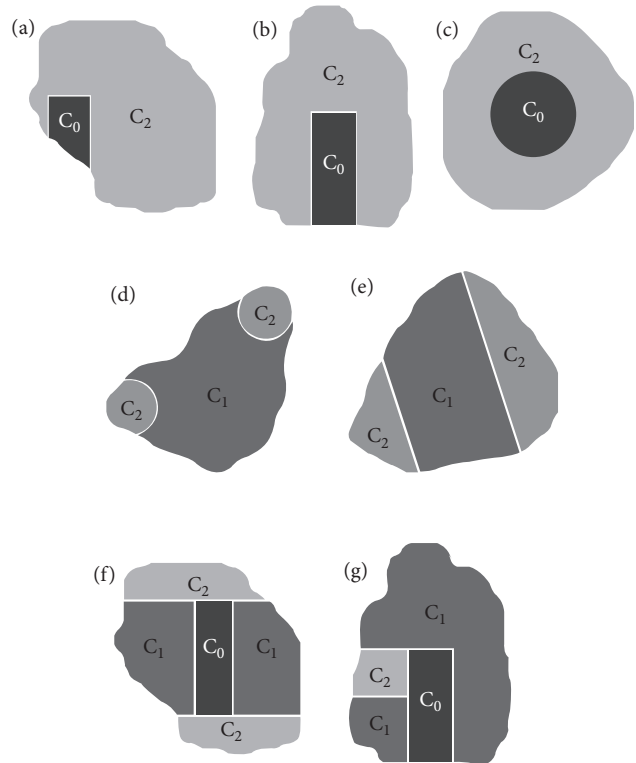


FIGURE 1: The seven different configurations to obtain lesions with different heterogeneous uptake. C_1 , C_2 , and C_0 represent areas with lower, higher, and no radioactivity concentration, respectively. (a–c) Strategies for reproducing necrotic tissue; (d, e) heterogeneous (multifocal) uptake; (f, g) heterogeneous uptake and necrotic tissue.

2.2. Image Segmentation. PET images of lesions were segmented in order to obtain the Metabolic Tumor Volume (MTV) from which extract the radiomic features. Segmentation methods used in this work included an adaptive threshold method and a fixed threshold method. The adaptive method was calibrated and validated on a variety of synthetic lesions miming real oncological lesions (i.e., with spherical and nonspherical shape and with homogenous and non-homogenous ^{18}F -FDG uptake), with an accuracy in the MTV measurement of 92% [4]. The fixed threshold method was implemented by using a cut-off of 60% from the maximum lesion uptake value. This threshold found a good compromise between a good estimate of the lesion volume and a good estimate of the lesion uptake, minimizing the possibility to include radioactivity background in the estimate [20, 21]. The two segmentation methods were implemented using Matlab and included in home-made software [22].

Since it has been shown that the use of thresholding approaches is appropriate for small lesions only when there is a good L/B , we calculated the percent error on the MTV estimate as a function of L/B for lesions with Volume GS < 10 cc, excluding from this computation the volume of necrotic regions ($C_0 = 0$) when present within a lesion.

2.3. Radiomic Feature Extraction. Radiomic imaging features were extracted from each segmented MTV as morphological

and statistical imaging features. Morphological imaging features (IF_M) were obtained starting from the shape and size characteristics of the segmented MTV [5, 8].

The statistical analysis of first-order histogram describing the distribution of voxel intensities in MTV enabled to extract first-order statistical imaging features (IF_{HIST}).

Texture analysis allowed obtaining statistical imaging features of higher orders. Images were resampled with an isotropic voxel size, considering the axial image size as resampled size. The MTV content was then resampled in 64 discrete gray-level values, and the texture analysis was performed with an in-house-developed MATLAB routine (v.2015b, MathWorks, Natick, MA, USA), largely based on a publicly available code [22]. Textural features were obtained from the analysis of the gray-level co-occurrence matrix ($IF_{TX-GLCM}$), the gray-level run-length matrix ($IF_{TX-GLRLM}$), the gray-level size zone matrix ($IF_{TX-GLSZM}$), and the neighborhood gray tone difference matrix ($IF_{TX-NGTDM}$). These matrices were obtained by the analysis of MTVs with 26-voxel connectivity, considering all possible direct connectivity with voxels in the same slice (8) and in the two adjacent slices ($9 + 9 = 18$).

2.4. Stability of Radiomic Features vs. Segmentation. In order to evaluate the impact of lesion volume segmentation (MTV) on the stability of radiomic features, the Friedman test was applied to the values of the radiomic features obtained for the two considered segmentation approaches (adaptive and fixed threshold methods, Section “Image Segmentation”).

2.5. Stability of Radiomic Features vs. Reconstruction. To study the impact of reconstruction settings on stability of radiomic features, PET images were reconstructed with reconstruction algorithms or parameters different with respect to the standard reconstruction protocols (section “Phantom setting and PET data acquisition”). For each reconstruction setting, lesions MTVs were extracted with the adaptive threshold segmentation method.

Reconstructions were performed with OSEM with or without PSF modelling and considering or omitting TOF. The impact of the matrix size of reconstructed images was also evaluated. Considering algorithm parameters, the influence of the number of iterations and subsets was assessed fixing a matrix size equal to 256×256 , because it is the most used size in clinical practice. In order to evaluate the impact of the full width at half maximum (FWHM) of Gaussian filter, matrix size was chosen such that the reconstructed voxel size is within 3.0–4.0 mm in any direction and FWHM not exceeding 7 mm, according to EANM guidelines [23].

Table 1 lists reconstruction algorithms and parameters used and their impact evaluated in this work.

For each radiomic feature, COV were calculated as average of all lesions to quantify variations over the different reconstruction settings, thus characterizing feature stability vs. reconstruction.

On the basis of COV results, radiomic features were categorized into 4 groups: stable ($COV \leq 5\%$), quite stable

($5\% < COV \leq 10\%$), poorly stable ($10\% < COV \leq 20\%$), and unstable ($COV > 20\%$).

For each feature, in order to provide representative information on its stability with respect to the different explored reconstruction settings, we considered the higher value of COVs obtained among all the reconstruction settings. A feature was considered quite stable when such COV value was found $\leq 10\%$.

2.6. Reproducibility of Radiomic Features. In order to explore reproducibility of radiomic features, a test-retest setting was used. Two sets of test-retest images were acquired approximately 30 min apart (acquisition time of 180 sec for each bed position). Lesions in the two sets of test-retest images were segmented with the adaptive threshold segmentation method.

For each feature, the pairwise intraclass correlation coefficient (ICC) was calculated [15]. Each feature with $ICC > 0.6$ in both of the two test-retest datasets (good or excellent agreement) was considered as a stable feature in test-retest setting.

2.7. Evaluation of Significance of Radiomic Features. A Mann–Whitney test was used to evaluate significant differences among each feature when calculated from heterogeneous vs. homogeneous uptake, thus measuring the potential of radiomic features in discriminating heterogeneous from homogeneous lesions.

The significance of each radiomic features in terms of capturing heterogeneity was evaluated also by testing correlation of each feature with H_{GS} .

The morphological radiomic feature “Sphericity” was evaluated in its ability to reflect geometrical characteristics as defined by S_{GS} , using a paired t -test. Heterogeneous lesions with a necrosis inside were excluded from the analysis since it was difficult to define the surface of active component of lesion.

3. Results

3.1. Phantom Setting and PET Data Acquisition. Table 2 reports GS values for each ^{18}F -FDG PET/CT acquisition.

Nine ^{18}F -FDG PET/CT acquisitions of the phantom have been performed including 38 lesions of different shape, size, radiotracer distribution, and L/B ratio in different locations of the phantom. Five different 3D-printed shells with irregular shape (A-E) were used as obtained from the PET image segmentation of real oncological lesions. Their V_{GS} ranged from 6.8 to 32.3 cc. Their sphericity index, S_{GS} , ranged from 0.49 to 0.74.

20/38 lesions were prepared with a uniform radiotracer uptake, while the remaining 18 lesions with a heterogeneous uptake. The H_{GS} for the 18 lesions with heterogeneous uptake ranged from 12.7 to 62.2, for lower to higher differences in radioactive uptake and its spatial distribution. In particular, their V_{GS} ranged from 6.4 to 29.4 cc (excluding from this computation the volume of necrotic regions when present within a lesion).

TABLE 1: Reconstruction settings.

Reconstruction algorithm	Number of iterations	Number of subsets	FWHM Gaussian filter (mm)	Reconstructed matrix size	
Impact of reconstruction algorithm	OSEM3D	3	18	5	256
	OSEM3D + PSF				
	OSEM3D + TOF				
	OSEM3D + PSF + TOF				
Impact of number of iterations		2	18	5	256
	OSEM3D + PSF + TOF	3			
Impact of number of subsets	OSEM3D + PSF + TOF	3	18	5	256
		4	24		
Impact of reconstructed matrix size	OSEM3D + PSF + TOF	3	18	5	128
					192
					256
Impact of FWHM of Gaussian filter	OSEM3D + PSF + TOF	3	18	5	192
				7	

OSEM = ordered subset expectation maximization; PSF = point spread functions; TOF = time of flight; FWHM = full width at half maximum.

TABLE 2: Summary of the 18F-FDG PET/CT acquisitions of the phantom, with gold standard values of each lesion.

Number of PET acquisition	Number of lesion acquired in PET	Shell type	V_{GS} (cc)	S_{GS}	H_{GS} (%)	V_{GS} excluding necrosis (cc)	L/B_{GS}
1	1	A	6.8	0.57	0	6.8	10
	2	B	10.5	0.62	0	10.5	10
	3	C	8.5	0.49	0	8.5	10
	4	D	12.5	0.74	0	12.5	10
2	5	A	6.8	0.57	0	6.8	10
	6	B	10.5	0.62	0	10.5	10
	7	C	8.5	0.49	0	8.5	10
	8	D	12.5	0.74	0	12.5	10
3	9	A	6.8	0.57	0	6.8	10
	10	B	10.5	0.62	0	10.5	10
	11	C	8.5	0.49	0	8.5	10
	12	D	12.5	0.74	0	12.5	10
4	13	A	6.8	0.57	0	6.8	27
	14	B	10.5	0.62	0	10.5	26
	15	C	8.5	0.49	21.1	7.4	9
	16	D	12.5	0.74	12.7	11.7	25
5	17	A	6.8	0.57	0	6.8	27
	18	B	10.5	0.62	0	10.5	26
	19	C	8.5	0.49	21.1	7.4	9
	20	D	12.5	0.74	12.7	11.7	25
6	21	A	6.8	0.57	0	6.8	27
	22	B	10.5	0.62	0	10.5	26
	23	C	8.5	0.49	21.1	7.4	9
	24	D	12.5	0.74	12.7	11.7	25
7	25	A	6.8	0.57	0	6.8	12
	26	B	10.5	0.62	0	10.5	11
	27	C	8.5	0.49	21.1	7.4	4
	28	D	12.5	0.74	12.7	11.7	11
8	29	A	6.8	0.57	14.9	6.4	18
	30	B	10.5	0.62	26.2	10.5	10
	31	C	8.5	0.49	24.8	7.6	7
	32	D	12.5	0.74	62.2	8.7	9
	33	E	32.3	0.73	16.3	29.4	25
9	34	A	6.8	0.57	14.9	6.4	12
	35	B	10.5	0.62	26.2	10.5	7
	36	C	8.5	0.49	24.8	7.6	5
	37	D	12.5	0.74	62.2	8.7	6
	38	E	32.3	0.73	16.3	29.4	16

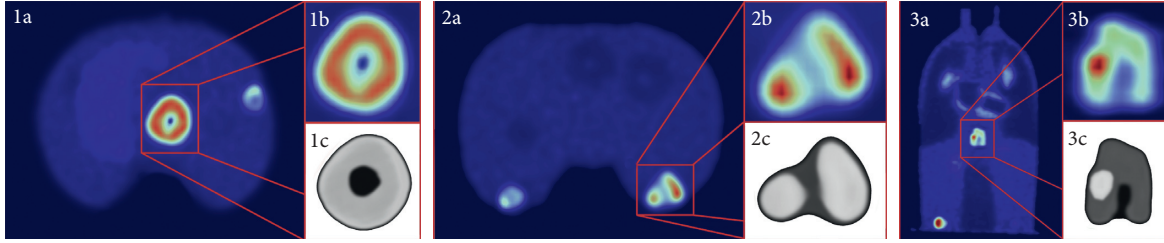


FIGURE 2: Examples of PET images of heterogeneous lesions (a-b, d-e, g-h), with 3D renders of lesions (c, f, i). (1) $V_{GS} = 32.3$ cc, $S_{GS} = 0.73$, $H_{GS} = 0.16$, $L/B_{GS} = 25$; (2) $V_{GS} = 10.5$ cc, $S_{GS} = 0.62$, $H_{GS} = 0.26$, $L/B_{GS} = 10$; (3) $V_{GS} = 8.6$ cc, $S_{GS} = 0.49$, $H_{GS} = 0.25$, $L/B_{GS} = 7$.

Explored L/B_{GS} ranged from 4 to 27.

L/B measured (L/B_m) on PET images by using a validated PET quantification technique [24] ranged from 4.3 to 27.6.

Image noise of each PET acquisition was evaluated as COV in uptake distribution inside a large region of the liver. The mean COV calculated on the 9 PET acquisition is $<8\%$.

Figure 2 shows 3 PET images of 3 representative heterogeneous lesions including necrosis and bifocal uptake.

3.2. Image Segmentation. Table 3 shows the mean percent error on the estimate of MTV for the two considered segmentation methods (adaptive and fixed threshold methods), for small lesions ($V_{GS} \leq 10$ cc, excluding from V_{GS} computation the volume of necrotic regions when present within a lesion), grouped as a function of L/B_m ($L/B_m \leq 5$, $5 < L/B_m \leq 10$, $10 < L/B_m \leq 15$).

The adaptive threshold method presents good results at higher L/B_m (mean percent error $<20\%$ for lesion with $L/B_m > 5$). The percent error of the fixed threshold method is larger (absolute mean percent error $>30\%$), irrespectively from L/B_m .

Generally, results show the tendency of the adaptive threshold method to overestimate the volume, while the fixed threshold segmentation method always underestimates it. However, the selection of the optimal segmentation method was not the purpose of this paper.

3.3. Radiomic Feature Extraction. Table 4 shows the radiomic features extracted from the segmented MTVs of each lesion.

In particular, five morphological features were extracted characterizing the shape and size of each lesion [8], 13 statistical features were extracted from the analysis of the intensity histogram of lesions, and 40 statistical features were obtained by the textural analysis (9 statistical features for GLCM, 13 for GLRLM, 13 for GLSZM, and 5 for NGTDM), for a total of 58 radiomic features.

3.4. Stability of Radiomic Features vs. Segmentation. Figure 3 shows the results of Friedman test for each of the 58 radiomic features.

By comparing the values of each feature extracted from the MTV as derived from the two segmentation approaches, it was found that many features have a large variability with respect to the applied segmentation method; thus the choice

TABLE 3: Mean percent error on the estimate of MTV of small lesions as a function of L/B_m , for the adaptive and fixed threshold segmentation methods.

L/B_m	Adaptive threshold mean percent error (%)	Fixed threshold mean percent error (%)
$L/B_m \leq 5$	27 ± 9	-33 ± 13
$5 < L/B_m \leq 10$	16 ± 30	-35 ± 26
$10 < L/B_m \leq 15$	17 ± 16	-31 ± 25

of the segmentation method have a strong impact on the stability of radiomic features.

In particular, results obtained on the whole datasets of both uniform and nonuniform lesions showed that less than 20% (11/58) of radiomic features can be considered full stable with respect the two considered segmentation methods.

In Figure 4, the results of Friedman test are presented only for the datasets of uniform lesions.

As expected, a larger number of radiomic features resulted stable (41%, 24/58).

3.5. Stability of Radiomic Features vs. Reconstruction. Results obtained considering variations of reconstruction parameter (i.e., reconstruction type, matrix size, FWHM of Gaussian filter, number of iterations, and number of subsets) are summarized in Figure 5, grouped with respect to COVs. Reconstruction strongly affects the estimate of radiomic features: 52% (30/58) of features showed a large variability with respect to a different reconstruction setting ($COV > 20\%$). Only 26% (15/58) showed a small variability among all the reconstruction setting variations ($COV \leq 10\%$). Most features are severely affected by variation in the dimension of reconstructed matrix. Features derived from the analysis of the intensity histogram (IF_{HIST}) are more influenced from reconstruction variation than the other features.

3.6. Reproducibility of Radiomic Features. Thirty-one of the 58 radiomic features (53%) resulted stable in the test-retest datasets ($ICC \geq 0.6$), as reported in Figure 6.

3.7. Evaluation of Significance of Imaging Features. Results from Mann-Whitney test showed that 24/58 (41%) of radiomic features have significantly different values in case of lesions

TABLE 4: The radiomic features considered in the work.

Feature name	Feature group	
MTV	IF _M	
Surface		
Spherical disproportion		
Sphericity		
Surface-volume ratio (SV)		
Maximum		
Minimum	IF _{HIST}	
Mean		
Median		
Mean absolute deviation (MAD)		
Root mean square (RMS)		
Energy		
Entropy		
Kurtosis		
Skewness		
Standard deviation		
Uniformity		
Variance		
Energy		IF _{TX-GLCM}
Contrast		
Entropy		
Homogeneity		
Correlation		
SumAverage		
Variance		
Dissimilarity		
Autocorrelation		
Short run emphasis (SRE)		
Long run emphasis (LRE)	IF _{TX-GLRLM}	
Gray-level nonuniformity (GLN)		
Run-length nonuniformity (RLN)		
Run percentage (RP)		
Low gray-level run emphasis (LGRE)		
High gray-level run emphasis (HGRE)		
Short run low gray-level emphasis (SRLGE)		
Short run high gray-level emphasis (SRHGE)		
Long run low gray-level emphasis (LRLGE)		
Long run high gray-level emphasis (LRHGE)		
Gray-level variance (GLV)	IF _{TX-GLSZM}	
Run-length variance (RLV)		
Small zone emphasis (SZE)		
Large zone emphasis (LZE)		
Gray-level nonuniformity (GLN)		
Zone-size nonuniformity (ZSN)		
Zone percentage (ZP)		
Low gray-level zone emphasis (LGZE)		
High gray-level zone emphasis (HGZE)		
Small zone low gray-level emphasis (SZLGE)		
Small zone high gray-level emphasis (SZHGE)	IF _{TX-NGTDM}	
Large zone low gray-level emphasis (LZLGE)		
Large zone high gray-level emphasis (LZHGE)		
Gray-level variance (GLV)		
Zone-size variance (ZSV)		
Coarseness		
Contrast		
Busyness		
Complexity		
Strength		

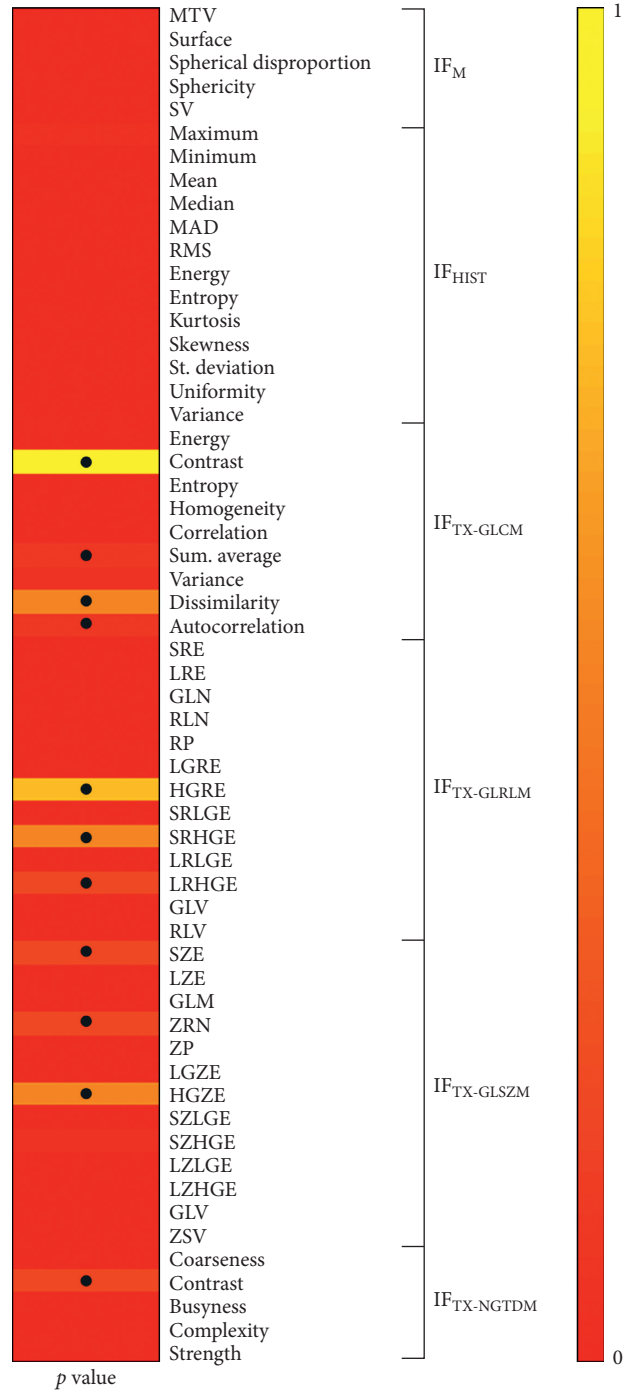


FIGURE 3: Stability of radiomic features on different segmentations. Friedman test results (p value), • indicates p value ≥ 0.05 .

with uniform versus nonuniform uptake (p value < 0.05) (Figure 7).

As shown in Figure 8, the correlation analysis performed between radiomic features and H_{GS} shows that 16 of them (28%) resulted significantly correlated (p value < 0.05).

Paired t -test on Gold Standard S_{GS} showed that the morphological IF “Sphericity” is able to reflect actual deviation from spherical shape, both in case of lesions with

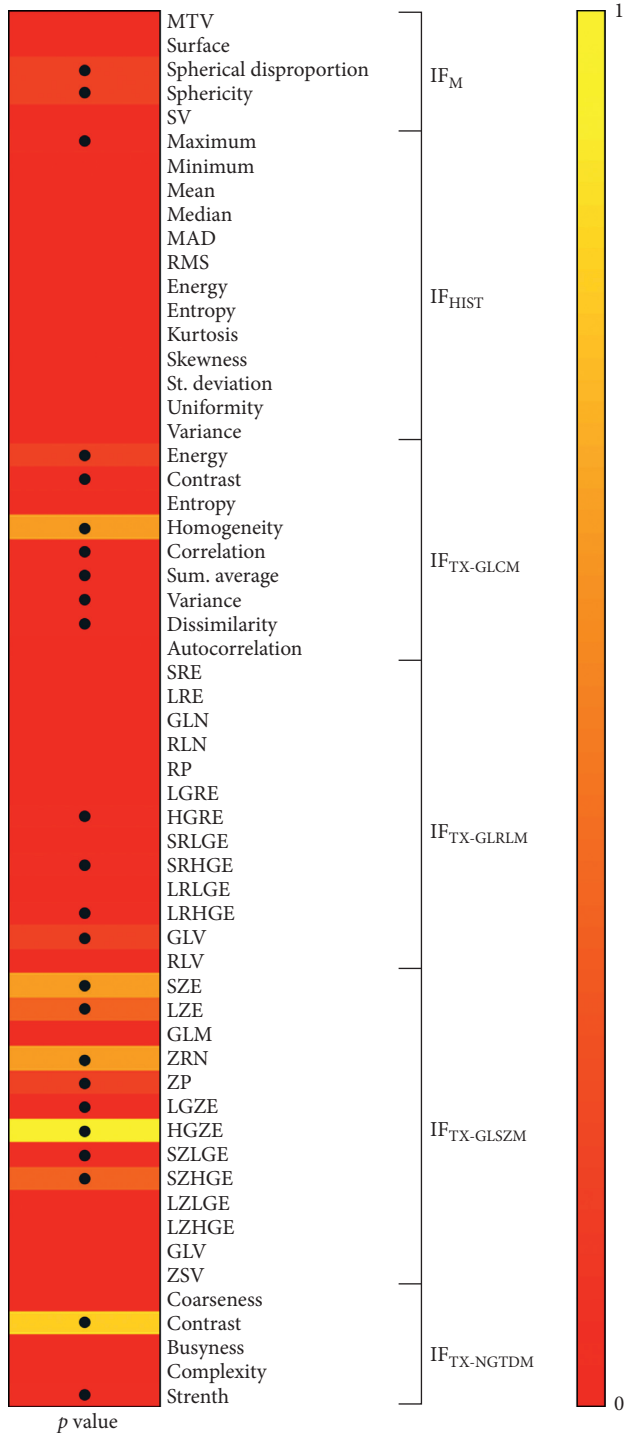


FIGURE 4: Uniform lesions. Stability of radiomic features on different segmentations. Friedman test results (p value), • indicates p value ≥ 0.05 .

uniform uptake and in case of lesions with nonuniform uptake.

4. Discussion and Conclusions

Despite the potential proven impact of radiomics, scientific evidences suggest that radiomic features extracted from PET

images of cancer lesions may have a large variability depending in particular on the different reconstruction settings and segmentation strategies used prior the radiomic analysis [25, 26]. Other studies report that radiomic features can be affected by a lack of inpatient reproducibility [27–29]. Furthermore, even if the radiomic hypothesis is that such features properly reflect tumor heterogeneity as measured on medical images (including PET images), no clear indications exist concerning which features can better reflect heterogeneous tumor uptake and which type and level of heterogeneity can be captured and quantified through PET.

Some published works were devoted to assess inpatient reproducibility or features stability with respect to both segmentation or reconstruction settings [11–13]. However, most of these works were performed on patients data or phantom data acquired in ideal conditions (e.g., in spherical synthetic lesions) [12, 30], or on simulated data, where it is difficult to reproduce noise and artifacts contributions, as in real clinical situations [10, 31–33]. Most of these works lack on details about the methodology adopted behind image processing and often evaluate only one aspect of the feature variability [11, 12]. Few works were able to deal with the interpretation of features with respect to tumor heterogeneity and with respect to which type and level of heterogeneity can be quantified through PET [10].

Consistently with other published studies [11, 25, 26, 34, 35], we found that different radiomic PET traits are influenced by the lesion volume delineation method (less than 20% features can be considered stable for the two methods assessed in this work), our results confirming that the choice of segmentation method severely affects the quantitative estimate of radiomic features. Such concern regards in particular the possibility to compare results obtained by radiomic studies in which different segmentation methods were used, as occurring in some multicenter evaluations proposing databases of reconstructed images with lesions segmented by operators and annotated on the archived images. To avoid bias in the results, our findings suggest the use of the same segmentation method to be applied with a standardized image processing procedure, possibly with the use of the same software tools, after image collection and archive.

The fixed threshold approach is widely used in the literature [12], and for this reason, we have used this method to segment the lesion volume. However, notwithstanding this was not the purpose of our paper, our results are in agreement with many published studies, showing that the accuracy in the definition of lesion volume is low [13]. In particular, the method risks to largely overestimate the lesion volume (large and negative percent errors), and this, in addition to a poor accuracy, can cause severe problems in the estimate of radiomic features due to the possible inclusion of signal uptake not linked to cancer tissue but to the surrounding tissues. The adaptive threshold approach seems to be more suitable for radiomic analysis since it is more conservative with respect to the estimate of cancer tissue volume.

We found, in agreement with previous reports [11] that also reconstruction largely affects the estimate of radiomic features (only 26% are stable with respect to different

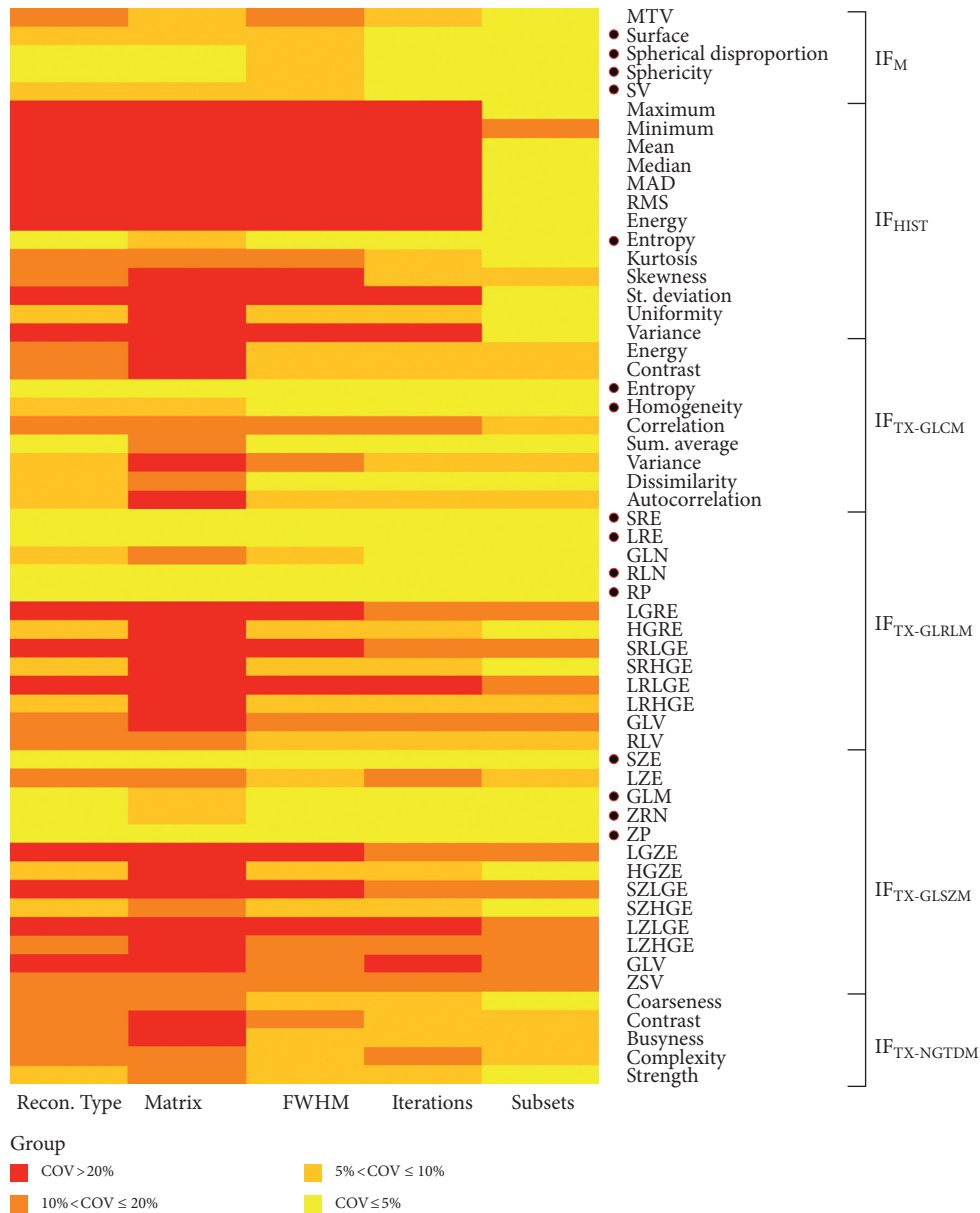


FIGURE 5: Stability of radiomic features on different reconstruction settings. COV results. • indicates COV ≤ 10%.

settings). In particular, the more impacting parameter is the reconstructed matrix size that leads to variations in the estimate of many radiomic features greater than 20%. To avoid bias in the results, similarly to what suggested as segmentation strategy, our findings would suggest the use of the same reconstruction method to be applied with the same image reconstruction setting, possibly with the use of the same reconstruction tool, after the acquisition and archive of raw data. Unfortunately, while a high level of standardization is possible for the segmentation step, this is difficult for the reconstruction step, for different reasons. Clinical images are reconstructed with different reconstruction algorithms depending from the physical characteristics of the imaging systems/models and various reconstruction settings defined in different imaging centers. This limits the possibility to standardize the reconstruction protocol in prospective

clinical studies or to have access to retrospective studies with the same reconstruction protocols used. The only way to perform valid radiomic studies should be to collect raw data from prospective patient studies and then to reconstruct them with the same reconstruction tool, but this is a very challenge task to be accomplished, in particular for the huge amount of resources (in space and time) required. Orlhac et al. [10] very recently have proposed a method based on the ComBat approach [36] used in genomics analysis that seems effective in standardizing radiomic features measured from PET images obtained using different imaging protocols.

Inpatient reproducibility can be a serious concern, but it could be properly managed. A good number of features (31) resulted reproducible from our results of test-retest setting, suggesting to consider this subset for further radiomic analysis. Among these reproducible features we

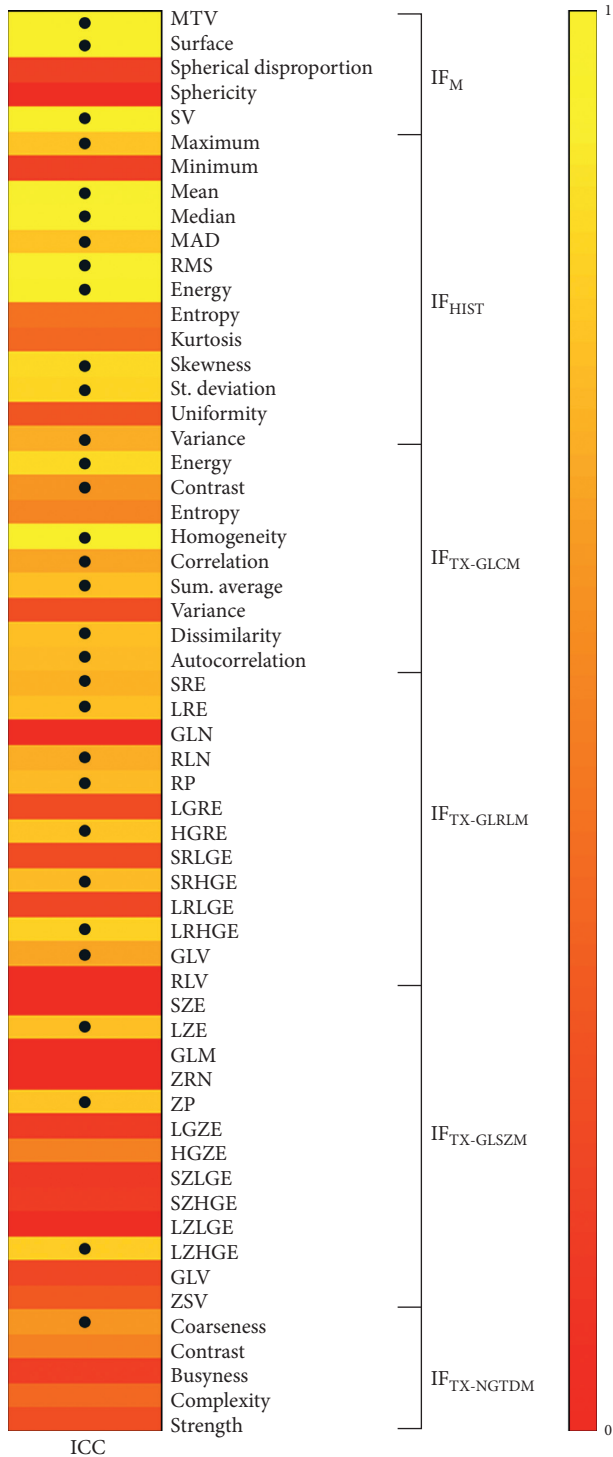


FIGURE 6: Reproducibility of radiomic features on test-retest datasets. ICC results. • indicates ICC ≥ 0.6 .

found most of morphological and histogram-derived features considered in this work, and some textural features from the gray-level co-occurrence matrix and gray-level run-length matrix.

Eleven of the 31 features were found also able to discriminate heterogeneous from uniform radioactivity uptake (p value from Mann-Whitney test < 0.05).

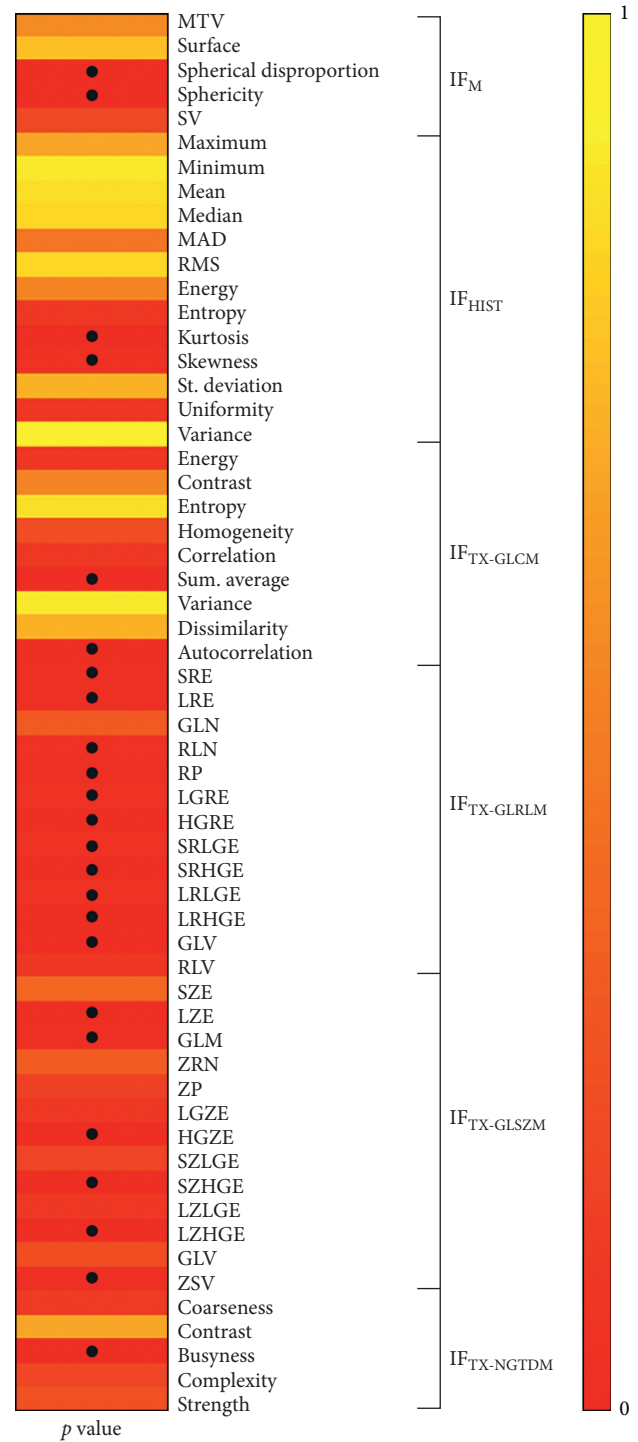


FIGURE 7: Mann-Whitney test results (p value), • indicates p value < 0.05 .

Furthermore, interesting results were obtained when comparing radiomic features with respect to gold standard indexes of heterogeneity and sphericity. Considering the uptake heterogeneity, we found 3 reproducible features (run-length-nonuniformity, run percentage, and large zone emphasis) among the 11 found above, which are also proven able to reflect the heterogeneity in the PET uptake (strongly correlated with the gold standard heterogeneity index).

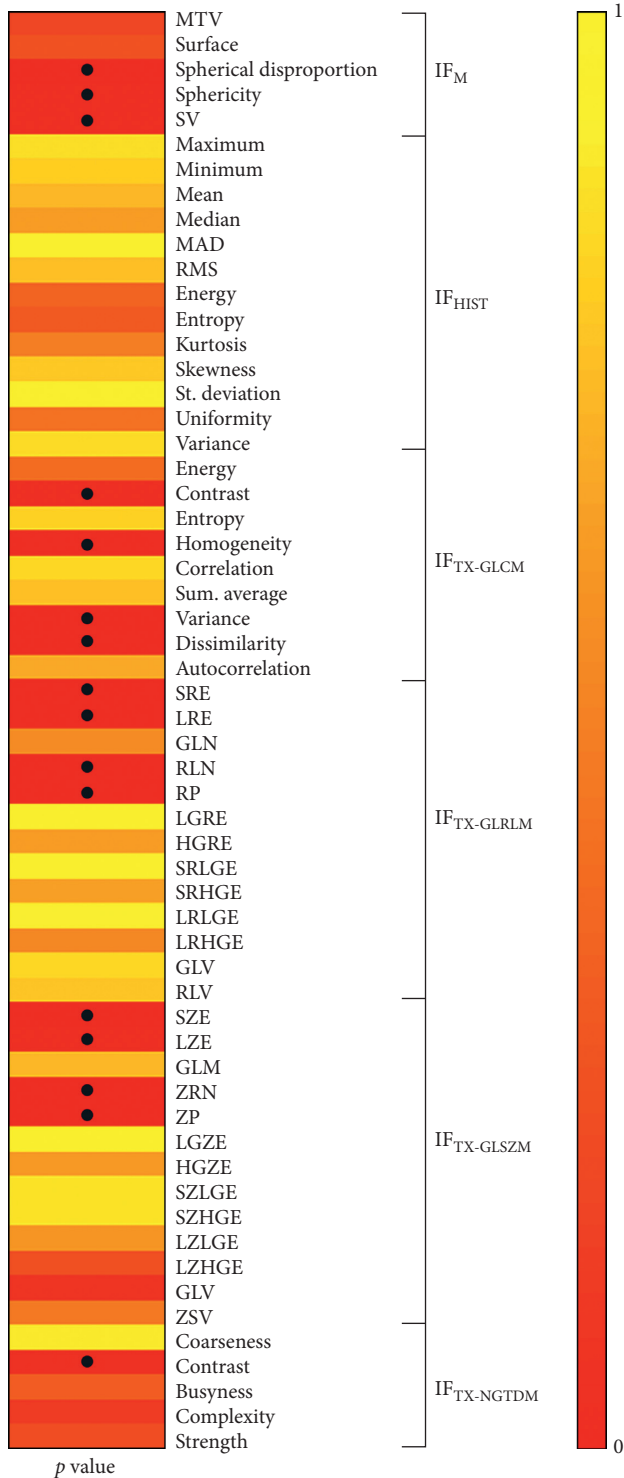


FIGURE 8: Results of correlation analysis between radiomic features and H_{GS} (p value), • indicates p value < 0.05.

These findings suggest that the 3 features can be considered as first choice when testing the hypothesis that PET heterogeneity could reflect real tumor heterogeneity.

In conclusions, in this work, we showed some limits and quantitative properties of the radiomics approach (with respect of the tested methods and parameters) that should be

overcome for a clinical translation of radiomics. Considering our findings, we suggest an optimal strategy for radiomic bias-free analysis to archive all raw data of PET acquisitions collected for a clinical study, to be then reconstructed and segmented by standardized reconstruction and segmentation protocols. We found a subset of thirty features that could be preferred for reproducible radiomic PET studies; 3 of them seeming particularly suitable for capturing tumor heterogeneity. However, our results need to be confirmed by other more extensive studies and cannot be exactly transferred to real or more complex clinical conditions.

Data Availability

An image set of our original anthropomorphic phantom is available to researchers, after registration, at http://inlab.ibfm.cnr.it/inlab/research_data.php.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the CNR Research Project “Aging: Molecular and Technological Innovations for Improving the Health of the Elderly” (no. DSB.AD009.001; activity no. DSB.AD009.001.043).

References

- [1] B. Sharma, A. Martin, S. Stanway, S. R. Johnston, and A. Constantinidou, “Imaging in oncology—over a century of advances,” *Nature Reviews Clinical Oncology*, vol. 9, no. 12, pp. 728–737, 2012.
- [2] M. E. Juweid, S. Stroobants, O. S. Hoekstra et al., “Use of positron emission tomography for response assessment of lymphoma: consensus of the imaging subcommittee of international harmonization project in lymphoma,” *Journal of Clinical Oncology*, vol. 25, no. 5, pp. 571–578, 2007.
- [3] M. D. Farwell, D. A. Pryma, and D. A. Mankoff, “PET/CT imaging in cancer: current applications and future directions,” *Cancer*, vol. 120, no. 22, pp. 3433–3445, 2014.
- [4] F. Gallivanone, M. Interlenghi, C. Canervari, and I. Castiglioni, “A fully automatic, threshold-based segmentation method for the estimation of the metabolic tumor volume from PET images: validation on 3D printed anthropomorphic oncological lesions,” *Journal of Instrumentation*, vol. 11, no. 1, article C01022, 2016.
- [5] F. Gallivanone, M. M. Panzeri, C. Canevari et al., “Biomarkers in vivo molecular imaging of breast cancer: pretreatment ^{18}F -FDG PET predicts patient prognosis, and pretreatment DWI-MR predicts response to neoadjuvant chemotherapy,” *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 30, no. 4, pp. 359–373, 2017.
- [6] F. Gallivanone, C. Canevari, I. Sassi et al., “Partial volume corrected ^{18}F -FDG PET mean standardized uptake value correlates with prognostic factors in breast cancer,” *Quarterly Journal of Nuclear Medicine and Molecular Imaging*, vol. 58, no. 4, pp. 424–439, 2014.
- [7] G. Decker, P. Mürtz, J. Gieseke et al., “Intensity-modulated radiotherapy of the prostate: dynamic ADC monitoring by

- DWI at 3.0 T,” *Radiotherapy and Oncology*, vol. 113, no. 1, pp. 115–120, 2014.
- [8] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar et al., “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature Communications*, vol. 5, no. 1, p. 4006, 2014.
- [9] M. Robertson-Tessi, R. J. Gillies, R. A. Gatenby, and A. R. Anderson, “Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes,” *Cancer Research*, vol. 75, no. 8, pp. 1567–1579, 2015.
- [10] F. Orlhac, C. Nioche, M. Soussan, and I. Buvat, “Understanding changes in tumor textural indices in PET: a comparison between visual assessment and index values in simulated and patient data,” *Journal of Nuclear Medicine*, vol. 58, no. 3, pp. 387–392, 2017.
- [11] J. Yan, J. L. Chu-Shern, H. Y. Loi et al., “Impact of image reconstruction settings on texture features in 18F-FDG PET,” *Journal of Nuclear Medicine*, vol. 56, no. 11, pp. 1667–1673, 2015.
- [12] I. Shiri, A. Rahmim, P. Ghaffarian, P. Geramifar, H. Abdollahi, and A. Bitarafan-Rajabi, “The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies,” *European Radiology*, vol. 27, no. 11, pp. 4498–4509, 2017.
- [13] F. Orlhac, M. Soussan, J. A. Maisonneuve, G. A. Garcia, B. Vanderlinden, and I. Buvat, “Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis,” *Journal of Nuclear Medicine*, vol. 55, no. 3, pp. 414–422, 2014.
- [14] P. Ghafarian, S. M. Aghamiri, M. R. Ay et al., “Is metal artefact reduction mandatory in cardiac PET/CT imaging in the presence of pacemaker and implantable cardioverter defibrillator leads?,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 38, no. 2, pp. 252–262, 2011.
- [15] F. Gallivanone, M. Interlenghi, D. D’Ambrosio et al., “An anthropomorphic phantom for advanced image processing of realistic 18F-FDG PET-CT oncological studies,” in *Proceedings of 2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, pp. 1–7, Strasbourg, France, October 2016.
- [16] C. Gini, “Measurement of inequality and incomes,” *Economic Journal*, vol. 31, no. 121, pp. 124–126, 1921.
- [17] V. Bettinardi, L. Presotto, E. Rapisarda, M. Picchio, L. Gianolli, and M. C. Gilardi, “Physical performance of the new hybrid PET/CT Discovery-690,” *Medical Physics*, vol. 38, no. 10, pp. 5394–5411, 2011.
- [18] S. Tong, A. M. Alessio, and P. E. Kinahan, “Image reconstruction for PET/CT scanners: past achievements and future challenges,” *Imaging in Medicine*, vol. 2, no. 5, pp. 529–545, 2010.
- [19] S. Surti and J. S. Karp, “Advances in time-of-flight PET,” *Physica Medica*, vol. 32, no. 1, pp. 12–22, 2016.
- [20] M. Soret, S. L. Bacharach, and I. Buvat, “Partial-volume effect in PET tumor imaging,” *Journal of Nuclear Medicine*, vol. 48, no. 6, pp. 932–945, 2007.
- [21] F. Gallivanone, C. Canevari, L. Gianolli et al., “A partial volume effect correction tailored for 18F-FDG-PET oncological studies,” *BioMed Research International*, vol. 2013, Article ID 780458, 12 pages, 2013.
- [22] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, “A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities,” *Physics in Medicine and Biology*, vol. 60, no. 14, pp. 5471–5496, 2015.
- [23] R. Boellaard, R. Delgado-Bolton, W. J. Oyen et al., “FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 42, no. 2, pp. 328–354, 2015.
- [24] F. Gallivanone, A. Stefano, E. Grosso et al., “PVE Correction in PET-CT Whole-body oncological studies from PVE-affected images,” *IEEE Transactions on Nuclear Science*, vol. 58, no. 3, pp. 736–747, 2011.
- [25] F. Yang, N. Dogan, R. Stoyanova, and J. C. Ford, “Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: a simulation study utilizing ground truth,” *Physica Medica*, vol. 50, pp. 26–36, 2018.
- [26] A. Midya, J. Chakraborty, M. Gönen, R. K. G. Do, and A. L. Simpson, “Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility,” *Journal of Medical Imaging*, vol. 5, no. 1, article 011020, 2018.
- [27] O. S. Al-Kadi, “Assessment of texture measures susceptibility to noise in conventional and contrast enhanced computed tomography lung tumour images,” *Computerized Medical Imaging and Graphics*, vol. 34, no. 6, pp. 494–503, 2010.
- [28] Y. Balagurunathan, Y. Gu, H. Wang et al., “Reproducibility and prognosis of quantitative features extracted from CT images,” *Translational Oncology*, vol. 7, no. 1, pp. 72–87, 2014.
- [29] Y. Balagurunathan, V. Kumar, Y. Gu et al., “Test-retest reproducibility analysis of lung CT image features,” *Journal of Digital Imaging*, vol. 27, no. 6, pp. 805–823, 2014.
- [30] L. Presotto, V. Bettinardi, E. De Bernardib et al., “PET textural features stability and pattern discrimination power for radiomics analysis: an “ad-hoc” phantoms study,” *Physica Medica*, vol. 50, pp. 66–74, 2018.
- [31] I. Buvat and I. Castiglioni, “Monte Carlo simulations in SPET and PET,” *Quarterly Journal of Nuclear Medicine*, vol. 46, no. 1, pp. 48–61, 2002.
- [32] I. Buvat, I. Castiglioni, J. Feuardent, and M. C. Gilardi, “Unified description and validation of Monte Carlo simulations in PET,” *Physics in Medicine and Biology*, vol. 50, no. 2, pp. 329–346, 2005.
- [33] I. Castiglioni, O. Cremonesi, M. C. Gilardi et al., “A Monte Carlo model of noise component in 3D PET,” *IEEE Transactions on Nuclear Science*, vol. 49, no. 5, pp. 2297–2303, 2002.
- [34] P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj, “Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters,” *Acta Oncologica*, vol. 49, no. 7, pp. 1012–1016, 2010.
- [35] T. Perrin, A. Midya, R. Yamashita et al., “Short-term reproducibility of radiomic features in liver parenchyma and liver malignancies on contrast-enhanced CT imaging,” *Abdominal Radiology*, pp. 1–8, 2018.
- [36] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.