Optimized Feature Extraction for Radiomics Analysis of ^{18}F-FDG PET Imaging

Laszlo Papp[1], Ivo Rausch[1], Marko Grahovac[2], Marcus Hacker[2], and Thomas Beyer[1]

[1]QIMP Team, Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria; and [2]Division of Nuclear Medicine, Department of Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Vienna, Austria

Radiomics analysis of ^{18}F-FDG PET/CT images promises well for an improved in vivo disease characterization. To date, several studies have reported significant variations in textural features due to differences in patient preparation, imaging protocols, lesion delineation, and feature extraction. Our objective was to study variations in features before a radiomics analysis of ^{18}F-FDG PET data and to identify those feature extraction and imaging protocol parameters that minimize radiomic feature variations across PET imaging systems. **Methods:** A whole-body National Electrical Manufacturers Association image-quality phantom was imaged with 13 PET/CT systems at 12 different sites following local protocols. We selected 37 radiomic features related to the 4 largest spheres (17–37 mm) in the phantom. On the basis of a combined analysis of voxel size, bin size, and lesion volume changes, feature and imaging system ranks were established. A 1-way ANOVA was performed over voxel size, bin size, and lesion volume subgroups to identify the dependency and the trend change in feature variations across these parameters. **Results:** Feature ranking revealed that the gray-level cooccurrence matrix and shape features are the least sensitive to PET imaging system variations. Imaging system ranking illustrated that the use of point-spread function, small voxel sizes, and narrow gaussian postfiltering helped minimize feature variations. ANOVA subgroup analysis indicated that variations in each of the 37 features and for a given voxel size and bin size can be minimized. **Conclusion:** Our results provide guidance to selecting optimized features from ^{18}F-FDG PET/CT studies. We were able to demonstrate that feature variations can be minimized for selected image parameters and imaging systems. These results can help imaging specialists and feature engineers in increasing the quality of future radiomics studies involving PET/CT.

Key Words: ^{18}F-FDG PET/CT; radiomics; feature extraction

J Nucl Med 2019; 60:864–872
DOI: 10.2967/jnumed.118.217612

Radiomics refers to the process of extracting and analyzing in vivo features from medical images for disease characterization (*1*). The radiomics approach was originally conceived for morphologic images only (*2,3*) but recently was adopted also for the analysis of ^{18}F-FDG PET/CT images, with promising results in various patient cohorts (*4–9*). It has been shown that several features, particularly textural indices, used in the radiomics approach are affected by, for example, variations in biologic factors (*10*), imaging and reconstruction protocols (*11,12*), delineation approaches (*13–15*), or feature extraction methods (*12,16–18*).

Feature variations challenge the reproducibility of radiomics assessments; therefore, standardized protocols related to patient preparation, imaging, and feature engineering are needed (*18,19*). In this context, Vallières et al. recently pointed to the importance of standardized image processing and feature computation for better addressing the "statistical quality of radiomics analyses" (*20*). Although individual feature computations in light of variable image resolutions (*12,21–26*) or bin sizes (*27–32*) have been investigated, optimized feature extraction after the combined analysis of voxel size, bin size, and lesion volume changes has not yet been reported. Instead, the choice of protocol parameters is still driven largely by the wish to maximize individual predictive performance. This is in contrast to the need for standards in radiomics analysis at the level of individual feature extraction parameters.

Our hypothesis was that feature extraction can be optimized through the analysis of ^{18}F-FDG PET image features derived from multiple scans of a standard phantom. We used multicenter data to provide a general solution to optimize feature extraction applicable mono- or multicentrically. We performed an in-depth analysis of features regarding voxel size, bin size, and lesion volume changes to support feature extraction optimization.

MATERIALS AND METHODS

Phantom Acquisition

The data used for this study were acquired in the context of a multicenter study across 12 PET imaging centers involving 13 imaging systems in Austria (*33*). A National Electrical Manufacturers Association (NEMA) image-quality phantom was filled with a background activity concentration

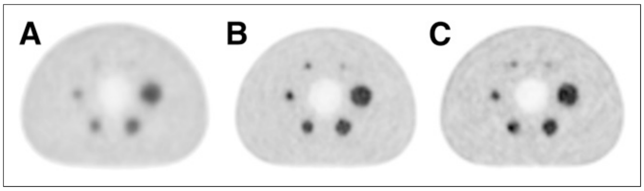

FIGURE 1. Central axial slices through reconstructed PET images of NEMA image-quality phantom acquired from 3 of the involved 13 PET/CT imaging systems (Table 1): PCS$_3$ (A), PCS$_{13}$ (B), and PCS$_8$ (C). Acquisitions followed local clinical standard protocols as part of previous study (*33*). PET image planes demonstrate typical variations in appearance of lesions and backgrounds.

Received Jul. 13, 2018; revision accepted Oct. 26, 2018.

For correspondence or reprints contact: Laszlo Papp, QIMP Team, Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria.

E-mail: laszlo.papp@meduniwien.ac.at

Published online Nov. 2, 2018.

TABLE 1

Image Acquisition and Reconstruction Protocols for NEMA Image-Quality Phantom Studies Using 13 PET/CT Systems (*33*)

System	Algorithm	PSF	TOF	Iterations	Subsets	Filter	FWHM	Voxel size (mm)	Time/bed position (min)	BckVar (%)
PCS_1	Blob-OS-TF	NA	Yes	NA	NA	Un	NA	4.00	1:15	2.80
PCS_2	OSEM	No	No	4	8	Ga	5	4.06	3:00	2.50
PCS_3	OSEM	No	No	2	8	Ga	5	5.31	2:00	2.97
PCS_4	LOR-RAMLA	No	No	NA	NA	Un	NA	4.00	1:30	4.51
PCS_5	TrueX	Yes	No	3	21	Ga	2	4.07	2:00	2.72
PCS_6	TrueX	Yes	No	4	21	None	NA	4.07	3:00	3.19
PCS_7	TrueX	Yes	No	4	21	None	NA	4.06	3:00	3.21
PCS_8	TrueX	Yes	No	3	21	Ga	2	4.07	2:00	3.22
PCS_9	TrueX (HD PET)	Yes	No	3	21	Ga	2	3.18	2:00	3.07
PCS_{10}	VUE Point	No	No	2	21	Ga	6	5.47	2:00	7.30
PCS_{11}	VUE Point FX	Yes	Yes	4	18	Ga	4	3.27	2:00	2.65
PCS_{12}	VUE Point FX	No	Yes	2	32	Ga	6.4	5.47	2:00	2.51
PCS_{13}	VUE Point HD	Yes	No	2	24	Ga	4	2.73	3:00	2.81

PSF = point spread function; TOF = time-of-flight; FWHM = full-width at half-maximum; BckVar (%) = background variability calculated according to NEMA NU2-2012; Blob-OS-TF = Blob-basis function ordered-subsets time of flight; NA = not applicable; OSEM = ordered-subset expectation maximization; Un = unknown; LOR-RAMLA = line-of-response–based row-action-maximum-likelihood algorithm; Ga = gaussian.

All imaging systems operated with uniform voxel sizes.

of about 5.3 kBq/mL as recommended by the NEMA NU2-2012 standard (*34*). The phantom contains 6 spheres (10–37 mm) that were filled with an activity concentration of 4 times the background concentration (Fig. 1). All phantom acquisitions and image reconstructions were performed by the same expert according to the on-site clinical standards for whole-body ^{18}F-FDG PET/CT imaging (Table 1).

Delineation

The delineation process was performed using the Hermes Hybrid 3D software, version 2.0 (Hermes Medical Solutions). First, a cuboid volume of interest (VOI, 5 × 5 × 5 voxels) was defined in the background area of each PET image. Then, the 4 largest spheres (spheres 1–4, with diameters of 37, 28, 22, and 17 mm) that were visually identifiable in all reconstructed PET images were delineated using a semiautomatic region-growing tool to generate corresponding VOIs (S_{37}, S_{28}, S_{22}, and S_{17}). Only voxels with values higher than the mean of the background VOI were included in a given VOI. The VOIs (S_{37}–S_{17}) were dilated by 5 voxels by an automated dilatation tool (DS_{37}, DS_{28}, DS_{22}, and DS_{17} VOIs; Fig. 2). This step was performed to avoid interpolation artifacts at border voxel positions in the S_{37}–S_{17} VOIs during the resampling.

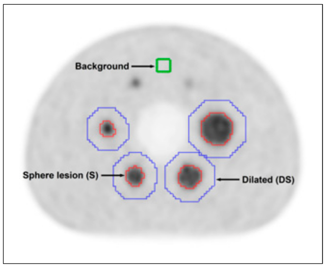

FIGURE 2. Axial slice of reconstructed NEMA image-quality PET phantom image with its overlaid delineated VOIs. Cuboid VOI (green) represents background region. Four small sphere VOIs (red) represent semiautomatically delineated spheres S_{17}, S_{22}, S_{28}, and S_{37} from left to right. Larger, dilated, VOIs (blue) are generated to avoid interpolation artifacts at border voxel positions in S_{37}–S_{17} VOIs during resampling.

Feature Extraction

For each acquisition, features were extracted from resampled images with 3 different voxel sizes (1, 2, and 4 mm) and combined with 4 different bin sizes (0.01, 0.025, 0.05, and 0.1 in units of tumor-to-background ratio). The combination of the image resolution and bin size parameters resulted in 12 feature extraction configurations ($C = \{c_1, \ldots, c_{12}\}$) (Fig. 3). The use of absolute bin sizes resulted in a variable number of bins (Fig. 4) (*27*).

To perform the feature extraction, the voxel values in the dilated VOIs (DS_{37}–DS_{17}) were normalized to the mean of the respective background VOI to calculate tumor-to-background ratios (*35,36*). The resampling to the given target resolution was then performed on the dilated VOIs (DS_{37}–DS_{17}) by ordinary kriging interpolation (*36,37*). The feature extraction was performed from the normalized, resampled DS VOIs, where the resampled S_{37}–S_{17} VOIs served as binary masks to identify voxels for the feature extraction (Fig. 4).

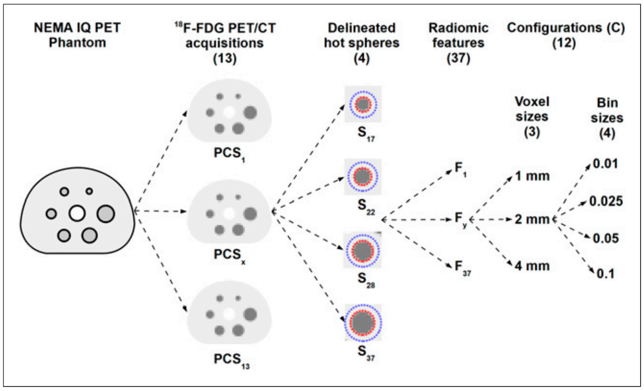

FIGURE 3. Representation of data acquisition and feature extraction processes. Same physical image-quality (IQ) phantom is used to acquire 13 ^{18}F-FDG PET/CT images from 12 imaging centers (PCS_1–PCS_{13}). Four largest visible hot spheres are delineated and analyzed. Thus, 37 radiomic features are extracted from each sphere with 3 voxel size and 4 bin size configurations.

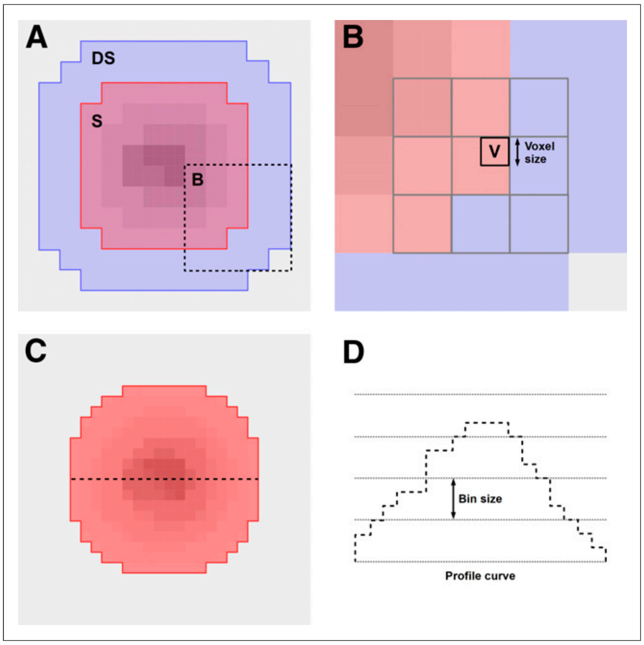

FIGURE 4. Explanation of resampling and binning steps that are performed for feature extraction. (A) Original image resolution with S VOI (red) and extended DS VOI (blue) regions (Fig. 2). Note, DS VOI also includes S VOI voxels. Dashed frame indicates zoomed subregion B. (B) Example target voxel (V in black frame) and original neighboring voxels (gray frames) that are involved in interpolation to determine V. Some of these voxels are outside S VOI; thus, resampling is performed from DS VOIs. (C) Radiomics analysis is performed from resampled DS VOI voxels that are inside resampled S VOI region (red). (D) Profile curve of voxels present at dashed line in C. Binning is characterized by choice of bin size, which defines which values are transformed to same bin. Feature extraction is performed over binned voxel values. This process results in variable number of bins per lesion.

Of the 37 features extracted from each of the 4 spheres (*36*), 34 were textural (*3,18*), whereas 3 features were shape-related and selected as independent features from binning (*18,31,38*) for reference comparison (Table 2). The feature extraction was performed by an in-house–developed program (feature extraction implementation properties are described in the supplemental definitions available at http://jnm.snmjournals.org).

Feature and PET/CT System Ranking

All 37 features and 13 PET/CT systems were ranked by a coefficient-of-variation (COV) analysis (*39,40*), where COV describes the SD of samples divided by their mean.

For each feature–PET/CT system pair, an individual COV was calculated over the 12 configurations (*C*). This step was performed for all 4 spheres, thus resulting in 4 feature–PET/CT system COV matrices. The ranking of the features was calculated for each sphere/VOI as the average COV over all PET/CT systems. Similarly, the ranking of the PET/CT systems was calculated for each sphere/VOI as the average COV across the respective 37 features (Fig. 5).

Feature Dependency on Voxel Size, Bin Size, and Volume

To assess the dependency of the features on voxel size, bin size, and sphere volume together, the COV of each of the 37 features was calculated across the 13 imaging systems for each sphere size and each of the 12 configurations (*C*). This resulted in 48 (12 configurations × 4 spheres) COVs. The COVs were subsequently grouped according to (a) voxel size, (b) bin size, and (c) sphere volume (Table 3). For each set of subgroups (a, b, and c) a 1-way ANOVA (*27,41*) was performed and the corresponding *P* value was used as a measure of dependence.

Feature Extraction Optimization

For each feature, the behavior of the COV changes in the 3 subgroups (a, b, and c) as a function of voxel size, bin size, and volume was assessed. To characterize the behavior trends, increasing, decreasing, inconsistent, and constant COV trend scenarios were considered. Last, the mean of S_{37}–S_{17} multicenter COVs for each of the 12 feature extraction configurations (*C*) was calculated. The configuration resulting in the smallest mean COV of the given feature was chosen as the optimal parameter set for feature extraction (Fig. 5).

RESULTS

Feature and PET/CT System Ranking

Information correlation (gray-level cooccurrence matrix [GLCM]) and shape features were least sensitive to feature extraction parameter (*C*) changes, followed by sum entropy (GLCM) and correlation (GLCM). The features that were most sensitive to feature extraction parameters were contrast and difference variance (GLCM) and contrast (neighborhood gray-tone difference matrix [NGTDM]), followed by 4 gray-level zone size matrix (GLZSM) features (Table 4, Supplemental Tables 1–4).

TABLE 2
Extracted Features from 4 Largest Spheres of Each PET Acquisition

Feature category	Feature name
GLCM (*18*)	Angular second moment, auto correlation, cluster prominence, cluster shade, contrast, correlation, difference entropy, difference variance, dissimilarity, entropy, information correlation, inverse difference, inverse difference moment, maximum probability, sum average, sum entropy, sum-of-squares variance, sum variance
GLSZM (*11*)	Gray-level nonuniformity, high gray-level zone emphasis, large zone high gray emphasis, large zone low gray emphasis, large zone size emphasis, low gray-level zone emphasis, small zone high gray emphasis, small zone low gray emphasis, small zone size emphasis, zone size nonuniformity, zone size percentage
NGTDM (*5*)	Busyness, coarseness, complexity, contrast, texture strength
Shape (*3*)	Compactness, spheric dice coefficient, volume

Details of feature calculations have been previously published (*18,36*).

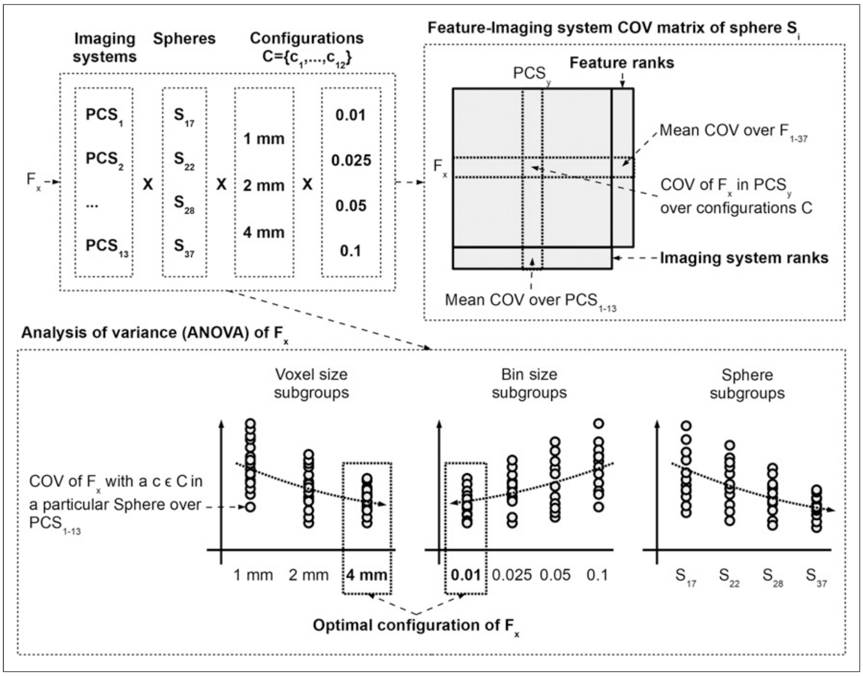

FIGURE 5. Each feature (F_x) has 13 imaging system, 4 sphere, and 12 configuration (3 voxel size and 4 bin size) variants. Feature and imaging system ranks are performed from feature-imaging system COV matrices. Each sphere (S_i) has its own COV matrix. Here, each matrix cell corresponds to COV of given feature F_x and PET/CT imaging system (PCS_y) over different feature extraction configurations (C). ANOVA analysis builds on subgrouping of COVs over PCS variants, as acquired by particular configuration ($c \in C$) in particular spheres. Optimal voxel size and bin size parameters are selected for F_x that minimize COV across imaging systems.

Table 5 summarizes the ranking of the 13 PET/CT imaging systems together with their standard imaging protocols (Table 1). The use of point-spread function modeling with a narrow gaussian postreconstruction filter (2–4 mm in full width at half maximum) together with large matrix sizes (192–256) led to higher imaging ranks for the individual PET/CT systems (Table 5). Imaging systems with time-of-flight capability did not generally rank higher. Likewise, the number of iterations, subsets, and time per bed position (Table 1) did not affect the imaging system ranks (Table 5). High-ranked imaging systems had lower background noise variation (Tables 1 and 5). In contrast, the low-ranked imaging systems represented no noticeable correlation with background noise variations, with the exception of the lowest-ranked system (PCS_{10}), which had the largest background variability, 7.3% (Tables 1 and 5).

Feature Dependency on Voxel Size, Bin Size, and Volume

Features independent of the voxel size were mainly from the GLCM category (Table 2): for example, correlation, sum variance, and cluster prominence had *P* values of 1.0, 0.995, and 0.992, respectively. In contrast, features most sensitive toward changes in voxel size were maximum probability (GLCM), angular second moment (GLCM), and compactness (shape), with *P* values of less than 0.001 (Supplemental Table 5).

Features from the GLCM category such as correlation, contrast, and cluster shade ($P \sim 1$) were independent of bin size, whereas GLZSM-based features were more dependent (Supplemental Table 6). Furthermore, large zone low gray emphasis ($P = 0.79$), large zone size emphasis ($P = 0.48$), and zone size percentage ($P = 0.44$) were less dependent on sphere volume. Dependencies on volume increased for GLCM features, such as cluster prominence, contrast, or sum variance, with *P* values near zero (Supplemental Table 7). Overall, the volume subgroup *P* values were considerably lower than *P* values of the voxel size and bin size subgroups (Supplemental Tables 5–7). Figure 6 shows an example of subgroup representation.

Feature Extraction Optimization

After the use of optimized feature extraction parameters, only 7 features resulted in small COVs (<5%), whereas 3, 3, and 27 were in the moderate (5% ≤ COV < 10%), elevated (10% ≤ COV < 20%), and large (COV ≥ 20%) categories, respectively (Table 6).

DISCUSSION

Quantitative radiomics analysis is challenged in multiple ways (*20,42*). In this study, we presented a holistic approach for analyzing and optimizing the process of feature extraction. By using a standard image-quality phantom, we ranked 37 popular radiomic features and 13 PET/CT imaging systems with regard to their stability. The imaging system ranks (Table 5) indicate that the influence of using point-spread function, a narrow gaussian postprocessing filter, and a large matrix has a larger impact on radiomics variations than the type of image reconstruction algorithm. Furthermore, we were able to demonstrate that multicenter feature variations can be minimized by preselecting feature-specific individual voxel size and bin size parameters based on their COV trends (Supplemental Tables 5–7; Table 6). With our feature optimization approach, 7 of our investigated features had a COV of less than 5%, and 3 of them had a COV of less than 10%. Without optimization, only one feature had a COV of less than 5%, and 4 had a COV of less than 10% (Table 6). The ANOVA subgroup analysis revealed that lesion volume was the most contributing factor in feature variations compared with voxel size and bin size

TABLE 3
Subgroups of COVs of Each Feature for 1-Way ANOVA

Groups	Voxel size	Bin size	Volume
Subgroups	3 (1 mm, 2 mm, 4 mm)	4 (0.01, 0.025, 0.05, 0.1)	4 (S_{37}, S_{28}, S_{22}, S_{17})
Subgroup elements	16 (4 volumes × 4 bin sizes)	12 (4 volumes × 3 voxel sizes)	12 (3 voxel sizes × 4 bin sizes)

TABLE 4
Feature Ranks with Regard to Average Absolute COV for 4 Largest Spheres (S_{37}–S_{17}).

Feature	Feature category	S_{37} COV rank	S_{28} COV rank	S_{22} COV rank	S_{17} COV rank
Information correlation	GLCM	0.00*	0.00*	0.00*	0.00*
Compactness	Shape	0.01*	0.02*	0.02*	0.03*
Volume	Shape	0.02*	0.02*	0.03*	0.03*
Spheric dice coefficient	Shape	0.03*	0.03*	0.07[†]	0.1[‡]
Sum entropy	GLCM	0.17[‡]	0.17[‡]	0.18[‡]	0.19[‡]
Correlation	GLCM	0.14[‡]	0.18[‡]	0.22	0.29
Entropy	GLCM	0.19[‡]	0.19[‡]	0.19[‡]	0.21
Small zone size emphasis	GLZSM	0.26	0.27	0.28	0.29
Difference entropy	GLCM	0.31	0.31	0.32	0.33
Zone size percentage	GLZSM	0.53	0.53	0.56	0.62
Inverse difference	GLCM	0.57	0.59	0.58	0.56
Coarseness	NGTDM	0.59	0.58	0.59	0.59
Inverse difference moment	GLCM	0.78	0.81	0.80	0.76
Sum average	GLCM	0.83	0.83	0.83	0.83
Dissimilarity	GLCM	1.07	1.07	1.07	1.08
Small zone low gray emphasis	GLZSM	1.12	1.10	1.10	1.11
Low gray-level zone emphasis	GLZSM	1.2	1.17	1.16	1.09
Maximum probability	GLCM	1.2	1.19	1.19	1.21
High gray-level zone emphasis	GLZSM	1.37	1.34	1.3	1.28
Angular second moment	GLCM	1.35	1.34	1.32	1.31
Auto correlation	GLCM	1.35	1.35	1.35	1.36
Texture strength	NGTDM	1.56	1.39	1.29	1.26
Sum variance	GLCM	1.35	1.35	1.35	1.36
Sum-of-squares variance	GLCM	1.35	1.35	1.35	1.36
Small zone high gray emphasis	GLZSM	1.42	1.39	1.37	1.35
Cluster prominence	GLCM	1.68	1.69	1.69	1.7
Cluster shade	GLCM	3.56	1.63	1.61	1.61
Zone size nonuniformity	GLZSM	1.7	1.76	1.92	1.85
Busyness	NGTDM	1.73	1.79	1.78	1.7
Complexity	NGTDM	2.12	1.86	1.72	1.65
Contrast	GLCM	2.03	2.03	2.04	2.06
Difference variance	GLCM	2.03	2.04	2.05	2.07
Contrast	NGTDM	1.69	2.10	2.35	2.46
Gray-level nonuniformity	GLZSM	2.1	2.12	2.17	2.21
Large zone high gray emphasis	GLZSM	2.75	2.65	2.55	2.41
Large zone size emphasis	GLZSM	3.23	3.24	3.22	3.13
Large zone low gray emphasis	GLZSM	3.29	3.28	3.26	3.21

*COV < 5%.

[†] 5% ≤ COV < 10%.

[‡] 10% ≤ COV < 20%.

COVs without footnotes are ≥20%. Smaller rank values correspond to smaller COV feature variations across their 12 feature extraction configurations and imaging systems.

changes (Supplemental Tables 5–7). Nevertheless, the multicentric variations of radiomic features generally vary as a function of activity distribution in the lesions. Furthermore, partial-volume effects (*15,43,44*) inherently increase heterogeneity in smaller lesions as well.

The clinical implications of our results are manifold. Since we involved 13 imaging systems applying clinical standard protocols, our trend analysis tables (Supplemental Tables 5–7) can serve as general lookup tables to understand the behavior of radiomic features as a function of voxel size, bin size, and volume changes.

TABLE 5
Imaging System (PCS) Protocol Parameter Ranks with Regard to Average Absolute COV for 4 Largest Spheres (S_{37}–S_{17})

PET/CT system	Algorithm	S_{37} COV	S_{28} COV	S_{22} COV	S_{17} COV
PCS_{13}	VUE Point HD	1.17	1.17	1.16	1.16
PCS_{11}	VUE Point FX	1.19	1.17	1.18	1.15
PCS_5	TrueX	1.18	1.18	1.16	1.18
PCS_6	TrueX	1.2	1.18	1.2	1.17
PCS_7	TrueX	1.18	1.2	1.2	1.22
PCS_8	TrueX	1.2	1.2	1.19	1.23
PCS_9	TrueX (HD PET)	1.85	1.17	1.2	1.2
PCS_1	Blob-OS-TF	1.22	1.21	1.23	1.21
PCS_4	LOR-RAMLA	1.23	1.24	1.23	1.22
PCS_2	OSEM	1.22	1.23	1.25	1.23
PCS_{12}	VUE Point FX	1.23	1.25	1.24	1.23
PCS_3	OSEM	1.27	1.26	1.25	1.23
PCS_{10}	VUE Point	1.25	1.27	1.26	1.26

Blob-OS-TF = Blob-basis function ordered-subsets time of flight; LOR-RAMLA = line-of-response–based row-action-maximum-likelihood algorithm; OSEM = ordered-subset expectation maximization.

Smaller ranks correspond to low COV variances in given sphere volume across each of 37 features and their 12 feature extraction configurations (*C*).

This information supports researchers in building more stable radiomic models in their studies. Although our results are based on tumor-to-background ratios, the fixed bin size approach preserved relative value range differences in our lesions; thus, our results are applicable to PET SUV units as well. With the help of our optimized COV table (Table 6), researchers can identify robust, reproducible features, whereas our imaging system ranks (Table 5) support imaging specialists in establishing new, radiomics-conforming PET acquisition protocols. In general, reducing feature variability supports the notion of standardizing the computation of radiomic features through standardized image processing, as suggested by the Image Biomarker Standardisation Initiative consortium (*42*). Accordingly, we consider our report a potential amendment to the Image Biomarker Standardisation Initiative guidelines.

To date, a wide range of studies have focused on the multicenter analysis of radiomic feature repeatability in PET (*20,28,45–47*). Fried et al. assessed the robustness of PET-based radiomic features when image reconstruction settings were varied across 3 PET/CT systems in lung cancer versus image-quality phantom acquisitions (*48*). Features that were reported as "reasonably robust" were contrast (GLCM), energy (GLCM), SD, and uniformity. In our study, contrast GLCM was one of the least reproducible feature (56% COV) even with optimized parameters (Table 6). However, Fried et al. involved 3 imaging systems only with variable reconstruction parameters and they did not incorporate different bin sizes in their analysis. Last, their textural feature equations are unknown; thus, differences in calculations may be present (*18*). Similarly, Yan et al. (*23*) investigated the variation in 55 textural features in light of different image reconstruction parameters in 20 lung cancer patients after ^{18}F-FDG PET/CT imaging. They reported inverse difference and low gray-level zone emphasis as robust features, whereas skewness, cluster shade, and zone percentage were the least robust (COV > 20%). In our study, we found similar results for cluster shade and zone percentage (COV > 20%). However, inverse difference moment (29.8% COV) and low gray-level zone emphasis (49.7% COV) were both highly variable. We consider 2 reasons for these discrepancies: first, Yan et al. applied a different delineation method, and second, they used a fixed number of bins (32, 64, and 128), whereas we used fixed bin sizes (*36*). In another study by Orlhac et al., 6 textural features were investigated in simulated and real patient data, including 10 sphere models with different activity distributions and 54 breast cancer PET/CT patients (*12*). The authors showed that all textural features were sensitive to voxel size differences (≤86%) and edge effects (≥29%). Our study confirmed that voxel size differences affect all features except GLCM correlation (Table 6). Shiri et al. (*24*) investigated variations in different intensity and radiomic

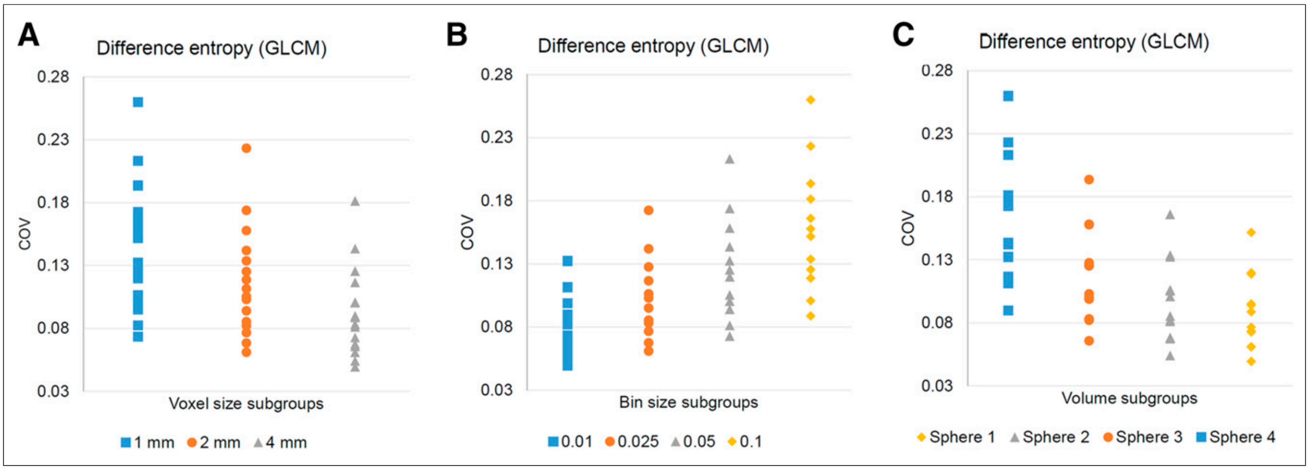

FIGURE 6. COV distributions of voxel size (A), bin size (B), and sphere volume (C) subgroups of feature difference entropy (GLCM). Each plotted sample corresponds to COV of given feature over $PCS_{1–13}$ with particular voxel size, bin size, and sphere volume configuration. Spheres 1–4 correspond to spheres S_{37}–S_{17}, respectively. Based on trend analysis, difference entropy has optimized voxel size of 4 mm (decreasing trend in function of increasing voxel size), optimized bin size of 0.01 (increasing trend in function of increasing bin size), and decreasing trend in function of decreasing volume.

TABLE 6
Features with Their Sphere S_{37}–S_{17} COVs (Mean ± SD), Optimal Voxel Size, Optimal Bin Size, and Resultant Optimized COV Across Imaging Systems

Feature	Feature category	COV (%)	Voxel size	Bin size	COV (%)
Information correlation	GLCM	0.0 ± 0.0*	4	0.01	0.0*
Compactness	Shape	0.6 ± 0.3*	1	NA	0.2*
Small zone size emphasis	GLZSM	12.3 ± 8.4‡	4	0.01	2.0*
Entropy	GLCM	6.9 ± 3.0†	4	0.01	2.1*
Zone size percentage	GLZSM	31.3 ± 27.5	4	0.01	3.6*
Sum entropy	GLCM	5.6 ± 1.3†	4	0.01	3.7*
Large zone size emphasis	GLZSM	100.3 ± 75.4	4	0.01	4.9*
Difference entropy	GLCM	11.7 ± 3.6‡	4	0.01	6.5†
Spheric dice coefficient	Shape	7.9 ± 1.4†	2	NA	6.8†
Coarseness	NGTDM	11.2 ± 4.7‡	1	0.01	7.45†
Correlation	GLCM	13.1 ± 0.1‡	1	0.1	12.9‡
Inverse difference	GLCM	21.3 ± 3.1	1	0.1	14.9‡
Angular second moment	GLCM	56.5 ± 17.7	4	0.01	17.6‡
Inverse difference moment	GLCM	29.8 ± 3.8	1	0.1	20.6
Volume	Shape	22.8 ± 0.5	4	NA	22.0
Sum average	GLCM	26.4 ± 0.7	2	0.01	25.3
Low gray-level zone emphasis	GLZSM	49.7 ± 28.5	4	0.01	26.9
Small zone low gray emphasis	GLZSM	48.6 ± 30.5	4	0.01	27.8
Busyness	NGTDM	60.5 ± 16.8	4	0.01	27.9
Gray-level nonuniformity	GLZSM	41.7 ± 3.9	4	0.01	28.7
Contrast	NGTDM	51.9 ± 12	1	0.1	29.0
Texture strength	NGTDM	43.9 ± 8.2	4	0.01	30.0
Dissimilarity	GLCM	31.4 ± 0.5	4	0.01	30.7
Large zone low gray emphasis	GLZSM	135.3 ± 75.4	4	0.01	30.8
Maximum probability	GLCM	52.4 ± 11	4	0.01	33.4
High gray-level zone emphasis	GLZSM	41.7 ± 3.9	1	0.05	35.6
Zone size nonuniformity	GLZSM	62.7 ± 17	4	0.01	38.3
Large zone high gray emphasis	GLZSM	76.6 ± 52	4	0.01	45.0
Auto correlation	GLCM	47.1 ± 0.8	2	0.01	45.7
Sum-of-squares variance	GLCM	47.7 ± 1.2	2	0.01	46.1
Sum variance	GLCM	47.9 ± 0.7	4	0.01	46.9
Small zone high gray emphasis	GLZSM	50.3 ± 7.8	4	0.01	47.4
Difference variance	GLCM	57.9 ± 1.3	1	0.1	54.0
Complexity	NGTDM	64.6 ± 4.8	4	0.01	55.1
Contrast	GLCM	57.1 ± 0.5	4	0.01	56.4
Cluster shade	GLCM	82.5 ± 4.7	1	0.01	76.9
Cluster prominence	GLCM	86.6 ± 1.2	4	0.01	84.7

*COV < 5%.
†5% ≤ COV < 10%.
‡10% ≤ COV < 20%.
NA = not applicable.
COVs without footnotes are ≥20%. List is sorted by increasing optimized COV.

features in 2 PET/CT systems using phantom acquisitions. Most textural features were reported to be sensitive (COV > 20%) with regard to voxel size changes, which we reconfirmed for features present in both studies (Supplemental Tables 5 and 6).

Lu et al. investigated the impact of delineation and binning methods including 40 ^{18}F-FDG patient studies, 5 delineation methods, and 88 features (*29*). Half the features depicted a higher intraclass correlation coefficient (≥0.8) with respect to segmentation, whereas 23% of features showed an intraclass correlation coefficient of at least 0.8 with regard to binning. Even though we did not compare different segmentation methods, our study reconfirmed that binning affects all textural features (Supplemental Table 6). Desseroit et al. studied the repeatability of shape and textural features of both low-dose CT and PET by means of different binning methodologies in a multicenter cohort of 74 ^{18}F-FDG PET/CT lung cancer patients (*31*). On the basis of their variable number of bins, they reported that all GLZSM features were poorly reliable and that angular second moment GLCM, contrast GLCM, and contrast NGTDM were the least repeatable, which was reconfirmed by our results (Table 6).

Altazi et al. (*32*) investigated 79 radiomic feature variations in light of different segmentation, reconstruction, and binning parameters in 88 cervical cancer patients having ^{18}F-FDG PET acquisitions. They reported inverse difference moment, entropy, difference entropy, and sum entropy (all GLCM) to be the most reproducible regarding binning variations, whereas none of the GLZSM and NGTDM features appeared to be reproducible. In our study, the above GLCM parameters were moderately reproducible as a function of binning variations (Supplemental Table 6), whereas GLZSM and NGTDM features were represented with lower reproducibility. Nevertheless, they used a fixed number of bins, which underestimates COV compared with the fixed bin size approach (*31*).

The effect of tumor size, image resolution, and noise levels in 66 ^{18}F-FDG PET radiomic features was investigated by van Velden et al. (*22*), who have shown that 37% and 73% of features were sensitive on resolution and volume changes, respectively. Our study reconfirmed that among voxel size, bin size, and volume changes, volume changes had the highest effect on feature variations (Supplemental Tables 5–7).

It currently appears more appropriate to follow a rigid methodologic approach toward sourcing robust and meaningful radiomic features (*18*). Our study addresses important quality factors in radiomics studies that relate to feature engineering. Specifically, we assessed the variability of popular radiomic features in light of clinically relevant combinations of quantification, image acquisition, and reconstruction settings (Table 1). As a result, we propose that radiomics studies should entail the dedicated selection of individual data processing configurations per feature, so that feature variations are minimized (Table 6). In general, a methodologic, high-quality approach to feature extraction should be preferred over reporting study-specific, fine-tuned performance results. In that regard, multicentric standardization efforts in compliance with responsible radiomics guidelines (*20,42,49*) should be promoted. Furthermore, we suggest that those features that had a high COV even after optimization should be normalized in the feature domain by methods such as ComBat, proposed by Orlhac et al. (*17*). Features that do not benefit from such approaches should be excluded from future studies. In the future, the selection of features that benefit from standardized feature extraction and feature normalization could contribute to the establishment of a type of "Radiomics NEMA" protocol in line with preestablished Image Biomarker Standardisation Initiative guidelines (*42*) that could represent one step toward the era of clinical radiomics.

CONCLUSION

Our results help in optimizing radiomics studies by selecting a priori features with known data acquisition and processing parameters that minimize individual feature variations. Our imaging system rank analysis aids imaging specialists in optimizing imaging protocol parameters to support repeatable radiomics analysis of ^{18}F-FDG PET/CT images. By selecting robust features that are aligned with the above concept and by following a responsible radiomics workflow, we can support the establishment of standardized radiomics approaches in clinical studies.

DISCLOSURE

The phantom data acquisition was supported by a grant from the Austrian Society of Nuclear Medicine and Molecular Imaging. This study utilized cloud computing resources that were granted by the Microsoft Research project CRM:0740606. No other potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We thank Markus Diemling and Kristin Papp (Hermes Medical Solutions, Sweden) for their support with the Hermes Hybrid 3D software workflow establishment and Albert Frömel (Microsoft Austria), Kenji Takeda (Microsoft Research), and Nicolas Sorger (Microsoft Austria) for their support with computational resources in the Microsoft Azure Cloud as part of the Microsoft Research Grant project CRM:0740606.

REFERENCES

1. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* 2012;48:441–446.
2. Gillies RJ, Anderson AR, Gatenby RA, Morse DL. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin Radiol.* 2010;65:517–521.
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* 2016;278:563–577.
4. Lee JW, Lee SM. Radiomics in oncological PET/CT: clinical applications. *Nucl Med Mol Imaging.* 2018;52:170–189.
5. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline ^{18}F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52:369–378.
6. Bundschuh RA, Dinges J, Neumann L, et al. Textural parameters of tumor heterogeneity in ^{18}F-FDG PET/CT for therapy response assessment and prognosis in patients with locally advanced rectal cancer. *J Nucl Med.* 2014;55:891–897.
7. Pyka T, Bundschuh RA, Andratschke N, et al. Textural features in pre-treatment [F18]-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy. *Radiat Oncol.* 2015;10:100.
8. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in ^{18}F-FDG-PET scans of oesophageal cancer. *Eur Radiol.* 2015;25:2805–2812.
9. George J, Claes P, Vunckx K, et al. A textural feature based tumor therapy response prediction model for longitudinal evaluation with PET imaging. *Proc Int Symp Biomed Imaging.* 2012:1048–1051.
10. Cortes-Rodicio J, Sanchez-Merino G, Garcia-Fidalgo MA, Tobalina-Larrea I. Identification of low variability textural features for heterogeneity quantification of ^{18}F-FDG PET/CT imaging. *Rev Esp Med Nucl Imagen Mol.* 2016;35:379–384.
11. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PLoS One.* 2016;11: e0166550.
12. Orlhac F, Nioche C, Soussan M, Buvat I. Understanding changes in tumor texture indices in PET: a comparison between visual assessment and index values in simulated and patient data. *J Nucl Med.* 2017;58:387–392.
13. Hatt M, Laurent B, Ouahabi A, et al. The first MICCAI challenge on PET tumor segmentation. *Med Image Anal.* 2018;44:177–195.
14. 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.

15. Leijenaar RTH, Carvalho S, Velazquez ER, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol.* 2013;52:1391–1397.
16. Sala E, Mema E, Himoto Y, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clin Radiol.* 2017;72:3–10.
17. Orlhac F, Boughdad S, Philippe C, et al. A post-reconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med.* 2018;59:1321–1328.
18. 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.
19. Larue RTHM, Defraene G, De Ruysscher D, Lambin P, Van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol.* 2017;90:20160665.
20. Vallières M, Zwanenburg A, Badic B, Cheze-Le Rest C, Visvikis D, Hatt M. Responsible radiomics research for faster clinical translation. *J Nucl Med.* 2018;59:189–193.
21. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol.* 2015;60:5471–5496.
22. van Velden F, Nissen I, Lammertsma A, Boellaard R. Dependence of various radiomics features on different imaging characteristics [abstract]. *J Nucl Med.* 2014;55(suppl):2071.
23. Yan J, Chu-Shern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in ^{18}F-FDG PET. *J Nucl Med.* 2015;56:1667–1673.
24. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on ^{18}F-FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol.* 2017;27:4498–4509.
25. Orlhac F, Theze B, Soussan M, Boisgard R, Buvat I. Multiscale texture analysis: from ^{18}F-FDG PET images to histologic images. *J Nucl Med.* 2016;57:1823–1828.
26. Bailly C, Bodet-Milin C, Couespel S, et al. Revisiting the robustness of PET-based textural features in the context of multi-centric trials. *PLoS One.* 2016;11:1–16.
27. Leijenaar RTH, Nalbantov G, Carvalho S, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep.* 2015;5:11075.
28. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in ^{18}F-FDG PET. *J Nucl Med.* 2012;53:693–700.
29. Lu L, Lv W, Jiang J, et al. Robustness of radiomic features in [^{11}C]choline and [^{18}F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. *Mol Imaging Biol.* 2016;18:935–945.
30. Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. ^{18}F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. *PLoS One.* 2015;10:e0145063.
31. Desseroit M-C, Tixier F, Weber WA, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. *J Nucl Med.* 2017;58:406–411.
32. Altazi BA, Zhang GG, Fernandez DC, et al. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. *J Appl Clin Med Phys.* 2017;18:32–48.
33. Rausch I, Bergmann H, Geist B, et al. Variation of system performance, quality control standards and adherence to international FDG-PET/CT imaging guidelines. *Nuklearmedizin.* 2014;53:242–248.
34. *Performance Measurements of Positron Emission Tomographs: NEMA NU 2.* Rosslyn, VA: National Electrical Manufacturers Association; 2013.
35. Hofheinz F, van den Hoff J, Steffen IG, et al. Comparative evaluation of SUV, tumor-to-blood standard uptake ratio (SUR), and dual time point measurements for assessment of the metabolic uptake rate in FDG PET. *EJNMMI Res.* 2016;6:53.
36. Papp L, Pötsch N, Grahovac M, et al. Glioma survival prediction with combined analysis of in vivo ^{11}C-MET PET features, ex vivo features, and patient features by supervised machine learning. *J Nucl Med.* 2018;59:892–899.
37. Laurenceau J, Sagaut P. Building efficient response surfaces of aerodynamic functions with kriging and cokriging. *AIAA J.* 2008;46:498–507.
38. Buvat I, Orlhac F, Soussan M. Tumor texture analysis in PET: where do we stand? *J Nucl Med.* 2015;56:1642–1644.
39. Van Velden FH, Cheebsumon P, Yaqub M, et al. Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. *Eur J Nucl Med Mol Imaging.* 2011;38:1636–1647.
40. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep.* 2015;5:13087.
41. van Elmpt W, Das M, Hüllner M, et al. Characterization of tumor heterogeneity using dynamic contrast enhanced CT and FDG-PET in non-small cell lung cancer. *Radiother Oncol.* 2013;109:65–70.
42. Zwanenburg A, Leger S, Vallières M, Löck S, Initiative for the IBS: image biomarker standardisation initiative. arXiv.org website. https://arxiv.org/abs/1612.07003. Published December 21, 2016. Revised September 17, 2018. Accessed January 17, 2019.
43. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol.* 2016;61:R150–R166.
44. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging.* 2013;40:1662–1671.
45. Lucia F, Visvikis D, Desseroit MC, et al. Prediction of outcome using pretreatment ^{18}F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy. *Eur J Nucl Med Mol Imaging.* 2018;45:768–786.
46. Scrivener M, de Jong EEC, van Timmeren JE, Pieters T, Ghaye B, Geets X. Radiomics applied to lung cancer: a review. *Transl Cancer Res.* 2016;5:398–409.
47. Hatt M, Majdoub M, Vallieres M, et al. ^{18}F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med.* 2015;56:38–44.
48. Fried D, Meier J, Mawlawi O, et al. MO-DE-207B-07: assessment of reproducibility of FDG-PET-based radiomics features across scanners using phantom imaging. *Med Phys.* 2016;43:3705–3706.
49. Decoding the tumor phenotype with non-invasive imaging. Radiomics website. http://www.radiomics.world/. Accessed January 17, 2019.