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

Marie-Charlotte Desseroit[1,2], Florent Tixier[2,3], Wolfgang A. Weber[4], Barry A. Siegel[5], Catherine Cheze Le Rest[2,3], Dimitris Visvikis[1], and Mathieu Hatt[1]

[1]Laboratory of Medical Information Processing, INSERM UMR 1101, IBSAM, University of Brest, Brest, France; [2]Medical School, University of Poitiers, Poitiers, France; [3]Nuclear Medicine, CHU Milétrie, Poitiers, France; [4]Memorial Sloan Kettering Cancer Center, New York, New York; and [5]Mallinckrodt Institute of Radiology and the Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri

The main purpose of this study was to assess the reliability of shape and heterogeneity features in both the PET and the low-dose CT components of PET/CT. A secondary objective was to investigate the impact of image quantization. **Methods:** A Health Insurance Portability and Accountability Act–compliant secondary analysis of deidentified prospectively acquired PET/CT test–retest datasets of 74 patients from multicenter Merck and American College of Radiology Imaging Network trials was performed. Metabolically active volumes were automatically delineated on PET with a fuzzy locally adaptive bayesian algorithm. Software was used to semiautomatically delineate the anatomic volumes on the low-dose CT component. Two quantization methods were considered: a quantization into a set number of bins (quantization B) and an alternative quantization with bins of fixed width (quantization W). Four shape descriptors, 10 first-order metrics, and 26 textural features were evaluated. Bland–Altman analysis was used to quantify repeatability. Features were subsequently categorized as very reliable, reliable, moderately reliable, or poorly reliable with respect to the corresponding volume variability. **Results:** Repeatability was highly variable among features. Numerous metrics were identified as poorly or moderately reliable. Others were reliable or very reliable in both modalities and in all categories (shape and first-, second-, and third-order metrics). Image quantization played a major role in feature repeatability. Features were more reliable in PET with quantization B, whereas quantization W showed better results in CT. **Conclusion:** The test–retest repeatability of shape and heterogeneity features in PET and low-dose CT varied greatly among metrics. The level of repeatability also depended strongly on the quantization step, with different optimal choices for each modality. The repeatability of PET and low-dose CT features should be carefully considered when selecting metrics to build multiparametric models.

Key Words: PET/CT; texture analysis; radiomics; repeatability

J Nucl Med 2017; 58:406–411
DOI: 10.2967/jnumed.116.180919

The crucial role of PET/CT with ^{18}F-FDG for diagnosis and staging of non–small cell lung cancer is established (*1*). Tumor metabolism is usually quantified with SUV metrics (e.g., maximum and mean) in PET, whereas the role of the low-dose CT component is limited to PET attenuation correction and anatomic localization.

Radiomics denotes the extraction of intensity, shape, and heterogeneity features from medical images (*2*). Its application to PET (*3*) and CT (*4*) has gained interest for characterizing non–small cell lung cancer tumors quantitatively, with potentially higher value than standard metrics, with the opportunity to combine features from both the PET and the low-dose CT components (*5*).

A first challenge is that numerous features can be calculated, most of which are sensitive to image noise, segmentation, or reconstruction settings (*7–11*). Their use for therapy response monitoring and early prediction faces another challenge: repeatability. Because metrics calculated in pre-, mid- and posttherapy images need to be compared, test–retest repeatability allows determining the cutoff above which a change is attributed to response or progression. This has been estimated at ±15% to 30% for SUV and volume (*12,13*). Regarding shape and heterogeneity metrics, several studies have investigated their repeatability in PET with ^{18}F-FDG or ^{18}F-fluorothymidine (*8,14–17*) and in diagnostic CT (*18,19*), dosimetry CT (*4,18*), contrast-enhanced CT (*18,20*), or cone-beam CT (*21*). These studies exploited small single-center cohorts (8 contrast-enhanced CT (*20*), 10 cone-beam CT (*21*), 11 ^{18}F-FDG PET (*8,15,17*), 11 ^{18}F-fluorothymidine PET (*16*), 16 ^{18}F-FDG PET (*14*), 20 CT and 13 contrast-enhanced CT (*18*), and 31 CT (*4,19*)) and never reported on the repeatability of features from the low-dose CT from PET/CT, which is important when combining features from both components (*5,6*).

Finally, it has been shown recently that the image quantization step in the calculation of textural features can have an impact on the relationship with other parameters (*3*) and on repeatability (*17,22*).

The primary goal of the present work was to evaluate the repeatability of shape and heterogeneity metrics from both the PET and the low-dose CT components in a large prospective multicenter cohort. A secondary goal was to evaluate the impact of the quantization step.

Received Jul. 11, 2016; revision accepted Aug. 29, 2016.
For correspondence or reprints contact: Marie-Charlotte Desseroit, LaTIM, INSERM UMR 1101, CHRU Morvan, 2 avenue Foch, 29609, Brest, France.
E-mail: Marie-Charlotte.Desseroit@etudiant.univ-brest.fr
Published online Oct. 20, 2016.

MATERIALS AND METHODS

Patient Cohort and Imaging

Patients with stage IIIB–IV non–small cell lung cancer were prospectively included in the multicenter Merck MK-0646-008 (40 patients in 17 sites) and American College of Radiology Imaging Network 6678 (34 patients in 14 sites) trials (NCT00424138 and NCT00729742, respectively) (*23*). The centers had to conform to the criteria of the American College of Radiology Imaging Network PET qualification (www.acrin.org/6678_protocol.aspx) to participate. Merck used a similar accreditation program. The PET/CT protocols were designed in accordance with National Cancer Institute guidelines (*24*). The institutional review board of each participating site approved the study, and all subjects gave written informed consent. The whole cohort of 74 patients had been included in a previous study (*23*), but that study analyzed only SUV measurements in PET whereas the present analysis also computed texture features and shape parameters both on the PET images and on the low-dose CT images. The present secondary analysis of deidentified PET/CT images from these trials was approved by the American College of Radiology Imaging Network and was performed in compliance with the Health Insurance Portability and Accountability Act.

PET and CT Analysis

For both the test and the retest datasets, the PET and the low-dose CT images were processed independently. In PET, the metabolically active volumes of the primary tumor and up to 3 additional lesions were segmented with the fuzzy locally adaptive bayesian algorithm previously validated for accuracy and robustness (*25,26*). In low-dose CT, the anatomic volume of primary tumors was delineated with a validated semiautomatic approach using 3D Slicer (*27*). Additional lesions were analyzed if they could be reliably delineated.

The following metrics were calculated on the delineated volumes. All features are described with their calculation formulas (*3*) in the supplemental material (available at http://jnm.snmjournals.org). Three-dimensional shape descriptors were included, such as sphericity, irregularity, and major axis (*4,28*).

First-order metrics (not accounting for the spatial distribution of voxels) in both Hounsfield units (low-dose CT) and SUV (PET) include maximum and mean values, as well as histogram-derived skewness, kurtosis, energy, entropy, and the area under the curve of the cumulative histogram (*29*). These metrics do not require quantization as a prior step. Quantization (not to be confused with quantification) is an intensity-resampling step applied to the image before building of texture matrices on which second and third order features rely. These matrix dimensions are determined by the number of intensity values obtained after this resampling. Several different quantization approaches have been proposed (*3*).

Second-order metrics from a gray-level-cooccurrence matrix and a neighborhood-gray-tone-difference matrix, and third-order metrics from a gray-level-zone-size matrix, were calculated in a single matrix considering all 13 orientations simultaneously (*30,31*). Quantization was performed in a set number of bins B (denoted from here onward as quantization B), as previously recommended (*14,18,30,32*) using Equation 1:

$$I_B = B \times \frac{I - I_{min}}{I_{max} - I_{min}}, \qquad \text{Eq. 1}$$

where I_{max} and I_{min} denote maximum and minimum intensity (Hounsfield units in low-dose CT and SUV in PET) and B is the number of bins (here, 64). Choosing a different B value can affect the repeatability of features (*14*). The results for a B value of 8 to 128 are presented in the supplemental material. It has been suggested that an alternative quantization using fixed-width bins (e.g., 0.5 SUV) can have an important impact (*17,22*). Results using this approach (denoted from here onward as quantization W) following Equation 2 were also generated.

$$I_W = \left\lceil \frac{I_O}{W} \right\rceil - \min\left(\left\lceil \frac{I_O}{W} \right\rceil \right) + 1, \qquad \text{Eq. 2}$$

where W is the bin width (here, 0.5 SUV for PET (*22*) and 10 Hounsfield units for low-dose CT). Note that a W value of 0.25 SUV and a W value of 5 Hounsfield units were also tested but no significant differences were observed. Supplemental Figure 1 shows a non–small cell lung cancer tumor imaged with both PET and low-dose CT, along with the corresponding quantization results and histograms.

Statistical Analysis

Statistical analyses were performed with MedCalc (MedCalc Software). The repeatability of each metric was assessed with Bland–Altman analysis by reporting the mean and SD of the differences between the two measurements. Lower and upper repeatability limits were calculated as $\pm 1.96 \times$ SD after log-transformation when not normal. Bland–Altman analysis was preferred over intraclass correlation coefficients on the basis of previous recommendations (*33*). Intraclass correlation coefficients are nonetheless provided in the supplemental material.

Correlations between metrics were assessed with Spearman rank coefficients.

Each metric was also categorized with respect to the repeatability (SD) of the corresponding volume of interest: very reliable ($\leq$0.5 times the repeatability), reliable ($>$0.5 to $\leq$1.5 times), moderately reliable ($>$1.5 to $\leq$2 times), and poorly reliable ($>$2 times).

RESULTS

The analysis was performed on only 73 datasets because 1 dataset was not available. In the PET images, 73 primary tumors and 32 additional lesions (nodal or distant metastases) were analyzed. Mean metabolically active volume was 47.8 cm^3 (median, 24.9 cm^3; SD, 55.4 cm^3). In the low-dose CT images, 2 patients were excluded because visual assessment of the images indicated that repeatable volume delineation could not be ensured (Supplemental Fig. 2). Seventy-one primary tumors and 5 additional lesions were analyzed. The mean anatomic volume was 52.4 cm^3 (median, 37.5 cm^3; SD, 53.0 cm^3).

Figure 1 displays the repeatability results for volume determination in both modalities, whereas Figures 2, 3, and 4 display the repeatability of first-order metrics and shape descriptors, second-order textural features, and third-order textural features, respectively. Tables containing all results along with other quantization values are in the supplemental material.

PET and Low-Dose CT Volumes

As shown in Figure 1, metabolically active volume determination had a repeatability of $-1.4\% \pm 11.1\%$, with upper and lower repeatability limits of $+20.3\%$ and -23.2%. Repeatability was dependent on the metabolically active volume, with smaller volumes exhibiting significantly poorer repeatability (Spearman rank coefficient $= -0.41$, $P < 0.0001$). The anatomic volume determination had a similar repeatability of $-0.4\% \pm 10.5\%$, with upper and lower repeatability limits of $+20.3\%$ and -21.0%. Repeatability was less dependent on volume (Spearman rank coefficient $= -0.32$, $P = 0.006$).

Each PET and low-dose CT feature was thus categorized as very reliable, reliable, moderately reliable, or poorly reliable, using similar but slightly different thresholds for each category (very reliable, $\leq$ 5.6% for PET and 5.3% for CT; reliable, $>$5.6% and $\leq$16.7% for PET and $>$5.3% and $\leq$ 15.8% for CT; moderately

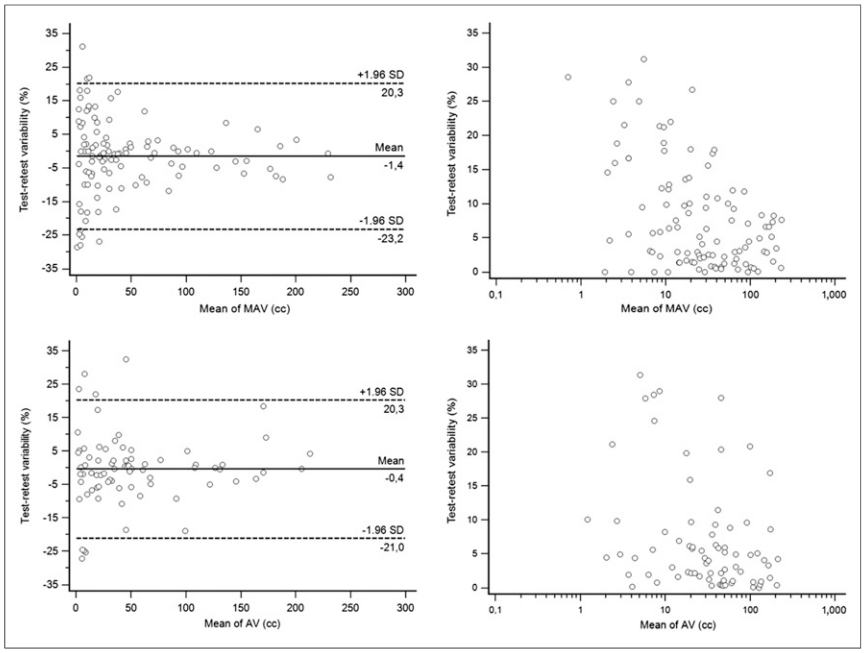

FIGURE 1. Bland–Altman analysis and correlation between volume and repeatability for metabolically active volume and anatomic volume determination. AV = anatomic volume; MAV = metabolically active volume.

reliable, >16.7% and ≤22.2% for PET and >15.8% and ≤21.0% for CT; poorly reliable, >22.2.% for PET and >21.0% for CT).

PET Features

Shape Descriptors and First-Order Metrics. Overall, the shape features in PET were very repeatable (Fig. 2). Irregularity and sphericity were very reliable, with only a 4.8% SD. Three-dimensional surface and major axis were reliable, although with higher variability (9.0% and 8.4%, respectively). Among intensity-based first-order features, the most repeatable were area under the curve of the cumulative histogram (−0.2% ± 3.6%) and histogram-derived entropy (−0.2% ± 3.6%), whereas the least repeatable were energy (−1.2% ± 23.8%) and skewness (−1.1% ± 33.7%). SUV_{mean} and SUV_{max} were moderately reliable, with upper and lower repeatability limits of −30.4% and 36.3%, respectively, for SUV_{mean} and −34.3% and 41.3%, respectively, for SUVmax.

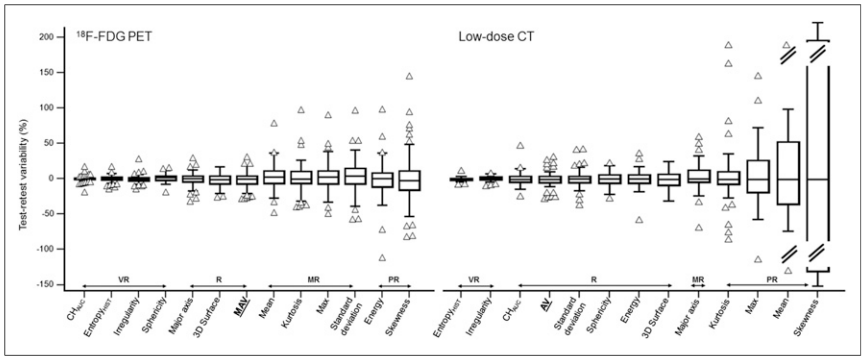

FIGURE 2. Repeatability of first-order metrics and 3-dimensional shape descriptors measured on ^{18}F-FDG PET and low-dose CT. Features are ranked left to right from highest to lowest repeatability. AV = anatomic volume; CH_{AUC} = area under the curve of the cumulative histogram; $entropy_{HIST}$ = histogram-derived entropy; MAV = metabolically active volume; MR = moderately reliable; PR = poorly reliable; R = reliable; VR = very reliable.

Second-Order Metrics. As shown in Figure 3, with quantization B, among gray-level-cooccurrence-matrix features, entropy (−0.1% ± 2.6%), sum entropy (−0.2% ± 2.1%), and difference entropy (−0.2% ± 3.0%) were the most repeatable, whereas most other features fell into the reliable category. Five were categorized as moderately reliable and 3 as unreliable. For correlation, the very poor repeatability was due to a few outliers for values around zero, to which Bland–Altman is very sensitive. After excluding them, correlation had reproducibility limits below ±20% and could be recategorized as moderately reliable. The 5 neighborhood gray-tone-difference-matrix features were less repeatable than the best gray-level-cooccurrence-matrix features although still categorized as reliable, all achieving an SD of around 14%–17%, except for neighborhood-gray-tone-difference-matrix contrast (27.6%).

The use of the alternate method, quantization W, changed both the above hierarchy and the absolute repeatability of the features. Overall, features calculated after quantization W were much less reliable and had notably more outliers, all exhibiting a higher variability than metabolically active volume.

Third-Order Metrics. As shown in Figure 4, among third-order metrics, quantization had a similar impact: with quantization W, all gray-level-zone-size-matrix features were categorized as poorly reliable, whereas with quantization B, 2 were very reliable (small-zone-size emphasis and zone-size percentage, with an SD of <4%) and 3 were reliable (large-zone-size emphasis, gray-level nonuniformity, and zone-size nonuniformity, with an SD of ~11%–14%). Among the least repeatable features were those focusing on small zones or low gray values (e.g., large-zone/low-gray emphasis, small-zone/low-gray emphasis, and low-gray-level-zone emphasis).

Low-Dose CT Features

Shape Descriptors and First-Order Metrics. As shown in Figure 2, morphologic irregularity, sphericity, and 3-dimensional surface were the most repeatable (SDs of 3.3%, 10.0%, and 11.6%, respectively). Major axis was less reliable (3.8% ± 18.4%).

On the one hand, 4 histogram metrics showed poor reliability: maximum (4.7% ± 38.6%) and mean (−4.2% ± 43.6%) intensity, kurtosis (4.8% ± 37.4%), and skewness (11.1% ± 202.2%). On the other hand, histogram-derived entropy and area under the curve of the cumulative histogram were very reliable (−0.1% ± 2.5% and 0.7% ± 9.1%, respectively).

Second-Order Metrics. Repeatability depended strongly on the quantization method, with quantization W providing better repeatability than quantization B (Fig. 3). Among gray-level-cooccurrence-matrix metrics, the most repeatable were

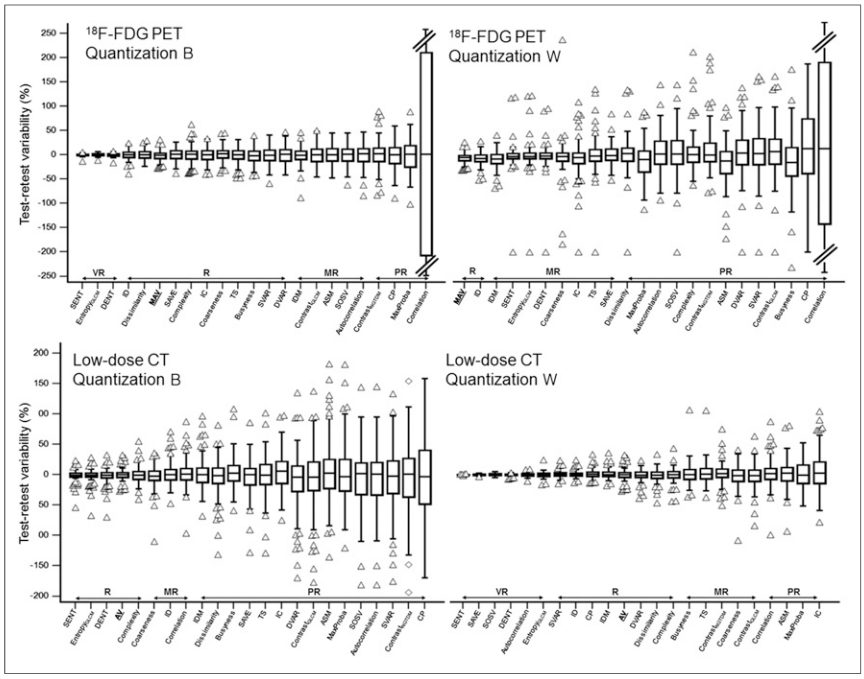

FIGURE 3. Repeatability of second-order metrics measured on ^{18}F-FDG PET and low-dose CT, using either quantization B or quantization W. Features are ranked left to right from highest to lowest repeatability. ASM = angular secondary moment; AV = anatomic volume; $contrast_{GLCM}$ = gray-level-cooccurrence-matrix contrast; $contrast_{NGTDM}$ = neighborhood-gray-tone-difference-matrix contrast; CP = cluster prominence; DENT = difference entropy; DVAR = difference variance; $entropy_{GLCM}$ = gray-level-cooccurrence-matrix entropy; IC = information correlation; ID = inverse difference; IDM = inverse different moment; MaxProba = maximum probability; R = reliable; MAV = metabolically active volume; MR = moderately reliable; PR = poorly reliable; SAVE = sum average; SENT = sum entropy; SOSV = sum of square variance; SVAR = sum variance; TS = texture strength; VR = very reliable.

entropy (−1.9% ± 12.0% vs. −0.4% ± 5.2% with quantizations B and W, respectively), sum entropy (−1.4% ± 10.0% vs. 0.1% ± 0.4%), and difference entropy (−2.3% ± 13.1% vs. −0.3% ± 1.9%). To a lesser extent, the same was observed for neighborhood-gray-tone-difference matrix, with higher repeatability using quantization W. Complexity was the only parameter with variability of less than 15.8% and was categorized as reliable (0.5% ± 14.3% and −0.5% ± 12.3% with quantizations B and W, respectively).

Third-Order Metrics. The quantization method also had an important impact (Fig. 4). Reliability was categorized as at least moderate for 8 parameters with quantization W but for only 2 parameters with quantization B. Small-zone-size emphasis (−0.6% ± 4.8% vs. −0.5% ± 2.6% with quantizations B and W, respectively) and zone-size emphasis (−2.8% ± 17.4% vs. −0.9% ± 11.9%) were the most repeatable features (Figs. 4D and 4E).

Impact of Quantization Method

Overall, the inverted impact of the quantization method in PET and low-dose CT can be explained by differences in correlation between the features and the corresponding volume and maximum intensity. In PET, we observed that features calculated with quantization W correlated with SUV_{max} but not with metabolically active volume. In contrast, features calculated with quantization B correlated with metabolically active volume but not with SUV_{max}. The higher repeatability obtained with quantization B can thus be explained by the fact that metabolically active volume repeatability was much higher than SUV_{max} repeatability. In contrast to PET, features in low-dose CT correlated with both volume and maximum intensity using quantization B but were less correlated or showed no correlation with either volume or intensity using quantization W. Because repeatability was much worse for maximum intensity than for CT volume, quantization B led to worse repeatability. This is illustrated in Figure 5 for the feature dissimilarity. For the PET component, the relative inversion of relationships with volume and SUV_{max} for quantization B compared with quantization W can be seen. In contrast, for the low-dose CT component, quantization B led to a higher correlation with maximum intensity than with volume, but quantization W led to lower correlation with volume and a nonsignificant correlation with maximum intensity.

DISCUSSION

In the present work, 73 test–retest PET/CT acquisitions from 31 centers (17 from the American College of Radiology Imaging Network in the United States and 14 from Merck in Asia and Europe) were analyzed for repeatability.

A similar variability in volume delineation was observed for both modalities. Metabolically active volume measured from PET was slightly smaller than anatomic volume measured from CT, mostly because more lymph nodes and metastases were delineated in PET than in CT and because portions of some large CT volumes had no ^{18}F-FDG uptake. Regarding SUV_{mean} and SUV_{max}, our results differ slightly from those previously published for the same cohort (*23*). Only lesions with an SUV_{max} of more than 4 were included in the previous analysis, whereas the current analysis did not apply this restriction. When we did restrict our analysis to lesions with an SUV_{max} of more than 4, our test–retest results for SUV_{max} were similar to those previously reported.

Regarding shape and heterogeneity features, our results confirm prior findings in PET (*8,14–17*). To the best of our knowledge, our study is the first to report on the repeatability of these features in the low-dose CT component.

Overall, the geometric features (shape descriptors) were found to be reliable (some with high repeatability) in both modalities, which can be related to the high repeatability of segmentation. This is in line with previous findings for PET (*8,17*) and with morphologic shape in other CT modalities (*4*). We emphasize that only one segmentation by one expert was considered. The variability might be higher when considering different segmentation approaches or several observers.

Regarding first-order metrics and textural higher-order features, our results confirm that the repeatability varies greatly among metrics. On the one hand, several features were confirmed to be unreliable in both modalities and should be systematically avoided—for example, first-order skewness; second-order angular second moment, gray-level-cooccurrence-matrix contrast, and neighborhood-gray-tone-difference-matrix contrast; and third-order metrics quantifying low gray values

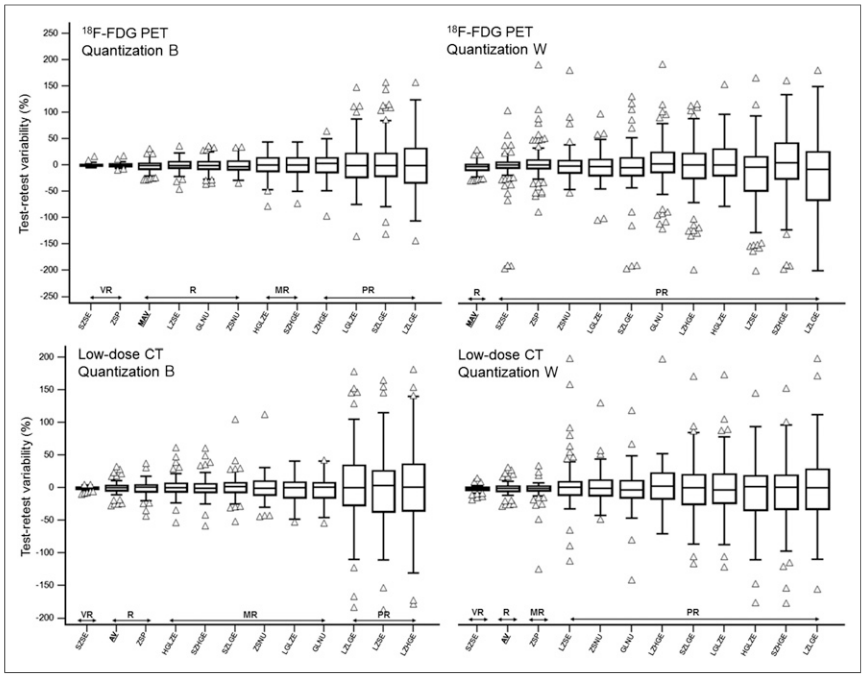

FIGURE 4. Repeatability of third-order metrics measured on ^{18}F-FDG PET and low-dose CT, using either quantization B or quantization W. Features are ranked left to right from highest to lowest repeatability. AV = anatomic volume; GLNU = gray-level nonuniformity; HGLZE = high-gray-level-zone emphasis; LGLZE = low-gray-level-zone emphasis; LZHGE = large-zone/high-gray emphasis; LZLGE = large-zone/low-gray emphasis; LZSE = large-zone-size emphasis; MAV = metabolically active volume; MR = moderately reliable; PR = poorly reliable; R = reliable; SZHGE = small-zone/high-gray emphasis; SZLGE = small-zone/low-gray emphasis; SZSE = small-zone-size emphasis; VR = very reliable; ZSNU = zone-size nonuniformity; ZSP = zone-size percentage.

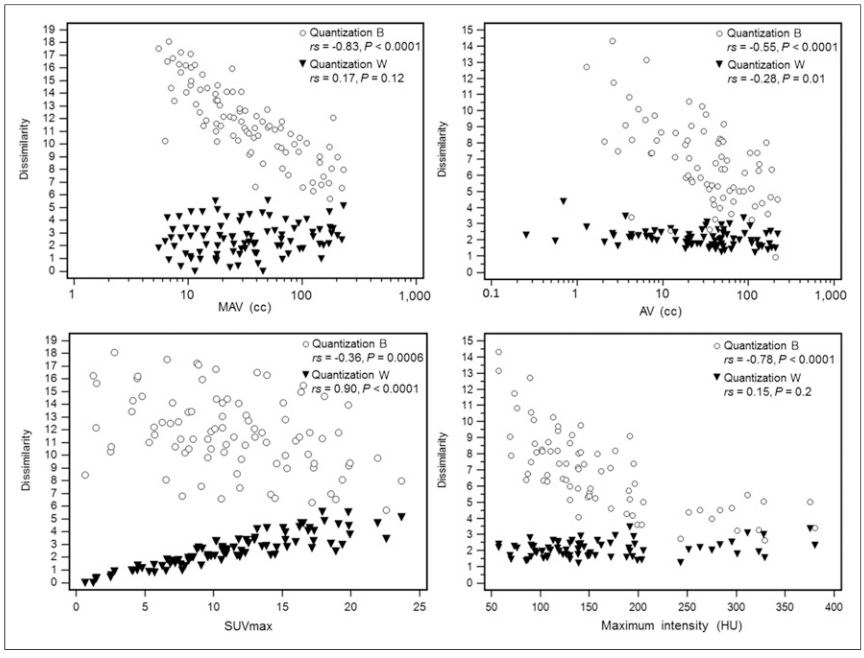

FIGURE 5. Correlation between textural feature (dissimilarity from gray-level-cooccurrence matrix) and volume (first row) or maximum intensity (second row), in both PET (first column) and low-dose CT (second column) components, depending on quantization approach. AV = anatomic volume; MAV = metabolically active volume; rs = Spearman rank coefficient.

or small zones. On the other hand, it should be emphasized that several features were identified as reliable, in all 3 categories and for both modalities. In between, other features with moderate repeatability should be used with caution as they exhibit larger variability than the corresponding volume determination.

We compared 2 different quantization methods. Quantization B is most often used. The impact of choosing another B value has been evaluated previously (*14*) and our results confirm these findings. Although a B of 64 is a good compromise and most features exhibited similar repeatability with different values, the repeatability of some metrics depended on B. We observed a different impact in PET and low-dose CT for quantization W, as it led to worse repeatability in PET but better repeatability in low-dose CT. This was explained by the different relationships between the features and the corresponding volume and maximum intensity. With more control over data acquisition and higher repeatability of SUV_{max}, quantization W may lead to higher repeatability. These results highlight the major impact of the quantization step and its variable impact depending on image modality that should thus not be overlooked.

Our results confirm that studies building clinical models by combining features from PET/CT images should carefully account for repeatability. This is mandatory when calculating the evolution of features across pre-, mid-, or posttherapy images. This is nonetheless an important factor when building models based on single-time-point images, as models built using robust and repeatable features are more likely to be generalizable and achieve good performance in external or testing cohorts. Repeatability is not the only criterion on which feature selection needs to be based, as discriminative power, robustness, and redundancy have to be considered also.

Our study has limitations. Low-dose CT and PET images were analyzed separately using different segmentation processes performed independently on the test and retest images. The repeatability evaluation therefore includes the intrinsic repeatability of the segmentation. We used robust segmentation approaches that should minimize variability. Another approach would consist in defining the volume on the test image and registering it on the retest image, which, however, requires accurate registration and raises other issues (*34*). In a clinical environment, the use of less accurate and less robust segmentation could lead to a lower repeatability, especially for volume-correlated features.

We chose to categorize the repeatability levels of each metric with respect to those of

the corresponding volume. The repeatability acceptance was similar for both modalities (reliability in PET was defined as an SD below 16.5%, compared with 15.8% for low-dose CT). These thresholds are arbitrary, and choosing different values would change the categorization of several metrics but without changing their hierarchy.

Finally, respiratory gating was not applied. In non–small cell lung cancer the lack of gating may lead to different levels of quantitative bias between the test and retest images, as well as between PET and low-dose CT. The repeatability we reported is therefore larger than what could ideally be obtained in other body regions where motion is less important, or if respiratory motion correction were applied (*35*).

CONCLUSION

The test–retest repeatability of shape and heterogeneity features in both components of PET/CT varied greatly among metrics. Repeatability also depended on the quantization step, with different optimal choices for PET or low-dose CT because of differences in the relationship between the metrics and volume or intensity. The repeatability of PET/CT features should be carefully accounted for when one is choosing metrics to combine in multiparametric models.

DISCLOSURE

Marie-Charlotte Desseroit's PhD is partly funded by Brest Métropôle Océane. Florent Tixier is funded by the association "Sport and Collection," CHRU Poitiers. This work has received a French government support granted to the CominLabs Laboratory of Excellence and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01. This work was also supported by the National Institute of Cancer (INCa project C14020NS). The original trials from which the images used in this study were obtained were supported by the U.S. National Cancer Institute through grants U01-CA079778 and U01-CA080098 and by Merck & Co., Inc. No other potential conflict of interest relevant to this article was reported.

REFERENCES

1. Sauter AW, Schwenzer N, Divine MR, Pichler BJ, Pfannenberg C. Image-derived biomarkers and multimodal imaging strategies for lung cancer management. *Eur J Nucl Med Mol Imaging.* 2015;42:634–643.
2. 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.
3. Hatt M, Tixier F, Pierce L, Kinahan P, Cheze Le Rest C, Visvikis D. Characterization of PET images using texture analysis: the past, the present… any future? *Eur J Nucl Med Mol Imaging.* June 6, 2016 [Epub ahead of print].
4. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006.
5. Desseroit MC, Visvikis D, Tixier F, et al. Development of a nomogram combining clinical staging with ^{18}F-FDG PET/CT image features in non-small cell lung cancer stage I-III. *Eur J Nucl Med Mol Imaging.* 2016;43:1477–1485.
6. Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. *Radiother Oncol.* 2012;102:239–245.
7. 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.
8. 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.
9. 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.
10. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol.* 2010;49:1012–1016.
11. 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.
12. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50(suppl 1):122S–150S.
13. Hatt M, Cheze-Le Rest C, Aboagye EO, et al. Reproducibility of ^{18}F-FDG and 3′-deoxy-3′-^{18}F-fluorothymidine PET tumor volume measurements. *J Nucl Med.* 2010;51:1368–1376.
14. 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.
15. Van Velden FHP, Nissen IA, Jongsma F, et al. Test-retest variability of various quantitative measures to characterize tracer uptake and/or tracer uptake heterogeneity in metastasized liver for patients with colorectal carcinoma. *Mol Imaging Biol.* 2014;16:13–18.
16. Willaime JMY, Turkheimer FE, Kenny LM, Aboagye EO. Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of ^{18}F fluorothymidine-positron emission tomography. *Phys Med Biol.* 2013;58:187–203.
17. 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. Fried DV, Tucker SL, Zhou S, et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. *Int J Radiat Oncol Biol Phys.* 2014;90:834–842.
19. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and prognosis of quantitative features extracted from CT images. *Transl Oncol.* 2014;7:72–87.
20. Yang J, Zhang L, Fave XJ, et al. Uncertainty analysis of quantitative imaging features extracted from contrast-enhanced CT in lung tumors. *Comput Med Imaging Graph.* 2016;48:1–8.
21. Fave X, Mackin D, Yang J, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? *Med Phys.* 2015;42:6784.
22. 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.
23. Weber WA, Gatsonis CA, Mozley PD, et al.; ACRIN 6678 research team, MK-0646-008 research team. Repeatability of ^{18}F-FDG PET/CT in advanced non-small cell lung cancer: prospective assessment in 2 multicenter trials. *J Nucl Med.* 2015;56:1137–1143.
24. Shankar LK, Hoffman JM, Bacharach S, et al.; National Cancer Institute. Consensus recommendations for the use of ^{18}F-FDG PET as an indicator of therapeutic response in patients in National Cancer Institute trials. *J Nucl Med.* 2006;47:1059–1066.
25. Hatt M, Cheze le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol Phys.* 2010;77:301–308.
26. Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. *Eur J Nucl Med Mol Imaging.* 2011;38:663–672.
27. Velazquez ER, Parmar C, Jermoumi M, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci Rep.* 2013;3:3529.
28. Apostolova I, Rogasch J, Buchert R, et al. Quantitative assessment of the asphericity of pretherapeutic FDG uptake as an independent predictor of outcome in NSCLC. *BMC Cancer.* 2014;14:896.
29. 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.
30. Hatt M, Majdoub M, Vallières 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.
31. 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.
32. Hunter LA, Krafft S, Stingo F, et al. High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images. *Med Phys.* 2013;40:121916.
33. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. *PLoS One.* 2012;7:e37908.
34. Yip SSF, Coroller TP, Sanford NN, et al. Use of registration-based contour propagation in texture analysis for esophageal cancer pathologic response prediction. *Phys Med Biol.* 2016;61:906–922.
35. Yip S, McCall K, Aristophanous M, Chen AB, Aerts HJWL, Berbeco R. Comparison of texture features derived from static and respiratory-gated PET images in non-small cell lung cancer. *PLoS One.* 2014;9:e115510.