Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in ^{18}F-FDG PET imaging

Florent TIXIER[1,2], Mathieu Hatt[1], Catherine Cheze Le Rest[1], Adrien Le Pogam[1], Laurent Corcos[2], Dimitris Visvikis[1]

[1] INSERM, U650, LaTIM, CHRU Morvan, Brest F-29200, France

[2] INSERM, U613, Faculty of Medicine, Brest F-29200, France

Corresponding author:

TIXIER Florent
LaTIM, INSERM U650,
CHRU Morvan, 5 avenue Foch,
29609 Brest Cedex, France
Tél: +33 (0)298-018-111
Fax: +33 (0)298-018-124
Email: florent.tixier@univ-brest.fr

Short running title: PET heterogeneities' reproducibility

ABSTRACT

^{18}F-FDG PET measurement of standardized uptake values (SUV) is increasingly used for monitoring therapy response or predicting outcome. Alternative parameters computed through textural analysis were recently proposed to quantify the tumor tracer uptake heterogeneity as significant predictors of response. The primary objective of this study was the evaluation of the reproducibility of these heterogeneity measurements. **Methods:** Double-baseline ^{18}F-FDG PET scans of 16 patients acquired within a period of 4 days prior to any treatment were considered. A Bland-Altman analysis was carried out on six parameters based on histogram measurements and 17 heterogeneity parameters based on textural features obtained after discretization with values between 8 and 128. **Results:** SUV_{max} and SUV_{mean} reproducibility were similar to previously reported studies with a mean percentage difference of 4.7±19.5% and 5.5±21.2% respectively. By comparison better reproducibility was measured for some of the textural features describing tumor tracer local heterogeneity, such as entropy and homogeneity with a mean percentage difference of -2±5.4% and 1.8±11.5% respectively. Several of the tumor regional heterogeneity parameters such as the variability in the intensity and size of homogeneous tumor activity distribution regions had similar reproducibility to the SUV measurements with 95% confidence intervals of -22.5% to 3.1% and -1.1% to 23.5% respectively. These parameters were largely insensitive to the discretization range values. **Conclusion:** Several of the parameters derived from textural analysis describing tumor tracer heterogeneity at local and regional scales had similar or better reproducibility as simple SUV measurements. These reproducibility results suggest that these FDG PET image derived parameters which have already been shown to have a predictive and prognostic value in certain cancer models, may be used within the context of therapy response monitoring or predicting patient outcome.

Keywords: PET, ^{18}F-FDG, texture analysis, reproducibility, oncology

INTRODUCTION

^{18}F-FDG PET imaging is well established in clinical practice for diagnosis and staging. On the other hand there is increasing interest in the use of this imaging modality within the context of therapy response assessment or patient follow-up. For such applications, standardized uptake value (SUVs) measurements are used, with the maximum of tumor activity concentration (SUV_{max}) being the most popular since it is the easiest to obtain. The use of the mean obtained in an $1cm^3$ sphere centered on the voxel of maximum activity concentration (SUV_{peak} (1)), has been proposed as an alternative since it should be more robust to noise compared to SUV_{max}, remaining at the same time easy to derive. Additional PET image derived parameters allowing a more complete lesion characterization include the mean SUV (SUV_{mean}), the metabolically active tumor volume (MATV, defined as the tumor volume that can be seen and delineated on a PET image) and the total lesion glycolysis (TLG, defined as the product of MATV and its associated SUV_{mean}), although they all require an accurate delineation of the functional tumor volume. Different studies have in the past explored the role of such PET image derived parameters for assessing response to therapy (2-6). More recently tracer uptake heterogeneity characterization based on textural analysis extracted from PET images has been also proposed, allowing an improved predictive and prognostic value to be derived from baseline PET scans (7,8).

Most frequently monitoring response to therapy involves a comparison of such PET image derived parameters between a baseline PET scan and a second scan carried out early or late during treatment, or after the end of treatment. In this case the variation of the parameters between the two scans is used to characterize response (1). Whether considering the % difference of PET image derived parameters between successive scans or the absolute values on a baseline scan the definition of thresholds in order to identify response or progressive disease requires, amongst others, an evaluation of the physiological reproducibility that characterizes them. Such evaluations are performed on double baseline scans acquired before any treatment within a few days interval from each other.

Until now only few studies have investigated the physiological reproducibility of such measurements, almost exclusively focusing on SUVs (9-11), and more recently on the MATV computed using different segmentation algorithms (12,13). Other authors have demonstrated the sensitivity of several textural feature parameters to PET acquisition and reconstruction settings (14), demonstrating the need for standardization in order for such image derived parameters to be used in therapy response assessment studies. However, the physiological reproducibility of these promising parameters extracted from the analysis of tumor activity distributions has never been investigated. The objective of our study was therefore to evaluate the reproducibility of textural features quantifying in a local, regional and global fashion the tumor tracer uptake heterogeneities, thereby identifying the potential of these parameters to be used for therapy response monitoring purposes. A comparison with the physiological reproducibility of SUVs using the same patient datasets was also performed since they are the most used parameters in current clinical practice and in order to facilitate a direct comparison with previous reproducibility studies.

MATERIALS AND METHODS

Patients

16 patients with newly diagnosed esophageal cancer were enrolled in this study. All of these patients underwent two ^{18}F-FDG PET baseline scans before initiating any treatment. The two scans were obtained within 2-7 days (median 4.2 days). PET images were acquired on a PET/CT scanner (Gemini; Philips), with 2-min acquisitions per bed position, 60 min after the injection of 6MBq/kg of ^{18}F-FDG. Data were reconstructed using a 3D row-action maximization-likelihood algorithm (RAMLA (15)) with standard clinical protocol parameters (2 iterations, relaxations parameter of 0.05, and 5mm full width at half maximum 3D Gaussian post-filtering). This analysis was carried out after obtaining the approval of the local Institutional Ethics Review Board.

Tumor Analysis

The primary lesions of each patient were delineated with the Fuzzy Locally Adaptive Bayesian (FLAB) algorithm which has been previously demonstrated to provide reproducible MATV automatic delineations (mean difference between baseline scans of 5±13%) (16). SUV_{max} and mean SUV within the delineated tumor (SUV_{mean}) were extracted from the primary tumor in each of the two baseline PET images for each patient. In addition, a number of tumor heterogeneity parameters shown in table 1, whose value for prognosis and prediction of outcome and treatment response on FDG PET images has been previously investigated (7,8), were calculated based on the delineated 3D functional volumes.

Textural Analysis

We define texture as a spatial arrangement of a predefined number of voxels allowing the extraction of complex image properties and we define a textural feature as a measurement computed using a texture matrix (8). Given that these features quantify the spatial relationship between voxels and their relative intensities, they can be associated to tracer heterogeneity patterns within the functional volume of the tumor at different scales, namely local and regional (using texture matrices) or global (using image-voxel-intensity histograms). The first type of matrices is used to quantify local heterogeneity as they allow characterization of the intensity variations between consecutive voxels. On the other hand, the second type of matrices allows characterization of arrangements of larger homogeneous areas (groups of voxels) within the tumors therefore providing information on tumor regional heterogeneity.

Local heterogeneity parameters were derived using the co-occurrences matrices (17) and were computed by considering a 26-connexity (i.e. neighboring voxels in all 13 directions in

three dimensions) and a 1-distance (i.e. no gap) relationship between consecutive voxels. On these matrices, 6 different parameters characterizing the local heterogeneity were calculated by averaging the values on the 13 directions for each feature. The other type of texture matrices is called intensity size-zone matrix (8, 18) and is constructed in two steps. First, homogenous areas are identified within the tumor and a matrix linking the size of each of these homogeneous areas to its intensity is constructed. 11 features characterizing the regional heterogeneity were calculated from this matrix. For example, parameters can quantifying the presence of large areas with high intensity (HILAE) or small areas with a low intensity (LISAE).

Other features characterizing regional heterogeneity include the variability in the size (SZV) and the intensity (IV) of identified homogeneous tumor zones, as well as the ratio between the number of homogeneous tumor zones and the overall tumor size (known as the zone-percentage (ZP)). Regional heterogeneity formulae were summarized in table 2 and the mathematical definition of all local features used in this study have been previously summarized in Haralick et al (17). A complete list of texture matrices and their associated features used in this work are included in table 1.

Building texture matrices on which the textural features are computed require a discretization of the voxel values within the previously delineated MATV on a specific range of values. This range has to be chosen as a power of two due to algorithmic constraints and in this study the features were extracted by considering downsampling to ranges of 8, 16, 32, 64 and 128 distinct values. Figure 1 illustrates on a transaxial tumor slice the resulting resampled MATV for each of these discretization ranges. This necessary downsampling step on the one hand reduces image noise while on the other normalizes the tumor voxel intensities across patients, subsequently facilitating the comparison of the extracted textural features. In a previous study (8) there were no statistically significant differences shown in the extracted textural feature values as a result of varying the number of discrete values in this resampling normalization process. 64 discrete values were considered sufficient for a range of SUVs

between 4 and 20. In the present study the influence of this parameter in the physiological reproducibility of the textural feature parameters was also assessed.

Statistical Analysis

The reproducibility of the quantitative values (q) for each parameter under investigation was assessed by calculating the mean percentage difference relative to the mean of both baseline scans using the following formula:

$$\Delta = \frac{(q_1 - q_2)}{(q_1 + q_2)/2} . 100 \qquad Eq. 1$$

This analysis was performed for all parameters and in the case of the textural features for all discretization values (from 8 to 128). A Kolmogorov-Smirnov test was first performed to verify the normality of the distribution of Δ. Bland-Altman analysis (19) was subsequently used to evaluate the differences for the image derived parameters considered. The mean and standard deviation (SD) and the associated 95% confidence intervals (CI) were obtained. Lower and upper reproducibility limits (LRL and URL), defining the reference range of spontaneous changes, were calculated as ±1.96 × SD provided that the distribution were not statistically different than a normal one. Intraclass correlation coefficients (ICC) were in addition calculated providing an evaluation of the reliability of measurements, whereas their reproducibility was estimated based on their precision (half the width of 95%CI * 100 %). The differences in the calculated reproducibility of the textural feature parameters as a function of the discretization values used in the normalization step was assessed using a paired student t-test. P values of less than 0.05 were considered statistically significant.

RESULTS

For all considered features, Δ showed no significant differences from a normal distribution according to the Kolmogorov-Smirnov test. Consequently, Bland-Altman analysis was performed on all parameters. All of the reproducibility results using the Bland-Altman analysis, including LRL and URL (and associated 95% CI), are provided in table 3 for both intensity histogram parameters and textural features, whereas the ICCs and associated 95% CI and precision are summarized in table 4. As figure 2A and Table 3 show SUV measurements exhibited reproducibility levels in line with previously published studies. A mean difference of 5±20% and associated LRL and URL of -34% and +43% were found for SUV_{max}, and 6±21% mean difference, with -36% LRL and +47% URL for SUV_{mean}. ICC was 0.94 (95% CI: 0.82-0.98; precision ±8%) and 0.92 (95% CI: 0.78-0.97; precision ±10%) for SUV_{max} and SUV_{mean} respectively. Amongst other global tumor heterogeneity characterization parameters derived using the intensity histogram, kurtosis was found to have similar reproducibility as SUV_{max} and SUV_{mean} but a lower ICC (0.80 with 95% CI between 0.44-0.93; precision ±25%; figure 2B). COV (Mean/SD) was characterized by reproducibility limits ranging between -43% and 51% and an ICC of 0.82 (95%CI: 0.49-0.94; precision ±23%). Standard deviation, skewness and minimum intensity had the highest reproducibility limits ranging between -45 and 60%.

Among the local heterogeneity parameters calculated on co-occurrence matrices, the entropy, homogeneity and dissimilarity were characterized by reproducibility limits below 30% and an ICC precision below ±16%, the most reproducible being the entropy, with LRL of -13% and URL of 9% (figure 2C). The other local features (2^{nd} angular moment, contrast and correlation) were characterized by lower reproducibility, with LRL and URL varying between -40.9% and 62.7%, which is comparable with the reproducibility achieved for some of the histogram based parameters such as skewness (LRL-URL between -54.2% and 53.6%) or minimum intensity (LRL-URL between -45.6% and 58.2%). Both the intensity and the size variability of uniform zones identified within the tumor, representing a measure of regional tumor heterogeneity and previously shown as significant predictors of response to therapy,

have shown a better physiological reproducibility with LRL and URL of-56.7% to 37.3% and -34.1% to 56.5% respectively (figure 2D). The respective ICCs for these measurements were 0.97 (95%CI: 0.93-0.99; precision ±3%) and 0.97 (95%CI: 0.91-0.99; precision ±4%). More specifically the SD of the mean percentage difference was 23.1% and 24% for the textural feature parameters related to the size and intensity variability of tumor uniform zones compared to 19.5% and 21.2% in the case of the SUV_{max} and SUV_{mean} respectively. Other regional heterogeneity features were not reproducible, as for example small area emphasis (LRL and URL of -113% and +100%), low-intensity emphasis (LRL and URL of -112% to +104%) and low-intensity small area emphasis (LRL and URL of -140% to +125%).

As illustrated in figure 3A, all of the textural parameters describing local tumor heterogeneity were found to be insensitive to the chosen discretization values. Within this context no statistically significant differences were found for the range of discretization values used (8 to 128) with a mean SD of 5% and 15% for 8 and 128 discretization values respectively. Several of the regional heterogeneity parameters calculated on intensity size-zone matrices were sensitive to the chosen discretization value, with statistically significant differences and SD values twice as high or low with varying discretization, as shown in figure 3B. The large area emphasis feature, for instance, was characterized by a mean difference of 29±79% and 4±30% using 8 and 64 values respectively. On the other hand, the intensity and size variability of uniform tumor areas as well as the high intensity emphasis zones where largely independent (SD differences <20%) of the discretization values with non-statistically significant differences.

DISCUSSION

Predicting and monitoring therapy response with PET imaging is one of the rising applications of this modality. Characterizing intra-tumor heterogeneity of the radiotracer uptake has been identified as a clinically relevant task and requires semi-automatic

validated, accurate, robust and reproducible tools (20). We have recently introduced the use of textural features for the characterization of tumor heterogeneity within the context of predicting tumor response to therapy using FDG PET imaging (8). It is clearly not straightforward to associate each of these heterogeneity features with one specific physiological process within the tumor, particularly in the case of FDG imaging. However, since all these different parameters represent measurements of tumor local and regional tracer uptake heterogeneity, a reasonable assumption is that their quantitation can be related to underlying physiological processes, such as vascularization, perfusion, tumor aggressiveness, or hypoxia (21, 22). All of these processes have been identified as potentially contributing to the way the FDG uptake is spatially distributed within a tumor volume.

A possible clinical significance of tumor uptake heterogeneity patterns can be related to the efficiency of a given treatment regime. One example is in the case of combined chemo-radiotherapy, where the delivery of a uniform radiation dose to a target tumor volume independently of the actual tracer distribution within the tumor may be responsible for possibly explaining failure of treatment (8, 20) Finer characterization of the heterogeneity as obtained through textural features could therefore help identifying potential responders or non responders before initiating treatment or early during treatment by characterizing the evolution of uptake heterogeneity during treatment.

As the features are calculated within a delineated MATV, it is important to reduce the potential variability that could arise from the reproducibility of the tumor volume delineation step. There is indeed a large variability in the reproducibility results observed depending on the segmentation algorithm used. It has been demonstrated that threshold-based delineation may lead to poorly reproducible delineated MATV on double baseline scans (12,13). On the other hand, the use of more sophisticated and robust segmentation algorithms (such as FLAB) has been demonstrated to lead to satisfactory results with similar reproducibility as

265 SUV_{max} (±30%) (13). This delineation method was therefore used in this study in order to
266 minimize the impact of MATV delineation to the textural features reproducibility.

267 The parameters extracted from the intensity histogram characterize the distribution of the
268 voxel intensities without taking into consideration spatial relationships between the voxels.
269 For this reason, the features extracted from the histogram can be denoted as global. The
270 maximum intensity of the histogram, corresponding to the SUV_{max}, had the best
271 reproducibility along with kurtosis and mean SUV with a SD of the mean percentage
272 difference of 19.5%, 18% and 21.2% with an ICC of 0.94, 0.80 and 0.92 respectively. These
273 reproducibility results are similar to these reported on previous reproducibility studies
274 concerning the SUVs measurements. The reproducibility for the other tumor global features,
275 namely the minimum intensity, standard deviation and skewness, was worse with LRL and
276 URL at -54% to 58%, which may compromise their potential for clinical use in order to
277 characterize tumor response or progression.

278 The local heterogeneity features derived from co-occurrence matrices provide far more
279 complex information than the intensity histogram as they are focusing on the relationship
280 between voxels and theirs neighbors at a local scale. Despite this characteristic of being very
281 specific and local parameters, some of these features (entropy, local homogeneity) exhibited
282 even better reproducibility than the SUV_{max}. These tumor local heterogeneity features were
283 previously identified amongst other tumor heterogeneity characteristics as being capable of
284 classifying esophageal cancer patients with high specificity and sensitivity regarding
285 response to combined radiochemotherapy. On the other hand, other local heterogeneity
286 features such as contrast, 2^{nd} angular moment or correlation were characterized by larger
287 reproducibility limits between -40% and 63% (ICC ≥ 0.94). Finally most of the local
288 heterogeneity parameters were found to be robust versus changes in the discretization
289 value.

290 Regarding regional heterogeneity features, several parameters (SAE, LAE, LIE, LISAE,
291 LILAE, HILAE and ZP) were found to be sensitive to the choice of the discretization value.
292 Some of them (particularly SAE, LIE and LISAE) were also found to have poor
293 reproducibility. All of these parameters are focusing on the smaller homogenous and lower
294 intensity regions, which on the one hand are expected to be less reproducible and on the
295 other hand not of the highest interest in terms of characterizing regional tumor FDG uptake
296 heterogeneities. Other regional heterogeneity parameters such as the features characterizing
297 large homogeneous and high intensity tumor regions (LAE, HIE, HILAE) may be more
298 interesting for predicting response to therapy. The high intensity areas, corresponding to high
299 radiotracer uptake regions, are associated to the more aggressive tumor parts. On the other
300 hand, the large homogeneous areas represent more robust tumor characteristics since they
301 are less likely to result from statistical noise or partial volume effects. Among these regional
302 heterogeneity parameters, only the high intensity regions feature exhibit a reproducibility
303 similar to the SUV_{max} (LRL -36% to URL +44%, ICC 0.82), and therefore sufficient to be
304 considered as a parameter of interest for characterizing patient response.

305 Finally, the parameters corresponding to the variability in the size or intensity (SZV and IV
306 respectively) of the homogeneous areas are also good indicators of the regional tumor
307 heterogeneity having already shown potential for patient differentiation in terms of response
308 to therapy. These parameters highlight the repartition of the intensity values or region sizes
309 within the tumor (high tumor heterogeneity corresponding to high variability of the radiotracer
310 distribution, corresponding in turn to high intensity variability). A good reproducibility with a
311 SD of the mean percentage difference of 24% and an ICC of 0.97 (compared to 19.5% for
312 the SUV_{max}) was measured for these regional heterogeneity features.

313 Our study suggests that a careful selection of the parameters to quantify local and regional
314 heterogeneity may provide both a complete and reproducible characterization of the tracer
315 uptake spatial heterogeneity within tumors in FDG PET images. It should be emphasized that
316 these parameters exhibiting the highest reproducibility in this study were also the ones that

were found to be significant predictors of patient response in a previous study (local homogeneity and entropy, intensity variability and size-zone variability) (8).

One of the limitations of the current study is the small sample of patients, which is however of the same size and in line with previously published reproducibility studies (9-11). On the other hand, although our reproducibility results were established on FDG PET images of esophageal cancer lesions, these lesions displayed a large range of sizes and tracer uptake heterogeneity patterns. These results obviously require confirmation for other cancer models and/or radiotracers. Partial volume effects (PVE) were not specifically investigated in this work, although since tumors were all larger than $10cm^3$ and in the same body region, PVE is expected to have a low impact on an inter-patient basis for this dataset as far as the reproducibility evaluation is concerned. On the other hand, PVE correction can be expected to have a potentially more important role on the absolute quantification of the heterogeneity parameters, and therefore the impact of partial volume effects correction within this context will be the focus of further investigations.

Finally, in this study we assumed that a satisfactory reproducibility range for textural features could be considered as ~±30-40% (SD of 15-20%) upper and lower limits. This was chosen accordingly to what was previously defined as reproducibility limits for the use of SUV and tumor metabolic volume measurements. This means that in order to be used for response monitoring purposes, a given parameter has to exhibit higher changes during treatment than its reproducibility range observed in double baseline scans. However, no study has yet to investigate the evolution of textural features on sequential PET scans and the correlation of these changes with therapy response. Such a study will provide an estimation of the range of changes for these parameters between a pre- and post- or early into treatment scans. This range of values, in comparison with the reproducibility limits of the same parameters as established in the present study, would allow evaluating the potential of using these heterogeneity measures within the context of assessing response to therapy with serial FDG PET scans.

CONCLUSIONS

The physiological reproducibility varied significantly among the various tumor heterogeneity features under investigation, only a few of them being identified as reproducible. Based on our results, heterogeneity parameters that should be preferentially considered for tumor heterogeneity characterization since they are the most reproducible include entropy, homogeneity and dissimilarity for local characterization, and variability in the size and intensity of homogeneous tumor areas for regional characterization.

References

1. 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.

2. Cazaentre T, Morschhauser F, Vermandel M et al. Pre-therapy 18F PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. *Eur J Nucl Med Mol Imaging*. 2010;37:494-504.

3. Rizk NP, Tang L, Adusumilli PS, et al. Predictive value of initial PET SUVmax in patients with locally advanced esophageal and gastroesophageal junction adenocarcinoma. *J Thoracic Oncol*. 2009;4:875-879.

4. Leibold T, Akhurst TJ, Chessin DB, et al. Evaluation of (18)F-FDG-PET for early detection of suboptimal response of rectal cancer to preoperative chemoradiotherapy : a prospective analysis. *Ann Surg Oncol.* 2011;18:2783-2789.

5. Shamim SA, Kumar R, Shandal V, et al. FDG PET/CT evaluation of treatment response in patients with recurrent colorectal cancer. *Clin Nucl Med*. 2011;36:11-16.

6. Hatt M, Visvikis D, Albarghach NM, et al. Prognostic value of 18F-FDG PET image-based parameters in oesophageal cancer and Impact of tumour delineation methodology *Eur J Nucl Med Mol Imaging.* 2011;38:1191-1202.

7. El Naqa I, Grigsby P, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit*. 2009;42:1162-1171.

8. Tixier F, Cheze-Le-Rest C, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52:369-378.

9. Weber WA, Ziegler SI, Thodtmann R, Hanauske AR, Schwaiger M. Reproductibility of metabolic measurements in malignant tumors using FDG PET. *J Nucl Med*. 1999;40:1771-1777.

10. Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurement determined by 18F-FDG PET in malignant tumors. *J Nucl Med*. 2008;49:1804-1808.

11. Paquet N, Albert A, Foidart J, Hustinx R. Within patient variability of FDG standardized uptake values in normal tissues. *J Nucl Med*. 2004;45:784-788.

12. Frings V, de Langen AJ, Smit EF, et al. Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. *J Nucl Med.* 2010;51:1870-1877.

13. Hatt M, Cheze-Le Rest C, Aboagye EO, et al., Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements. *J Nucl Med.* 2010;51:1368-1376.

14. 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.

15. Browne J, de Pierro AB. A row-action alternative to the EM algorithm for maximizing likelihood in emission tomography. *IEEE Trans Med Imaging*. 1996; 15:687-699.

16. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE Trans Med Imaging*. 2009; 28(6):881-893.

17. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;3:610-621.

18. Thibault G, Fertil B, Navarro C, Pereira S. Texture indexes and gray level size zone matrix: application to cell nuclei classification. *Pattern Recognition Inf Process.* 2009;140-145.

19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986; 327:307-310.

20. Basu S, Kwee TC, Gatenby R, et al. Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and metastatic

406 sites, a plausible explanation for failed attempts to cure malignant disorders. *Eur J Nucl Med*
407 *Mol Imaging.* 2011; 38:987-991.
408 21. Rajendran JG, Schwartz DL, O'Sullivan J, et al. Tumour hypoxia imaging with 18F
409 fluoromisonidazole positron emission tomography in head and neck cancer. Clin Cancer
410 Res. 2006; 12:5435–5441.
411 22. Kunkel M, Reichert TE, Benz P, et al. Overexpression of Glut-1 and increased glucose
412 metabolism in tumours are associated with a poor prognosis in patients with oral squamous
413 cell carcinoma. Cancer. 2003; 97:1015–1024.
414

Figure captions

Figure 1: Illustration of one of the tumors considered in this study (sagittal slice) for varying discretization values (from 8 to 128 distinct values).

Figure 2: Bland-Altman plots of intensity histogram parameters: SUV_{max} (A) and kurtosis (B); as well as textural features heterogeneity parameters: entropy (C) and size-zone variability (D). Lines show combined mean, 95%CI, as well as upper and lower reproducibility limits

Figure 3: Plots showing the standard deviation of the mean percentage difference as a function of the discretization value for parameters derived from co-occurrences matrices (entropy, dissimilarity, contrast) (A) and intensity size-zone matrices (LISAE: Low-intensity small-area emphasis, SZV: Size-zone variability, ZP: zone percentage) (B).

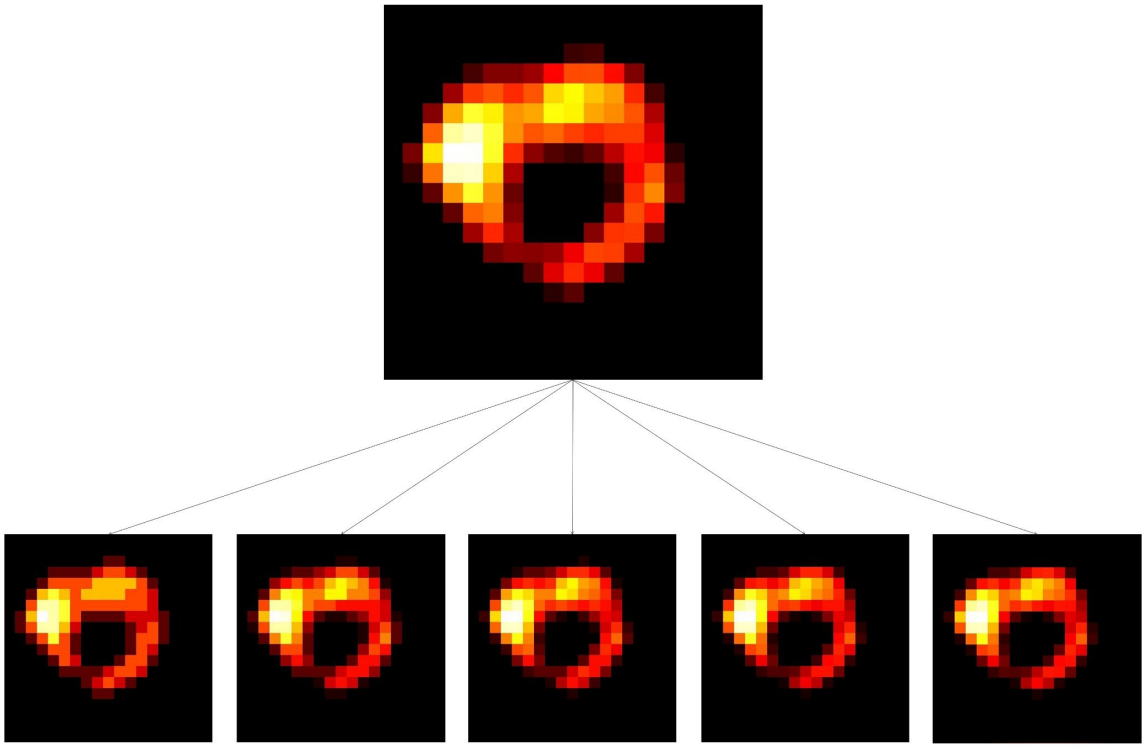


Figure 1

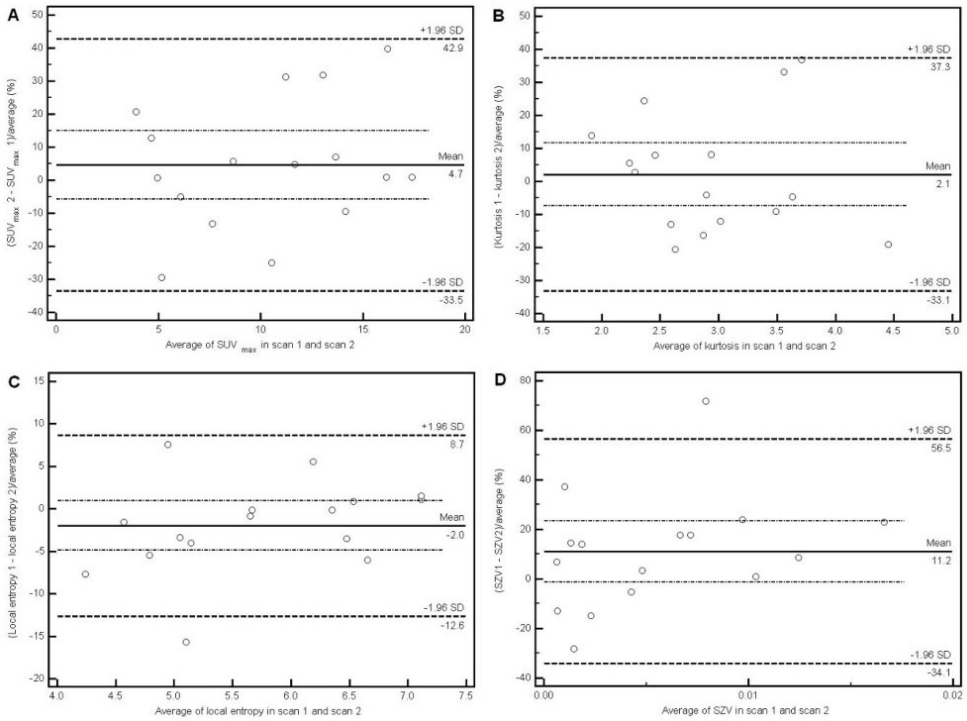
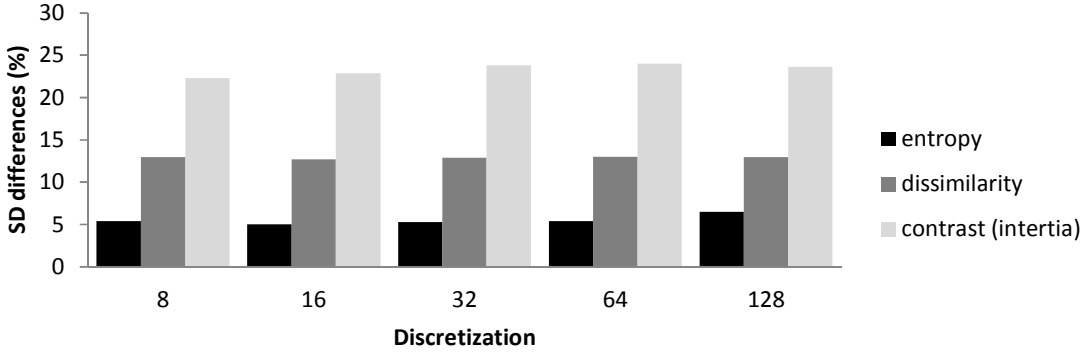

A
+1.96 SD
42.9
Mean
4.7
-1.96 SD
-33.5
(SUVmax 2 - SUVmax 1)/average (%)
Average of SUVmax in scan 1 and scan 2
B
+1.96 SD
37.3
Mean
2.1
-1.96 SD
-33.1
(Kurtosis 1 - kurtosis 2)/average (%)
Average of kurtosis in scan 1 and scan 2
C
+1.96 SD
8.7
Mean
-2.0
-1.96 SD
-12.6
(Local entropy 1 - local entropy 2)/average (%)
Average of local entropy in scan 1 and scan 2
D
+1.96 SD
56.5
Mean
11.2
-1.96 SD
-34.1
(SZV1 - SZV2)/average (%)
Average of SZV in scan 1 and scan 2

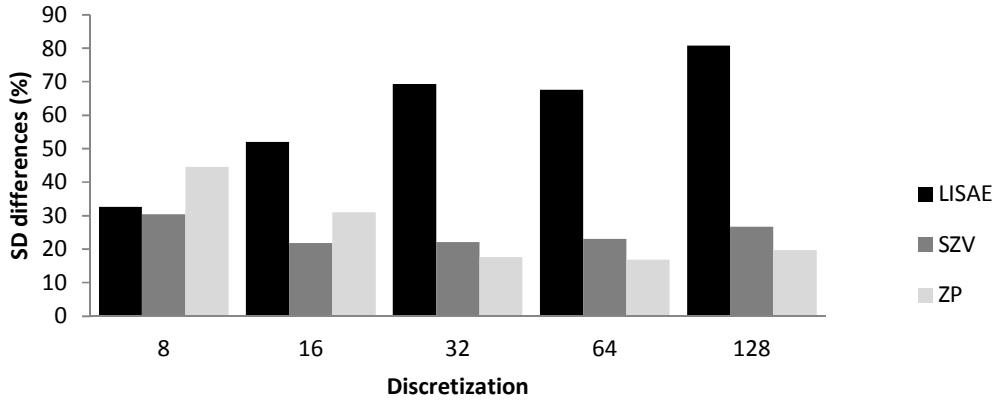
Figure 2

435 **A**



436

437 **B**



438

439 **Figure 3**

440

441 Texture type, scale and associated features

Type	Feature	Scale
Features based on intensity histogram	Minimum intensity	Global
	Maximum intensity (SUV_{max})	
	Mean intensity (SUV_{mean})	
	Variance	
	SD	
	Skewness	
	Kurtosis	
	Mean/SD	
Features based on intensity-size-zone matrix	Small-area emphasis (SAE)	Regional
	Large-area emphasis (LAE)	
	Intensity variability (IV)	
	Size-zone variability (SZV)	
	Zone percentage (ZP)	
	Low-intensity emphasis (LIE)	
	High-intensity emphasis (HIE)	
	Low-intensity small-area emphasis (LISAE)	
	High-intensity small-area emphasis (HISAE)	
	Low-intensity large-area emphasis (LILAE)	
	High-intensity large-area emphasis (HILAE)	
Features based on co-occurrence matrices	Second angular moment	Local
	Contrast (inertia)	
	Entropy	
	Correlation	
	Homogeneity	
	Dissimilarity	

442

443 **Table 1**

444

445 Regional Heterogeneity features formulae

$SAE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{z(i,j)}{j^2}$	$HIE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N} i^2.z(i,j)$
$LAE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N} j^2.z(i,j)$	$LISAE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{z(i,j)}{i^2.j^2}$
$IV = \frac{1}{\Omega}\sum_{i=1}^{M}\left[\sum_{j=1}^{N}\frac{z(i,j)}{i^2}\right]^2$	$HILAE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N} i^2.j^2.z(i,j)$
$SZV = \frac{1}{\Omega}\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\frac{z(i,j)}{i^2}\right]^2$	$LILAE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{j^2.z(i,j)}{i^2}$
$ZP = \Omega \Big/ \sum_{i=1}^{M}\sum_{j=1}^{N} j^2.z(i,j)$	$HILAE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{i^2.z(i,j)}{j^2}$
$LIE = \frac{1}{\Omega}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{z(i,j)}{i^2}$	

Ω : number of homogeneous areas within the tumor
z : intensity size-zone matrix
M : used discretization value
N : size of the largest homogeneous area within the tumor
z(i,j) represents the number of areas with an intensity I and a size j

446

447

448 **Table 2**

449

Reproducibility results for all considered image derived parameters, including SUVs and textural features (calculated using a downsampling range of 64 values).

Texture	Feature	Mean±SD	95% CI	LRL	95% CI for LRL	URL	95% CI for URL
Global	Minimum intensity	6.3 ± 26.5	-7.8 to 20.4	-45.6	-70.2 to -20.9	58.2	33.6 to 82.8
	Maximum intensity (SUV_{max})	4.7 ± 19.5	-5.7 to 15.0	-33.5	-51.7 to -15.4	42.9	24.7 to 61.0
	Mean intensity (SUV_{mean})	5.5 ± 21.2	-5.8 to 16.8	-36.1	-55.8 to 16.4	47.1	27.3 to 66.8
	SD	1.2 ± 23.2	-11.1 to 13.6	-44.18	-65.7 to -22.6	46.6	25.1 to 68.2
	Skewness	-0.3 ± 27.5	-15.0 to 14.3	-54.2	-79.8 to -28.6	53.6	28.0 to 79.2
	Kurtosis	2.1 ± 18.0	-7.4 to 11.7	-33.1	-49.8 to -16.4	37.3	20.6 to 54.0
	Mean/SD	4.1 ± 24.1	-8.8 to 16.9	-43.2	-65.6 to -20.7	51.3	28.9 to 73.7
Local	2nd ang moment	10.9 ± 26.4	-3.2 to 25.0	-40.9	-65.5 to -16.3	62.7	38.1 to 87.3
	Contrast (intertia)	5.4 ± 24.0	-18.1 to 7.4	-52.3	-74.6 to -30.0	41.6	19.3 to 63.9
	Entropy	-2.0 ± 5.4	-4.9 to 0.9	-12.6	-17.7 to -7.6	8.7	3.6 to 13.8
	Correlation	-0.6 ± 27.7	-15.3 to 14.1	-54.8	-15.3 to 14.1	53.6	27.9 to 79.3
	Homogeneity	1.8 ± 11.5	-4.4 to 7.9	-20.8	-31.5 to -10.1	24.4	13.6 to 35.1
	Dissimilarity	-2.1 ± 13.0	-9.0 to 4.9	-27.6	-39.7 to -15.5	23.5	11.4 to 35.6
Regional	Small Area Emphasis (SAE)	-6.0 ± 54.3	-35.0 to 22.9	-112.5	-163.0 to -62.0	100.4	49.9 to 150.9
	Large Area Emphasis (LAE)	3.6 ± 30.0	-12.4 to 19.6	-55.2	-83.1 to -27.3	62.4	34.5 to 90.3
	Intensity Variability (IV)	-9.7 ± 24.0	-22.5 to 3.1	-56.7	-79.0 to -34.4	37.3	15.0 to 59.6
	Size-Zone Variability (SZV)	11.2 ± 23.1	-1.1 to 23.5	-34.1	-55.6 to -12.6	56.5	35.0 to 78.0
	Zone Percentage (ZP)	-2.7 ± 16.9	-11.7 to 6.2	-35.8	-51.5 to -20.1	30.3	14.6 to 46.0
	Low-Intensity Emphasis (LIE)	-4.0 ± 55.3	-33.5 to 25.4	-112.4	-163.9 to -61.0	104.4	525.9 to 155.8
	High-Intensity Emphasis (HIE)	3.9 ± 20.4	-7.0 to 14.8	-36.1	-55.1 to -17.1	44.0	24.9 to 63.0
	Low-Intensity Small Area Emphasis (LISAE)	- 7.0 ± 67.6	-43.1 to 29.0	-139.5	-202.4 to -76.6	125.4	62.5 to 188.3
	High-Intensity Small Area Emphasis (HISAE)	1.0 ± 31.2	-15.6 to 17.6	-60.1	-89.1 to -31.1	62.0	33.0 to 91.0
	Low-Intensity Large Area Emphasis (LILAE)	1.8 ± 28.9	-13.6 to 17.2	-54.9	-81.8 to 28.0	58.5	31.6 to 85.4
	High-Intensity Large Area Emphasis (HILAE)	3.5 ± 35.8	-15.6 to 22.6	-66.7	-100.1 to -33.4	73.7	40.4 to 107.1

Table 3

Reliability of measurements using ICCs (calculated using a downsampling range of 64 values).

Texture	Feature	ICC	95% CI	Precision
Global	Minimum intensity	0.99	0.92 to 0.99	± 4%
	Maximum intensity (SUV_{max})	0.94	0.82 to 0.98	± 8%
	Mean intensity (SUV_{mean})	0.92	0.78 to 0.97	± 10%
	SD	0.99	0.96 to 0.99	± 2%
	Skewness	0.82	0.49 to 0.94	± 23%
	Kurtosis	0.80	0.44 to 0.93	± 25%
	Mean/SD	0.82	0.49 to 0.94	± 23%
Local	2nd ang moment	0.95	0.85 to 0.98	± 7%
	contrast (inertia)	0.94	0.82 to 0.98	± 8%
	Entropy	0.98	0.93 to 0.99	± 3%
	correlation	0.98	0.94 to 0.99	± 3%
	homogeneity	0.88	0.64 to 0.96	± 16%
	dissimilarity	0.93	0.81 to 0.98	± 9%
Regional	Small Area Emphasis (SAE)	0.61	-0.11 to 0.86	± 38%
	Large Area Emphasis (LAE)	0.89	0.70 to 0.96	± 13%
	Intensity Variability (IV)	0.97	0.93 to 0.99	± 3%
	Size-Zone Variability (SZV)	0.97	0.91 to 0.99	± 4%
	Zone Percentage (ZP)	0.84	0.55 to 0.95	± 20%
	Low-Intensity Emphasis (LIE)	0.68	0.08 to 0.89	± 41%
	High-Intensity Emphasis (HIE)	0.82	0.48 to 0.94	± 23%
	Low-Intensity Small Area Emphasis (LISAE)	0.59	-16 to 0.86	± 35%
	High-Intensity Small Area Emphasis (HISAE)	0.83	0.52 to 0.94	± 21%
	Low-Intensity Large Area Emphasis (LILAE)	0.93	0.80 to 0.98	± 9%
	High-Intensity Large Area Emphasis (HILAE)	0.78	0.36 to 0.92	± 28%

Table 4