Reproducibility of ^{18}F-FDG and 3′-Deoxy-3′-^{18}F-Fluorothymidine PET Tumor Volume Measurements

Mathieu Hatt[1], Catherine Cheze-Le Rest[1,2], Eric O. Aboagye[3], Laura M. Kenny[3], Lula Rosso[3], Federico E. Turkheimer[3], Nidal M. Albarghach[1,4], Jean-Philippe Metges[4], Olivier Pradier[1,4], and Dimitris Visvikis[1]

[1]INSERM, U650, LaTIM, CHU Morvan, Brest, France; [2]Academic Department of Nuclear Medicine, CHU Morvan, Brest, France; [3]MRC Clinical Sciences Centre, Imperial College London, Hammersmith Hospital, London, United Kingdom; and [4]Institute of Oncology, CHU Morvan, Brest, France

The objective of this study was to establish the repeatability and reproducibility limits of several volume-related PET image–derived indices—namely tumor volume (TV), mean standardized uptake value, total glycolytic volume (TGV), and total proliferative volume (TPV)—relative to those of maximum standardized uptake value (SUV_{max}), commonly used in clinical practice. **Methods:** Fixed and adaptive thresholding, fuzzy C-means, and fuzzy locally adaptive Bayesian methodology were considered for TV delineation. Double-baseline ^{18}F-FDG (17 lesions, 14 esophageal cancer patients) and 3′-deoxy-3′-^{18}F-fluorothymidine (^{18}F-FLT) (12 lesions, 9 breast cancer patients) PET scans, acquired at a mean interval of 4 d and before any treatment, were used for reproducibility evaluation. The repeatability of each method was evaluated for the same datasets and compared with manual delineation. **Results:** A negligible variability of less than 5% was measured for all segmentation approaches in comparison to manual delineation (5%–35%). SUV_{max} reproducibility levels were similar to others previously reported, with a mean percentage difference of 1.8% ± 16.7% and −0.9% ± 14.9% for the ^{18}F-FDG and ^{18}F-FLT lesions, respectively. The best TV, TGV, and TPV reproducibility limits ranged from −21% to 31% and −30% to 37% for ^{18}F-FDG and ^{18}F-FLT images, respectively, whereas the worst reproducibility limits ranged from −90% to 73% and −68% to 52%, respectively. **Conclusion:** The reproducibility of estimating TV, mean standardized uptake value, and derived TGV and TPV was found to vary among segmentation algorithms. Some differences between ^{18}F-FDG and ^{18}F-FLT scans were observed, mainly because of differences in overall image quality. The smaller reproducibility limits for volume-derived image indices were similar to those for SUV_{max}, suggesting that the use of appropriate delineation tools should allow the determination of tumor functional volumes in PET images in a repeatable and reproducible fashion.

Key Words: oncology; PET; other; delineation; ^{18}F-FDG; ^{18}F-FLT; reproducibility; tumor volume

J Nucl Med 2010; 51:1368–1376
DOI: 10.2967/jnumed.110.078501

Received Apr. 28, 2010; revision accepted Jun. 10, 2010.
For correspondence or reprints contact: Mathieu Hatt, LaTIM, Inserm U650, CHU Morvan, 5 Ave. Foch, 29609 Brest, France.
E-mail: hatt@univ-brest.fr

Most current PET clinical practices for diagnosis, staging, prognosis, therapy-response assessment, and patient follow-up rely on manual and visual analysis (*1*). The index most commonly used in PET clinical studies is the standardized uptake value (SUV). To obtain this index of activity accumulation, a region of interest (ROI) should be determined, usually drawn manually or using some fixed threshold. Although an ROI is not the only factor that can affect the accuracy of SUVs, the type and size of an ROI are large contributors to the variability of such measurements, as has been previously demonstrated (*2,3*). A popular alternative is the use of the pixel with the maximum activity value, usually referred to as the maximum SUV (SUV_{max}). Many studies have demonstrated the prognostic and predictive value of SUV_{max}, despite the fact that it is sensitive to image noise (*4,5*). On the other hand, a few, mostly recent, studies have explored the use of overall tumor volume (TV) as an index for prognosis and response assessment (*6–8*). These studies considered the TV either alone or in combination with the mean SUV (SUV_{mean}), to form the total glycolytic volume (TGV) and total proliferative volume (TPV) (for ^{18}F-FDG and 3′-deoxy-3′-^{18}F-fluorothymidine [^{18}F-FLT], respectively), defined as the product of TV × SUV_{mean} (*9–11*).

The accuracy, robustness, repeatability, and reproducibility of image delineation are directly responsible for the reduced use of functional volumes derived from PET images. On the one hand, manual delineation of functional volumes using PET images leads to high inter- and intraobserver variability (*3*), principally arising from the poor quality of PET images. On the other hand, current state-of-the-art algorithms for functional-volume segmentation consist of fixed- (*12*) or adaptive-threshold approaches (*13,14*). Although fixed-threshold approaches are attractive because of their simplicity, their drawbacks are numerous given that the value of the threshold to be used for each lesion clearly depends on multiple factors, such as lesion contrast and size and image noise (*15*). The solutions based on the use of adaptive thresholding consider the contrast between the object to delineate and its surrounding background. However, adaptive thresholding requires imaging system–specific optimization using uniformly filled spheric lesions, hence reducing the robustness of the approach, particularly in the case of multicenter trials. In addition, this method depends on the background ROI choice, which can in turn lead to reduced interobserver reproducibility for functional-volume determination. A few automatic algorithms have been proposed (*16–19*). The main difference between these algorithms and the threshold-based approaches is that the algorithms automatically estimate the parameters of interest and find the optimal regions' characteristics in a given image, without system-dependent parameters. This technique may reduce issues associated with deterministic approaches based on thresholding, potentially increasing the robustness and reproducibility of PET functional-volume determination (*20*).

Establishing the level of reproducibility and repeatability is essential in the use of any image-derived index in prognostic or therapy-response studies, allowing the evaluation of which change between 2 studies can be considered significant. To date, only a few reproducibility studies (*21–25*), almost exclusively concentrating on SUV_{max} and SUV_{mean} variability in double-baseline ^{18}F-FDG PET scans, have shown a relative absolute percentage difference of up to 13%, with an SD of 10%. The reproducibility of quantitative indices (Patlak influx constant), associated with the acquisition of dynamic datasets, has also been assessed (*21,22*), showing similar levels of reproducibility (mean percentage difference, 8%–10%). Studies on the reproducibility of such indices in the case of ^{18}F-FLT PET have shown that changes larger than 15%–20% and 25%–30% may be considered significant in SUV_{mean} (obtained using a 41% fixed threshold) and SUV_{max} or Patlak influx constant, respectively (*26,27*).

In most of these studies, SUV_{mean} has been calculated using manually drawn ROIs or a single fixed threshold (varying from 40% to 75% of the maximum activity). Among these studies, only 1 has considered the reproducibility of metabolic functional volumes using a fixed threshold. Krak et al. (*3*) have shown a mean percentage difference in the ROI volumes of 23% ± 20% and 55% ± 35% for a fixed threshold of 50% and 75%, respectively. Finally, to our knowledge there has been no published study evaluating the reproducibility of TGV and TPV.

To date, despite numerous studies assessing the accuracy of different segmentation algorithms, there is a lack of evaluation of the repeatability and reproducibility of these algorithms relative to different threshold- and automatic-based delineation approaches. Therefore, the main objective of this study was to assess the repeatability and reproducibility in determining 3-dimensional (3D) functional volumes and associated indices (SUV_{mean}, TGV, and TPV) in PET using different algorithms. The reproducibility of SUV_{max} was also included because it represents the index most used today in clinical practice and facilitates a direct comparison with previous studies. This evaluation was performed on double-baseline ^{18}F-FDG and ^{18}F-FLT clinical PET datasets.

MATERIALS AND METHODS

Segmentation Algorithms Considered

Four approaches were used in this work. Two different fixed thresholds (*12*) were considered, at 42% (T42) and 50% (T50) of the maximum voxel value, using a region-growing algorithm with the maximum-intensity voxel as seed.

An adaptive-threshold method (TSBR, for threshold source–to–background ratio) (*13*) was also included:

$$I_{\text{threshold}} = a + b\frac{1}{\text{SBR}}. \qquad \text{Eq. 1}$$

SBR is the source-to-background ratio, defined as the contrast between a manually defined background ROI and the mean of the maximum-intensity voxel and its 8 surrounding neighbors in the same slice. The parameters a and b are optimized through linear regression analysis for a given scanner using phantom acquisitions of various sphere sizes and contrast.

For automatic-segmentation approaches, the fuzzy C-means (FCM) (*28*) clustering algorithm, with 2 clusters (background and lesion), was considered. This algorithm has been previously used for functional-volume segmentation tasks in both brain and oncology applications (*29,30*) and iteratively minimizes a cost function of the voxel-intensity values to estimate the center of each cluster and membership of each voxel to these clusters. The second automatic algorithm considered was the fuzzy locally adaptive Bayesian (FLAB) (*19*) methodology, based on a combination of statistical models with a fuzzy measure to simultaneously address issues of both noise and blur resulting from partial-volume effects in PET images. FLAB is also able to deal with strongly heterogeneous uptake in tumors of complex shape and generate nonbinary segmented volumes by considering 3 classes and the associated fuzzy transitions (*31*). The parameters required for the segmentation (gaussian mean and variance of each class and spatial priors for each voxel) were estimated using the iterative stochastic expectation maximization procedure. For all approaches, the tumors were delineated after having been isolated in a 3D box of interest previously defined and fixed for all segmentation methodologies (manual and automatic).

Repeatability and Reproducibility: Definitions

Within the context of this study, repeatability is defined as the ability of a given segmentation algorithm to reach the same result regarding the definition of a functional volume when applied multiple times on a single image. In such a task, entirely deterministic fixed-threshold approaches (T42, T50) will always give the same result. On the other hand, more advanced methods—for example, the adaptive thresholding or automatic algorithms such as FCM and FLAB considered here—are susceptible to giving different results when applied multiple times on the same image. The adaptive-threshold segmentation, for instance, depends on a manually drawn background ROI and may thus result in variable delineation depending on the choice of this ROI. On the other hand, FCM and FLAB are iterative procedures that may not converge to the same result at each execution. Finally, manual delineation may be considered as the least repeatable, even when considering a single operator (intraoperator variability). A second aspect considered in this study was the impact of a segmentation algorithm on the reproducibility of determining functional volumes from 2 baseline PET scans.

Two different clinical datasets—comprising esophageal and breast cancer patients scanned with ^{18}F-FDG and ^{18}F-FLT, respectively—were used. In both cases, 2 consecutive PET scans were acquired at an interval of a few days. We therefore studied the differences in derived functional TVs, lesion SUV_{mean}, and TGVs and TPVs extracted from both images. The repeatability of measuring TVs using the various delineation approaches considered in this study was investigated for the same clinical datasets.

Validation Studies

Fourteen whole-body ^{18}F-FDG PET/CT images acquired for patients with esophageal cancer (n = 17 lesions) and nine ^{18}F-FLT PET/CT images acquired for breast cancer patients (n = 12 lesions) were considered. Esophageal cancer patients' images were acquired at 3.4 ± 2.2 d on a PET/CT scanner (Gemini; Philips), with 2-min acquisitions per bed position, 60 min after the ^{18}F-FDG injection (6 MBq/kg). Data were reconstructed using a 3D row-action maximization-likelihood algorithm with standard clinical protocol parameters (2 iterations, relaxation parameter of 0.05, 5 mm in full width at half maximum, 3D gaussian postfiltering). ^{18}F-FLT PET images were acquired for patients with breast cancer (*27*); 2 scans were obtained within 2–7 d (median, 4.1 d) before treatment. All patients received a single bolus intravenous injection of ^{18}F-FLT (153–381 MBq) over 30 s, and dynamic PET was performed for 95 min. Patients were scanned on a PET scanner (ECAT962/HR+; CTI/Siemens), and data were reconstructed using ordered-subset expectation maximization (360 iterations, 6 subsets, no postfiltering).

In both cases, 2 baseline scans were acquired within an average of 3–4 d of each other. Because no treatment was administered between the 2 baseline scans, and considering the short time between the 2 acquisitions, the assumption was that no significant physiologic changes occurred in between the time the scans were obtained. A similar assumption had been previously used in all other studies evaluating the reproducibility and repeatability of different SUV measurements in PET, with double-baseline scans obtained within 5–10 d (*21–25*). Figure 1 shows the 2 baseline scans—1 for an esophageal cancer (Fig. 1A) and 1 for a breast cancer (Fig. 1B) patient.

Analysis

For the repeatability evaluation, the tumors in the first image for each patient were segmented 10 times each with FCM, FLAB, and TSBR. In addition, manual delineation was performed by 2 nuclear medicine experts. More specifically, the 2 experts performed 10 different slice-by-slice manual delineations for the different lesions considered in a randomized fashion, ensuring a minimum of a week between 2 consecutive delineations of the same lesion. All these manual segmentations were performed under the same conditions as those of full-range contrast display. The mean percentage variability and associated SD with respect to the mean segmented volume was computed for each of the lesions and segmentation approaches across the 10 executions and across the 10 manual delineations, to assess the repeatability of the approaches. The repeatability of the manual delineations of the 2 experts were compared separately (intraobserver variability) and with each other (interobserver variability) using intraclass coefficients.

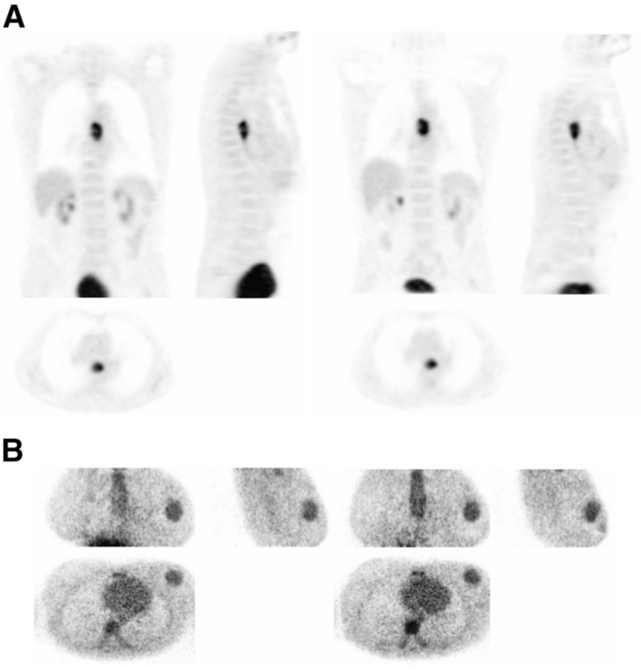

FIGURE 1. Baseline images: ^{18}F-FDG (esophagus) (A) and ^{18}F-FLT (breast) (B).

To study the relative impact of the different segmentation algorithms on the reproducibility of deriving different PET image indices, TVs were segmented independently on both baseline scan images for each lesion, using the different automatic-segmentation approaches. Subsequently, TV (in cm^3), SUV_{mean}, TGV or TPV, and SUV_{max} quantitative values (M) were computed for each delineated lesion and compared between the 2 scans using the mean percentage difference relative to the mean of both baseline scans:

$$(M_{scan2} - M_{scan1}) \Bigg/ \frac{(M_{scan1} + M_{scan2})}{2} \times 100. \qquad \text{Eq. 2}$$

The distribution of the differences between each pair of measurements was assessed for each index using the Kolmogorov–Smirnov test, showing no significant differences from a normal distribution (Fig. 2). Bland–Altman analysis (*32*) was subsequently used to highlight differences between segmentation methodologies. Mean and SD of differences and the respective 95% confidence intervals (CIs) were obtained. To define the reproducibility limits (reference range of spontaneous changes), the 95% CIs for the difference between 2 measurements were computed as the mean difference ± 1.96 times the SD of the difference. To investigate any potential correlations in the measured reproducibility, the magnitude of the percentage difference for the TV, SUV_{max}, and SUV_{mean} measurements was compared with the average of the TVs using the Pearson correlation coefficient r. This analysis was repeated to investigate the correlation of the reproducibility of the different parameters with the SUV_{mean}.

RESULTS

Table 1 contains the mean variability and SD around the mean segmented volume across the 10 manual delineations

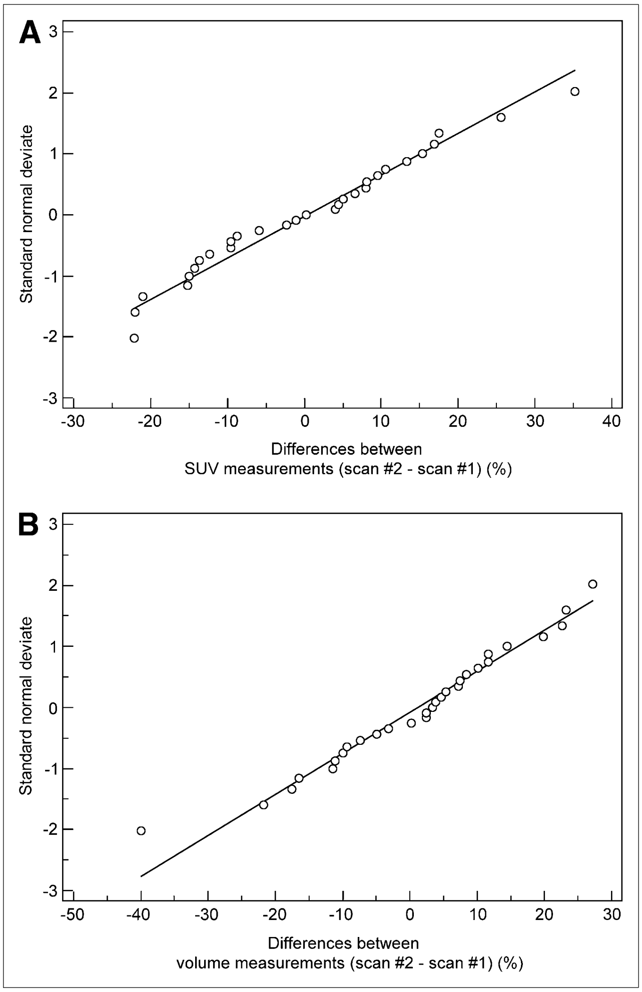

FIGURE 2. Plots showing that distributions of differences for SUV_{mean} (FLAB) (A) and TV (FLAB) (B) between 2 scans were not significantly different from normal.

performed by each of the 2 nuclear medicine experts and 10 repeated executions of the FLAB, FCM, and TSBR algorithms. Results for both clinical datasets are presented separately. FLAB demonstrated highly repeatable results in all of the studied cases, with negligible variability (1%) around the mean segmented 3D volumes across the different repeated executions. FCM also led to satisfactory repeatability results (1.4% ± 1.6% for the ^{18}F-FDG cases and 2.3% ± 1.9% for the ^{18}F-FLT cases). In comparison, the use of the TSBR led to more than twice as high variability (2.9% ± 2.7% and 4.7% ± 3.6% for the ^{18}F-FDG and ^{18}F-FLT cases, respectively). By contrast, manual segmentation by the 2 experts showed high intraobserver variability for ^{18}F-FDG esophageal lesions (14.1% ± 12.1% and 16.4% ± 11.3% for experts 1 and 2, respectively). Interobserver variability was 17.1% ± 14.3%, with an intraclass coefficient of 0.67 (95% CI, 0.39–0.89). In the case of ^{18}F-FLT, this variability was even higher, with an intraobserver variability of 22.1% ± 18.7% and 23.8% ± 17.8% for experts 1 and 2, respectively, and an interobserver variability of 27.4% ± 21.9%, with an intraclass coefficient of 0.59 (95% CI, 0.31–0.84).

Tables 2 and 3 contain a summary of the reproducibility results for the different parameters computed from Bland–Altman plots on the 2 consecutive baseline scans for ^{18}F-FDG esophageal and ^{18}F-FLT breast lesions, respectively. The observed reproducibility of SUV_{max} and SUV_{mean} measurements for the volumes obtained using TSBR and FLAB is illustrated in Figure 3. The corresponding plots for TV are shown in Figures 4A and 4B using TSBR and FLAB, respectively.

Concerning the reproducibility of SUV_{max}, similar percentage differences were measured for the ^{18}F-FDG and ^{18}F-FLT datasets, with an SD of the mean percentage difference of 16.7% and 14.9%, respectively. The upper and lower percentage reproducibility limits for the SUV_{max} were −31% to 35% and −30% to 28% for the ^{18}F-FDG and ^{18}F-FLT datasets, respectively. On the other hand, the automatic approaches led to ^{18}F-FDG TV measurement reproducibility limits of −21% to 31% and −51% to 52% for the FLAB and the FCM algorithms, respectively. A poorer reproducibility of the ^{18}F-FDG TV measurements was observed for the threshold-based approaches, with upper and lower reproducibility limits of −90% to 51% and −69% to 73% for the adaptive and T42, respectively. In the case of ^{18}F-FLT TV measurements, the reproducibility was similar to that of ^{18}F-FDG for the threshold-based approaches, whereas a deterioration in the reproducibility obtained with the automatic approaches was observed, particularly for the FCM algorithm (with reproducibility limits of −66% to 74%).

SUV_{mean} measurements using FLAB exhibited reproducibility levels similar in magnitude to that for the TV definition, with an SD of the mean percentage difference of 15.6% and 14.1% for the ^{18}F-FDG and ^{18}F-FLT datasets, respectively. This was, however, not the case for the other tumor-delineation algorithms considered, with the larger SUV_{mean} reproducibility limits using the FCM tumor definition (−77% to 62% and −59% to 59% for the ^{18}F-FDG and ^{18}F-FLT datasets, respectively). Finally, the smaller SUV_{mean} reproducibility for the threshold-based approaches was obtained using T50 for both the ^{18}F-FDG and the ^{18}F-FLT datasets, with a mean percentage difference of −10.5% ± 23% and −13.3% ± 16.8%, respectively.

The reproducibility of TGV and TPV, being the product of TV and SUV_{mean}, was dependent on the direction of changes for both TV and SUV_{mean}. As an increase of TV was correlated with a decrease of SUV_{mean} and vice versa ($P < 0.002$; $r = 0.54$, 0.67, and 0.72 for FLAB, TSBR, and T42, respectively), TGV and TPV reproducibility levels were generally similar in magnitude to the TV and SUV_{mean} considered separately. However, in certain cases there were more increases or decreases of both TV and SUV_{mean} for a given patient, resulting in larger variability of the TGV and TPV measurements (e.g., the TSBR measurements of the

TABLE 1. Repeatability Evaluation

	Esophageal lesion		Breast lesion	
Method	Mean variability (%)	SD	Mean variability (%)	SD
FLAB	0.6	0.3	1.1	0.7
FCM	1.4	1.6	2.3	1.9
Fixed threshold	0	0	0	0
Adaptive threshold	2.9	2.7	4.7	3.6
Manual delineation (expert 1)	14.1	12.2	22.1	18.7
Manual delineation (expert 2)	16.4	11.3	23.8	17.8
Manual delineation (expert 2 with respect to 1)	17.1	14.3	27.4	21.9

Data are mean variability and SD around mean segmented volume for repeated delineations of 17 esophageal and 12 breast lesions on first baseline ^{18}F-FDG and ^{18}F-FLT scans, respectively.

^{18}F-FLT breast lesions, with 22.1% ± 48.9% for the TPV, whereas TV and SUV_{mean} were 11.3% ± 31.4% and −3.2% ± 26.5%, respectively).

The TV reproducibility results were dependent on the measured TV, with a larger variability seen for smaller tumors. This dependence was statistically significant for the adaptive thresholding ($r = 0.37$, $P = 0.046$; Fig. 5A), with differences higher than 30% on average (≤75%) in several of the tumors below 50 cm^3. On the other hand, this dependence was not significant for FLAB ($r = 0.27$, $P = 0.16$; Fig. 5B), with most differences less than 30%—irrespective of TV—further demonstrating improved robustness, as previously shown (*19,20*). In terms of the SUV_{max} reproducibility results, no statistically significant trend with either the lesion size ($r = 0.016$, $P = 0.93$; Fig. 5C) or the mean of the 2 SUV_{mean} measurements ($r = 0.14$, $P = 0.49$) was observed. Finally, no statistically significant trends were found for the SUV_{mean} reproducibility depending on the lesion size, irrespective of the segmentation algorithm used ($r = 0.2$, $P = 0.3$, and $r = 0.23$, $P = 0.23$, for TSBR and FLAB, respectively).

DISCUSSION

Functional-volume delineation today represents an area of interest for multiple clinical (routine and research) applications of PET (prognosis, response prediction, therapy assessment, radiotherapy treatment planning). In all of these applications, the repeatability and reproducibility with which functional volumes can be determined under different imaging conditions play a predominant role, allowing a level of confidence to be established in the use of such TV measurements. Volume-definition methodologies currently used in clinical practice are based on the use of manual delineation or fixed and adaptive thresholding (*12–14*), whereas several promising automatic algorithms have been proposed (*16–19*). The major drawback of manual delineation is high inter- and intraobserver variability; in addition, the approach is time-consuming. On the other

TABLE 2. Reproducibility Results Using ^{18}F-FDG for Esophageal Lesions

Method	Parameter	Mean ± SD	95% CI	LRL	95% CI for LRL	URL	95% CI for URL
	SUV_{max}	1.8 ± 16.7	−6.8 to 10.4	−30.9	−45.9 to −16	34.6	19.9–49.6
FLAB	TV	5 ± 13.3	−1.8 to 11.9	−21.1	−33 to −9.1	31.1	19.2–43
	SUV_{mean}	0 ± 15.6	−8 to 8	−30.5	−44.4 to −16.6	30.5	16.5–44.4
	TGV	5.1 ± 10.6	−0.4 to 10.5	−15.8	−25.3 to −6.3	25.9	16.4–35.5
FCM	TV	0.4 ± 26.4	−13.2 to 14	−51.4	−75.1 to −27.7	52.2	28.5–75.9
	SUV_{mean}	−7.8 ± 35.5	−26 to 10.5	−77.4	−109.2 to −45.5	61.8	30–93.7
	TGV	−7.4 ± 30.2	−22.9 to 8.2	−66.6	−93.7 to −39.5	51.9	24.8–78.9
TSBR	TV	−19.4 ± 36	−37.9 to −0.9	−89.9	−122.1 to −57.6	51.1	18.9–83.3
	SUV_{mean}	6.3 ± 27.4	−7.8 to 20.4	−47.4	−72 to −22.8	60.1	35.5–84.6
	TGV	−13 ± 28.2	−27.5 to 1.5	−68.2	−93.4 to −42.9	42.2	17–67.4
T42	TV	2.1 ± 36.1	−16.5 to 20.7	−68.7	−101.2 to −36.3	72.9	40.5–105.3
	SUV_{mean}	−10.5 ± 30	−25.9 to 5	−69.3	−96.2 to −42.4	48.4	21.5–75.3
	TGV	−8.4 ± 23.4	−20.5 to 3.6	−54.3	−75.3 to −33.3	37.5	16.5–58.5
T50	TV	0.9 ± 32.9	−16 to 17.8	−63.5	−92.9 to −34	65.3	35.9–94.8
	SUV_{mean}	−10.5 ± 23	−22.6 to 1.6	−56.5	−77.6 to −35.5	35.6	14.5–56.6
	TGV	−9.5 ± 23.1	−21.4 to 2.4	−54.9	−75.6 to 34.1	35.8	15.1–56.6

LRL = lower reproducibility limit; URL = upper reproducibility limit.
Data are percentage differences between scan 2 and scan 1 measurements.

TABLE 3. Reproducibility Results Using ^{18}F-FLT for Breast Lesions

Method	Parameter	Mean ± SD	95% CI	LRL	95% CI for LRL	URL	95% CI for URL
	SUV_{max}	−0.9 ± 14.9	−10.4 to 8.5	−30	−46.6 to −13.4	28.2	11.6–44.8
FLAB	TV	4.3 ± 15.7	−5.7 to 14.3	−26.5	−44.1 to −8.9	35.2	17.6–52.8
	SUV_{mean}	−0.6 ± 14.1	−9.6 to 8.3	−28.2	−44 to −12.5	27	11.2–42.7
	TGV	3.7 ± 17.2	−7.2 to 14.6	−30	−49.2 to −10.8	37.4	18.2–56.6
FCM	TV	4.2 ± 35.7	−18.4 to 26.9	−65.6	−105.5 to −25.8	74.1	34.3–114
	SUV_{mean}	0.3 ± 30.1	−18.8 to 19.4	−58.6	−92.2 to −25	59.2	25.6–92.8
	TGV	4.6 ± 29.8	−14.3 to 23.6	−53.9	−87.2 to −20.5	63.1	29.7–96.4
TSBR	TV	11.3 ± 31.4	−8.7 to 31.2	−50.4	−85.5 to −15.2	72.8	37.7–108
	SUV_{mean}	−3.2 ± 26.5	−20 to 16.6	−55.1	−84.7 to −25.5	48.7	19.1–78.3
	TGV	22.1 ± 48.9	−9 to 53.2	−73.8	−128.5 to −19.1	118	63.3–172.7
T42	TV	9.8 ± 35	−12.4 to 32.1	−58.7	−97.8 to −19.6	78.4	39.3–117.5
	SUV_{mean}	−9.4 ± 20.9	−22.7 to 3.9	−50.3	−73.7 to −27	31.6	8.2–54.9
	TGV	0.7 ± 27.3	−16.7 to 18	−52.8	−83.3 to −22.3	54.1	23.6–84.6
T50	TV	11.2 ± 31.4	−8.8 to 31.1	−50.5	−85.6 to −15.3	72.8	37.6–107.9
	SUV_{mean}	−13 ± 16.8	−24 to −2.7	−46.2	−64.9 to −27.4	19.5	0.8–38.3
	TGV	−1.8 ± 26	−18.4 to 14.7	−52.8	−81.9 to −23.7	49.1	20.1–78.2

LRL = lower reproducibility limit; URL = upper reproducibility limit.
Data are percentage differences between scan 2 and scan 1 measurements.

hand, currently considered state-of-the art adaptive threshold–based algorithms have been shown to accurately define functional volumes under certain imaging conditions of spheric and homogeneous-activity-distribution lesions. However, adaptive-threshold approaches usually involve some user interaction to select background ROIs, which can potentially lead to user-introduced variability. Although signal intensity reproducibility, predominantly considering the use of SUV_{max}, has been previously assessed, the potential of new indices such as TV or TGV and TPV can be considered only after the assessment of their reproducibility, which has not been previously widely assessed. Therefore, in this study the reproducibility limits of these indices, in comparison to other indices considered as the current gold standard, have been assessed using different tumor-delineation methodologies on double-baseline ^{18}F-FDG and ^{18}F-FLT datasets.

In terms of repeatability, all algorithms exhibited mean differences of less than 5%, with automatic approaches coming closer to the perfect repeatability that can be achieved by deterministic approaches such as a fixed threshold. The repeatability of both threshold and automatic-segmentation approaches was superior to that of manual delineation. This should, of course, be considered within the context of the limited absolute accuracy of thresholding, particularly for lesions not homogeneous in form and activity distribution (*31*).

The variability in the SUV_{max} observed in this work is similar to that measured in previous reproducibility studies, with comparable percentage differences for ^{18}F-FDG and ^{18}F-FLT datasets. These percentage differences suggest that differences larger than −30% can be considered as significant in treatment response, whereas changes above 35% (30% for ^{18}F-FLT) may be indicative of no response. Depending on the delineation algorithm used, the mean percentage difference and corresponding SD for TV measured on the 2 baseline scans varied from 5% ± 13% to −19% ± 36% for the ^{18}F-FDG and from 4% ± 16% to 10% ± 35% for the ^{18}F-FLT datasets. The smallest TV reproducibility limits obtained were similar to those for SUV_{max}. These limits ranged from −21% to 31% and −27% to 35% for ^{18}F-FDG and ^{18}F-FLT, respectively, suggesting in turn that, depending on the segmentation algorithm used and similar to SUV_{max}, CIs may be considered for monitoring therapy response based on functional TV. Similarly, in the case of TGV and TPV the smallest reproducibility limits measured were between −16% to 26% and −30% to 37% for ^{18}F-FDG and ^{18}F-FLT, respectively. On the other hand, the largest reproducibility limits for the ^{18}F-FDG TV and TGV ranged from −90% to 73% and from −68% to 52%, respectively.

Reproducibility ranges obtained for the ^{18}F-FDG esophageal lesions were almost systematically smaller than the ones obtained on the ^{18}F-FLT breast lesions—which can be attributed to the higher level of noise and overall lower contrast observed in the ^{18}F-FLT cases, resulting in less robust delineations. In addition, ^{18}F-FDG esophageal lesions tended to appear more homogeneous than breast lesions. For instance, FCM—which incorporates neither noise nor spatial modeling—is associated with a larger mean TV variability of the ^{18}F-FLT dataset relative to ^{18}F-FDG, whereas FLAB exhibited similar reproducibility levels for both. The variability in reproducibility highlights the need for a robust delineation tool ensuring high reproducibility in an environment of substantial image-quality variability—likely, for example, to be encountered in multicenter trials in which the use of functional TV as a measure of response to therapy may be considered.

T50 uses a more restrictive threshold than 42% and is therefore less prone to large overevaluation of low contrast

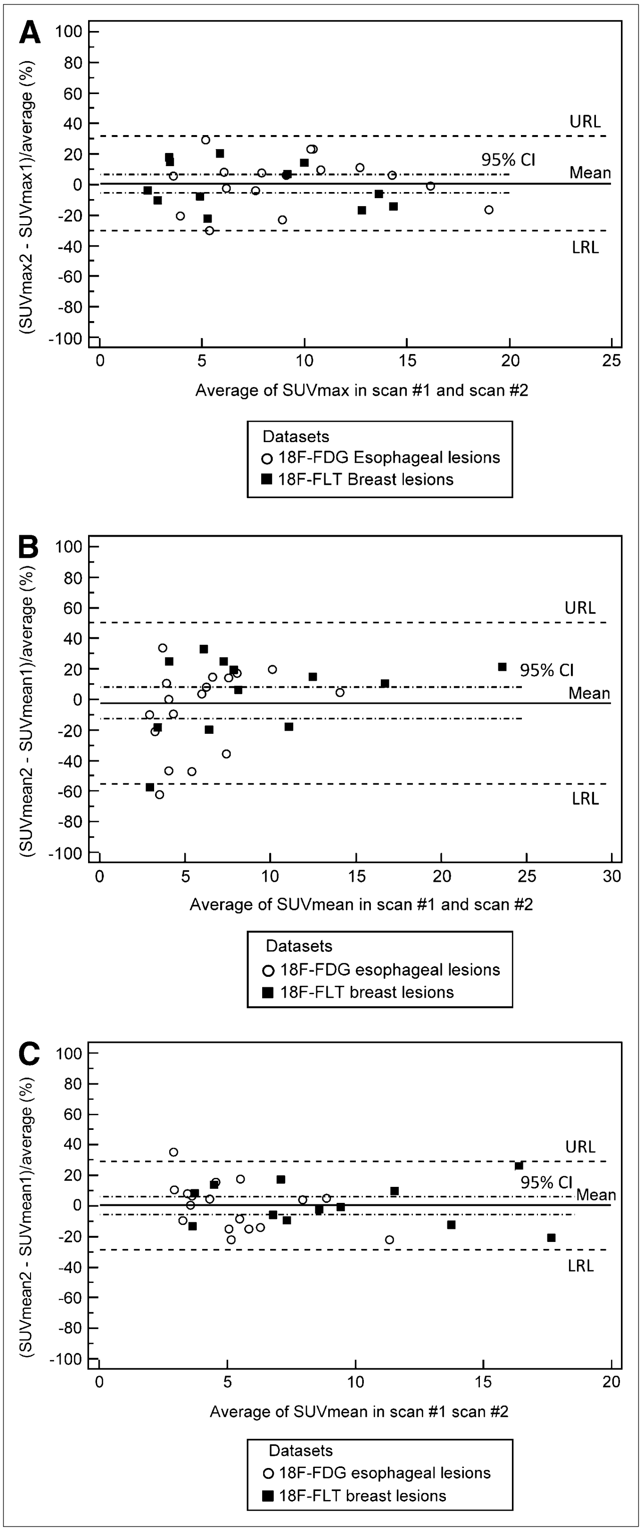

FIGURE 3. Bland–Altman plots of SUV_{max} (A), SUV_{mean} using adaptive thresholding (B), and SUV_{mean} using FLAB (C) for both ^{18}F-FDG and ^{18}F-FLT lesions. Lines show combined mean, 95% CI, and upper and lower reproducibility limits. Individual values for ^{18}F-FDG and ^{18}F-FLT lesions are shown in Tables 2 and 3, respectively. LRL = lower reproducibility limit; URL = upper reproducibility limit.

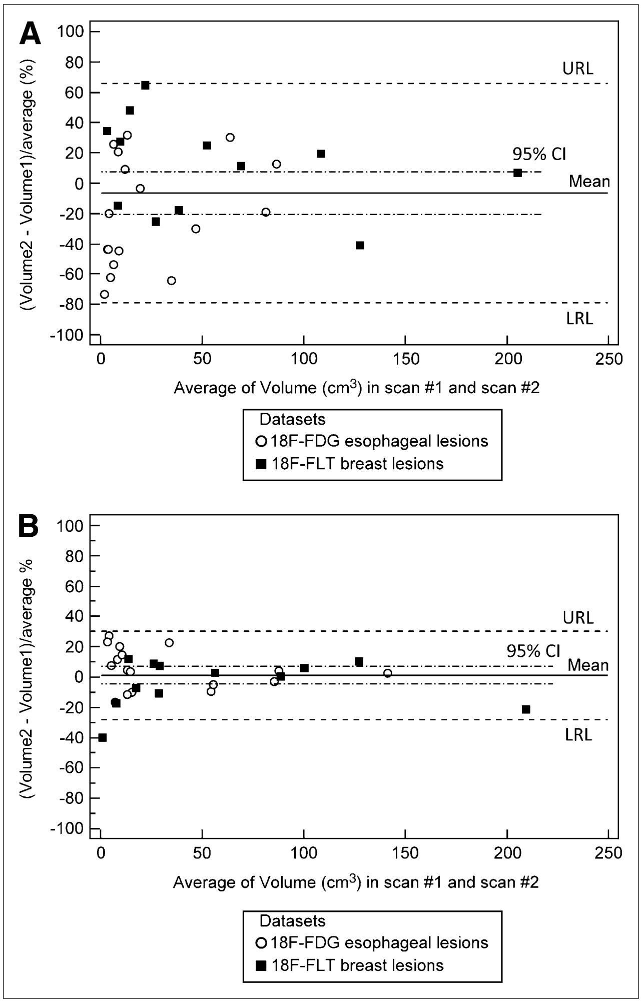

FIGURE 4. Bland–Altman plots of TV using adaptive thresholding (A) and TV using FLAB (B) for both ^{18}F-FDG and ^{18}F-FLT lesions. Lines show combined mean, 95% CI, and upper and lower reproducibility limits. Individual values for ^{18}F-FDG and ^{18}F-FLT lesions are shown in Tables 2 and 3, respectively. LRL = lower reproducibility limit; URL = upper reproducibility limit.

(<4:1) or small-size (<2 cm in diameter) TVs. T50 led to systematically lower variability than T42. Finally, the adaptive-threshold methodology did not demonstrate better reproducibility than did fixed thresholding, which can be attributed to the use of the background ROI placed manually on both scans, combined with the fact that background activity may also vary between the 2 scans.

Although a potential criticism for the current study can be the lack of ground-truth for the functional volumes, the aim of this work was not to assess the absolute accuracy of algorithms, which has been assessed previously for the approaches used in this work (*19,31*). The objective was to assess the reproducibility limits of functional-volume–related indices that can be attained depending on the algorithm. Within this context, the repeated studies of the double-baseline acquisitions have been performed within an

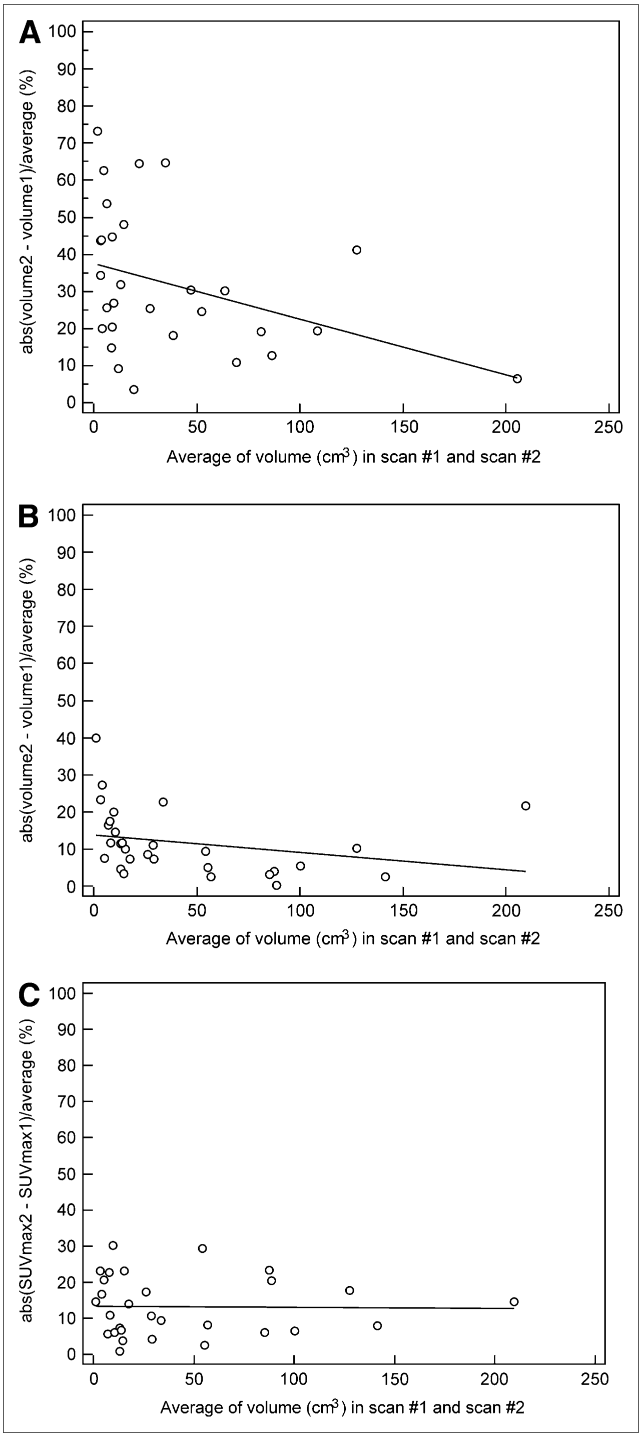

FIGURE 5. Differences between TVs (A and B) and SUV_{max} (C) measured in 2 baseline scans in relation to average TV obtained using adaptive thresholding (A) and FLAB (B and C). abs = absolute.

average of 3–4 d, without any treatment between them, matching the method used by all other reproducibility studies to date (*21–25*). Finally, the reproducibility of SUV_{max} was included in this work as the current gold standard, facilitating at the same time the comparison of our reproducibility study to those performed previously. The SUV_{max} reproducibility limits obtained in this work for both ^{18}F-FDG and ^{18}F-FLT agree closely with those of previous studies.

CONCLUSION

The smaller reproducibility ranges obtained for the different image indices considered in this study, similar to those of SUV_{max}, suggest that new automatic-segmentation approaches may facilitate the introduction of TVs or a combination of TVs and signal intensity in the form of TGVs and TPVs derived from PET images for therapy-response studies. However, our results also demonstrate that the reproducibility of different quantitative parameters associated with functional volumes depends significantly on the delineation approach.

ACKNOWLEDGMENTS

We gratefully acknowledge funding by the Ligue Contre le Cancer (Finistère Committee), French National Research Agency (ANR-08-ETEC-005-01), Cancéropôle Grand Ouest (R05014NG), CR-UK & EPSRC Cancer Imaging Centre (Imperial College, London), U.K. Medical Research Council, and Department of Health (C2536/A10337, U.1200.02.005.00001.01).

REFERENCES

1. Kelloff GJ, Hoffman JM, Johnson B, et al. Progress and promise of FDG PET imaging for cancer patient management and oncologic drug development. *Clin Cancer Res.* 2005;11:2785–2808.
2. Visvikis D, Cheze-Le Rest C, Costa DC, Bomanji J, Gacinovic S, Ell PJ. Influence of OSEM and segmented attenuation correction in the calculation of standardised uptake values for ^{18}FDG-PET. *Eur J Nucl Med Mol Imaging.* 2001; 28:1326–1335.
3. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Imaging.* 2005; 32:294–301.
4. Lucignani G, Larson SM. Doctor, what does my future hold? The prognostic values of FDG-PET in solid tumours. *Eur J Nucl Med Mol Imaging.* 2010;37: 1032–1038.
5. 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.
6. Seol YM, Kwon BR, Song MK, et al. Measurement of tumor volume by PET to evaluate prognosis in patients with head and neck cancer treated by chemo-radiation therapy. *Acta Oncol.* 2010;49:201–208.
7. Chung MK, Jeong HS, Park SG, et al. Metabolic tumor volume of [^{18}F]-fluorodeoxyglucose positron emission tomography/computed tomography predicts short-term outcome to radiotherapy with or without chemotherapy in pharyngeal cancer. *Clin Cancer Res.* 2009;15:5861–5868.
8. Hyun SH, Choi JY, Shim YM, et al. Prognostic value of metabolic tumor volume measured by ^{18}F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. *Ann Surg Oncol.* 2010;17:115–122.
9. Larson SM, Erdi Y, Akhurst T, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET FDG imaging: the visual response score and the change in total lesion glycolysis. *Clin Positron Imaging.* 1999;2:159–171.
10. Francis RJ, Byrne MJ, Van der Schaaf AA, et al. Early prediction of response to chemotherapy and survival in malignant pleural mesothelioma using a novel semiautomated 3-dimensional volume-based analysis of serial ^{18}F-FDG PET scans. *J Nucl Med.* 2007;48:1449–1458.
11. Cazaentre T, Morschhauser F, Vermandel M, et al. Pre-therapy ^{18}F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. *Eur J Nucl Med Mol Imaging.* 2010;37:494–504.

12. Erdi YE, Mawlawi O, Larson SM, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. *Cancer.* 1997;80 (suppl 12):2505–2509.

13. Daisne J-F, Sibomana M, Bol A, et al. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radiother Oncol.* 2003;69:247–250.

14. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of ^{18}F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J Nucl Med.* 2005;46: 1342–1348.

15. Biehl KJ, Kong FM, Dehdashti F, et al. ^{18}F-FDG PET definition of gross tumor volume for radiotherapy of non-small cell lung cancer: is a single standardized uptake value threshold approach appropriate? *J Nucl Med.* 2006;47:1808–1812.

16. El Naqa I, Yang D, Apte A, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. *Med Phys.* 2007;34: 4738–4749.

17. Montgomery DWG, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys.* 2007;34:722–736.

18. Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur J Nucl Med Mol Imaging.* 2007;34:1427–1438.

19. Hatt M, Turzo A, Roux C, et al. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. *IEEE Trans Med Imaging.* 2009;28:881–893.

20. Hatt M, Bailly P, Turzo A, Roux C, Visvikis D. Automatic delineation of functional volumes in PET: a robustness study [abstract]. *J Nucl Med.* 2009;50 (suppl 2):282P.

21. Minn H, Clavo AC, Grenman R, Wahl RL. In vitro comparison of cell proliferation kinetics and uptake of tritiated fluorodeoxyglucose and L-methionine in squamous-cell carcinoma of the head and neck. *J Nucl Med.* 1995;36:252–258.

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

23. Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by ^{18}F-FDG PET in malignant tumors. *J Nucl Med.* 2008;49:1804–1808.

24. 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.

25. Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of ^{18}F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. *J Nucl Med.* 2009;50:1646–1654.

26. De Langen AJ, Klabbers B, Lubberink M, et al. Reproducibility of quantitative 18FLT measurements using positron emission tomography. *Eur J Nucl Med Mol Imaging.* 2009;36:389–395.

27. Kenny L, Coombes RC, Vigushin DM, et al. Imaging early changes in proliferation at 1 week post chemotherapy: a pilot study in breast cancer patients with FLT positron emission tomography. *Eur J Nucl Med Mol Imaging.* 2007;34:1339–1347.

28. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybernet.* 1973;31:32–57.

29. Zhu W, Jiang T. Automation segmentation of PET image for brain tumors. *IEEE Nucl Sci Symp Conf Rec.* 2003;4:2627–2629.

30. Belhassen S, Zaidi H. Segmentation of heterogeneous tumors in PET using a novel fuzzy C-means algorithm [abstract]. *J Nucl Med.* 2009;50(suppl 2):286P.

31. 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.

32. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–310.