Repeatability of Metabolically Active Volume Measurements with ^{18}F-FDG and ^{18}F-FLT PET in Non–Small Cell Lung Cancer

Virginie Frings[1], Adrianus J. de Langen[2], Egbert F. Smit[2], Floris H.P. van Velden[1], Otto S. Hoekstra[1], Harm van Tinteren[3], and Ronald Boellaard[1]

[1]Department of Nuclear Medicine and PET Research, VU University Medical Centre, Amsterdam, The Netherlands; [2]Department of Pulmonary Diseases, VU University Medical Centre, Amsterdam, The Netherlands; and [3]Comprehensive Cancer Centre, VU University Medical Centre, Amsterdam, The Netherlands

In addition to tumor size measurements with CT, there is a need for quantitative measurements of metabolic active volumes, possibly adding to tracer uptake measurements in oncologic response evaluation with PET. The aim of this study was to evaluate the metabolic volume test–retest variability in ^{18}F-FDG and 3′-deoxy-3′-^{18}F-fluorothymidine (^{18}F-FLT) PET studies for various commonly used volumes of interest (VOIs) and the dependence of that variability on lesion size and relative radiotracer uptake. **Methods:** Twenty non–small cell lung cancer patients were scanned twice with ^{18}F-FDG (n = 11) or ^{18}F-FLT (n = 9). VOIs were defined on images reconstructed with normalization- and attenuation-weighted ordered-subset expectation maximization using 4 isocontours (A41%, A50%, and A70% thresholds, adapted for local background, and 50% threshold, uncorrected for background). Statistical analysis comprised intraclass correlation coefficients and Bland–Altman analysis. **Results:** In the ^{18}F-FDG and ^{18}F-FLT groups, 34 and 20 lesions, respectively, were analyzed. Median volumes at the A50% threshold were 3.31 and 2.19 mL (interquartile range, 1.91–8.90 and 1.52–7.27 mL) for ^{18}F-FDG and ^{18}F-FLT, respectively. Intraclass correlation coefficients were greater than 0.9, with the exception of the A70%-based metabolic volumes for ^{18}F-FLT. For lesions greater than 4.2 mL, repeatability coefficients (RCs = 1.96 × SD) of the percentage difference ranged from 22% to 37% for ^{18}F-FDG and from 39% to 73% for ^{18}F-FLT, depending on the VOI method being used. Repeatability was better for larger tumors, but there was no dependence on absolute uptake (standardized uptake value). **Conclusion:** Results indicate that changes of greater than 37% for ^{18}F-FDG and greater than 73% for ^{18}F-FLT (1.96 × SD) for lesions with A50% metabolic volumes greater than 4.2 mL represent a biologic effect. For smaller lesions (A50% VOI < 4.2 mL), an absolute change of 1.0 and 0.9 mL for ^{18}F-FDG and ^{18}F-FLT, respectively, is biologically relevant. Considering the balance between the success rate of automatic tumor delineation and repeatability of metabolic volume, a 50% threshold with correction for local background activity (A50%) seems optimal among the VOI methods evaluated.

Received Mar. 15, 2010; revision accepted Sep. 10, 2010.
For correspondence or reprints contact: Ronald Boellaard, Department of Nuclear Medicine and PET Research, VU University Medical Centre, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands.
E-mail: r.boellaard@vumc.nl

Key Words: positron emission tomography; non–small-cell lung cancer; repeatability; metabolic volume; ^{18}F-FLT; ^{18}F-FDG

J Nucl Med 2010; 51:1870–1877
DOI: 10.2967/jnumed.110.077255

Response to therapy in cancer patients can be monitored with several methods. Traditionally, tumor size measurement with CT is the standard. At present, uptake of ^{18}F-FDG is seen as an investigational tool, just as MRI methods and serum markers (*1,2*). Ideally, robust methodologies enable individual therapy guidance and evaluation of drug efficacy early in the development process.

The recently revised Response Evaluation Criteria in Solid Tumors indicated the potential for PET studies for monitoring disease progression, based on visual assessment (*1*), but also revealed that both widespread standardization and availability are still lacking. Wahl et al. pointed out the opportunities of PET in response evaluation with the capability of imaging metabolic activity (*3*). So far, most PET efforts in this context focused on measurements of tracer uptake. However, metabolic volume measurement might add relevant information because it represents the amount of tumor tissue that is tracer-avid and enables size or volume measurements of viable tumor (*4,5*). In this paper, we will use the term *metabolic volume* to indicate tumor volumes that are derived directly from the PET studies alone, whereas the term *tumor size* refers to CT-based size or volume assessments. In this paper and for the tracers used (^{18}F-FDG and 3′-deoxy-3′-^{18}F-fluorothymidine [^{18}F-FLT]), the term *metabolic volume* may be justified because both tracers are trapped in tissue by metabolic (kinase) activity. Yet, this term should be used with care because volume assessment using PET with other tracers, such as those that bind to receptors or measure perfusion, should not be indicated by metabolic volume.

A major obstacle to the introduction of PET measures into response criteria such as Response Evaluation Criteria in Solid Tumors is the lack of evidence beyond proof of

principle. Because of the relatively small sample size of most observational PET studies, meta-analysis would be beneficial but is hampered by methodologic heterogeneity, especially in acquisition, reconstruction, and data-analysis methods. Awareness that standardization of procedures is a key issue has only recently seemed to grow (*6*).

Repeatability (a function of biologic, technical, and observer variation) is an important feature of response-evaluation tools. Knowledge of normal variation helps to identify the relevant change in parameter value caused by an intervention.

The aim of this study was to evaluate the metabolic volume test–retest reproducibility of 4 frequently used semiautomatic 3-dimensional volume-of-interest (VOI) methodologies in ^{18}F-FDG and ^{18}F-FLT PET studies and the potential dependence of that variability on lesion size and relative uptake.

MATERIALS AND METHODS

Patients

Twenty patients with histology- or cytology-diagnosed non–small cell lung cancer (NSCLC) were included prospectively. All patients signed a written informed consent form in accordance with approval by the institutional review board.

PET

Patients were scanned twice within 7 d (mean, 1.7 d; median, 1 d) before any therapy—11 with ^{18}F-FDG and 9 with ^{18}F-FLT. Repeatability of ^{18}F-FDG and ^{18}F-FLT quantitative methods has been published elsewhere (*7,8*), and for full detail we refer to those publications. Scans were obtained on an ECAT EXACT HR+ scanner (Siemens/CTI). A 10- to 15-min transmission scan was followed by a 60-min emission scan in 2-dimensional mode (*9*). At the start of the dynamic 60-min emission scan, a bolus injection of 370 MBq of ^{18}F-FLT or ^{18}F-FDG in 5 mL of saline was administered through an injector (model H5GPE MCT Plus, 200 mL; Medrad International) at 0.8 mL/s, after which the line was flushed with 42 mL of saline (2.0 mL/s).

The last 15 min of the scan (the last 3 frames of the sinogram) were summed and used for VOI delineation. The images were reconstructed using normalization- and attenuation-weighted ordered-subset expectation maximization with 2 iterations and 16 subsets, followed by postsmoothing of the reconstructed images using a gaussian filter of 5 mm in full width at half maximum (FWHM), resulting in a spatial resolution of approximately 6.5 mm in FWHM near the center of the field of view.

Phantom Experiment: Volumetry

A phantom study using the NU2-2001 image-quality phantom of the National Electrical Manufacturers Association was performed to assess various VOI methods. The phantom background compartment was filled with an ^{18}F-FDG solution (2 kBq/mL). Spheres were filled with an ^{18}F-FDG solution, resulting in sphere-to-background ratios (SBRs) equal to 4.5 and 9 (thus, 2 experiments), covering lesion-to-nontumor ratios, as seen in patient studies. All phantom experiments were performed using the same scanner (HR+) and same reconstruction parameters as applied during patient studies.

In addition to this phantom experiment, a second experiment was performed using the same phantom to assess the repeatability of observed metabolic volumes. The background compartment was filled with an ^{18}F-FDG solution (2 kBq/mL), and the spheres were filled so that an SBR of 9 was obtained. An identical experiment, with an SBR of 4.5 was performed. These phantoms were each scanned 6 times using the same scanner (HR+), procedure, and reconstruction parameters as applied during patient studies. For each of the 6 experiments, the phantoms were positioned at (slightly) different locations in the scanner. In this way, the axial slices and image matrix will cut through or sample the spheres differently during each study, thereby resembling the conditions met during clinical studies.

Image Analysis

For VOI definition, a semiautomatic delineation tool was used, applying predefined thresholds of the maximum voxel value within the tumor (*6,10*). In this study, 4 predefined threshold VOIs consisting of 41%, 50%, and 70% of the maximum voxel value, with correction (adaption) for local background (A41%, A50%, and A70%, respectively), and a 50% uncorrected threshold VOI of the maximum voxel value (50%) were used to define the lesion volume. Rather than showing results for a single (optimal) threshold, we chose to use several VOIs to illustrate the effect of different VOI thresholds on volumetric accuracy, precision, and success rate. The 4 volumes (Fig. 1) were analyzed after a visual check, to ensure that nontumor tissue was not included. All evaluable lesions in the field of view that had adequate focal uptake and were thus delineable with the semiautomatic VOI tool for at least 1 of the studied VOI methods were included.

For the phantom experiment, VOIs equal to those described for the patient study were used. Volume recovery coefficients were obtained by dividing the observed VOIs by the true phantom sphere volumes. Volume recovery coefficients were plotted as a function of true sphere volume and SBR. In addition, the repeatability of observed volumes seen during the second series of experiments will be reported.

Statistical Analysis

The repeatability of the measurements was estimated by calculating the mean and SD of the difference between test and retest values. In addition, the percentage difference was calculated as the absolute value of the difference between test and retest values, divided by the mean of both measurements. For both quantities, the repeatability coefficient (RC) was calculated as 1.96 × SD, as adopted by the British Standards Institution (*11*). Assuming that data are normally distributed with a mean of 0, in 95% of the cases the difference between the 2 measurements will be less than the RC. A *t* test was used to test the null hypothesis that the mean difference between test and retest metabolic volumes is 0. A rejection of this hypothesis would indicate that significant systematic bias and repeatability would not be assessable. To address the clinical question of repeatability across VOIs, both the absolute difference and the percentage difference were plotted against the mean of the test–retest value. The RCs for absolute and percentage differences were calculated for 2 subgroups of lesion VOIs of less and more than 4.2 mL. This volume threshold corresponds by approximation to a diameter of 2 cm (for spheric tumors). The threshold diameter of 2 cm equals about 3 times the spatial resolution of the PET images (FWHM, 6.5 mm near the center of the field of view) below which quantification, VOI definition, and detectability are hampered by partial-volume effects.

The impact of clustered observations (multiple pairs of measurements of different lesions on the same subject) was studied by means of mixed-effects models and, if necessary, corrected (*12,13*).

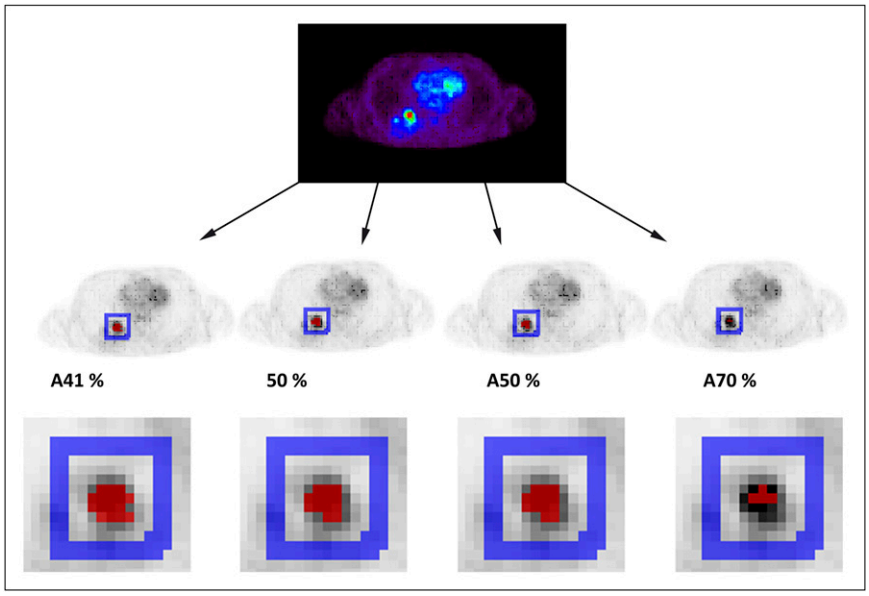

FIGURE 1. Typical example of 4 threshold-defined VOIs for ^{18}F-FDG scan, for which red voxels represent resulting VOI and blue voxels local background, used for background correction.

RESULTS

Phantom Studies

Figure 2 shows the volume recovery coefficients observed for the phantom studies. It was found that an A41% threshold most closely provided true sphere volume for spheres larger than 17 mm in diameter (or 2.6 mL), especially for an SBR of 4.5. Yet, for small spheres all methods, including the A41% threshold, seem not to provide reliable sphere volumes (and thus these points are missing in Fig. 2). When the SBR equaled 9, VOI A41% still provided the most accurate sphere volumes, although some bias up to −20% was observed. Use of higher threshold values (A50%–A70%) obviously results in smaller volume recovery coefficients. Yet, these higher thresholds are included in the patient studies because we hypothesized that relatively low thresholds may be more sensitive to lesion and nontumor uptake heterogeneity.

In Table 1, the coefficient of variation ([COV] %) of observed volumes is given for each of the VOI methods studied. In general, COV increases with smaller VOIs (or with higher VOI thresholds), for lower SBRs and smaller spheres. In Table 2, the SDs of observed volumes are shown. In this case, SD seems to decrease for smaller spheres, but a change of SD with higher-threshold VOI showed a less clear trend.

Patient Studies

The test–retest variabilities of four 3-dimensional VOI methods were analyzed in 20 NSCLC patients (16 men; age range, 45–78 y). Thirty-four lesions were identified in 11 patients scanned with ^{18}F-FDG and 20 lesions in 9 patients scanned with ^{18}F-FLT.

Table 3 shows the feasibility of successful VOI definition as a function of threshold: the A70% threshold was successfully identified for all lesions, whereas A41%, 50%, and A50% thresholds sometimes failed because of low tumor-to-background contrast. An A41% threshold identified 18 of 34 (53%) and 11 of 20 (55%) lesions for ^{18}F-FDG and ^{18}F-FLT, respectively. The 50% VOI identified 53% and 60% of the total lesions and A50% VOI identified 94% and 95%, respectively. Considering (equally weighted) balance between success rate and repeatability of tumor metabolic volume measurements, VOI A50% seems most optimal of the tested VOI methods. This VOI makes little concession

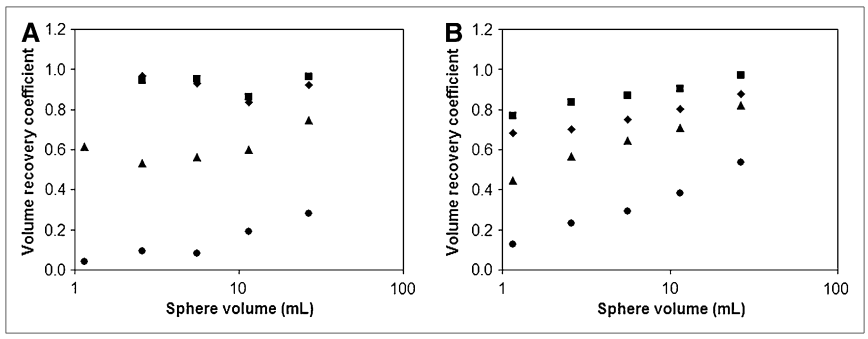

FIGURE 2. Plot of volume recovery coefficients per sphere volume in phantom study for different thresholds (same thresholds applied for patient study) with SBRs of 4.5 (A) and 9 (B). In both A and B, VOI A41%, 50%, A50%, and A70% (upper to lower datasets) are represented by ■, ♦, ▲, and ●, respectively.

TABLE 1
Percentage COV of Observed VOIs as Function of Actual Sphere Volume and VOI Method

Sphere volume	Percentage COV							
	SBR = 9				SBR = 4.5			
	A41	50	A50	A70	A41	50	A50	A70
26.52	2.5	2.8	3.5	14.3	6.8	8.0	8.4	40.6
11.49	5.1	7.3	5.2	6.7	6.5	6.9	4.7	40.8
5.57	7.3	4.2	6.6	16.3	6.1	6.5	8.2	30.2
2.57	9.5	8.1	5.9	23.0	19.2	20.7	13.0	54.7
1.15	17.0	11.6	14.1	24.4	16.4	41.6	18.3	53.6

on lesion detection (32/34 and 19/20 for ^{18}F-FDG and ^{18}F-FLT, respectively) and has sustainable or acceptable repeatability. Higher thresholds will underestimate true volume and have lower repeatability, similar to VOI A70%, whereas lowering the threshold will fail in success rate of lesion detectability, as seen for VOIs A41% and 50%.

The estimated median lesion diameter using the A50% threshold for ^{18}F-FDG and ^{18}F-FLT was 1.85 and 1.61 cm, respectively (interquartile range, 1.54–2.57 and 1.43–2.40, respectively). No statistical difference was observed between the metabolic volumes in the first and second scans for any VOI method.

The absolute and percentage differences with RC are shown in Table 4 and plotted for ^{18}F-FDG A50% and ^{18}F-FLT A50% in Figure 3. RCs of the percentage differences ranged from 44% to 71% for ^{18}F-FDG and 35% to 94% for ^{18}F-FLT, depending on VOI used. For ^{18}F-FDG A50%, for example, the range of percentage differences indicates that the measurement value would have to change more than 62% before one could be confident that the change represented more than measurement variation. There was a trend of increasing percentage RC with higher thresholds.

From Figure 3, it is clear that the percentage difference was to some extent inversely related to VOI in the case of ^{18}F-FDG PET, although this was not seen for ^{18}F-FLT PET, for which it seems to be suitable to use both absolute and relative differences for all VOIs. The RCs for absolute and percentage difference were also calculated for subgroups of

TABLE 2
SD of Observed VOIs as Function of Actual Sphere Volume and VOI Method

Sphere volume	SD							
	SBR = 9				SBR = 4.5			
	A41	50	A50	A70	A41	50	A50	A70
26.52	0.63	0.65	0.73	1.49	1.71	1.84	1.60	2.03
11.49	0.56	0.71	0.44	0.29	0.69	0.67	0.35	0.81
5.57	0.37	0.18	0.24	0.27	0.31	0.30	0.26	0.29
2.57	0.21	0.14	0.08	0.16	0.52	0.54	0.21	0.20
1.15	0.22	0.13	0.11	0.05	0.28	0.87	0.13	0.06

TABLE 3
Feasibility of VOI Definition and Spectrum of VOIs (mL)

Tracer	VOI	Scan	No. of lesions	Q1	Median	Q3	Missing
^{18}F-FDG	A41%	1	18	3.29	5.75	8.87	16
		2	18	2.86	5.59	11.25	16
		All	36	2.96	5.75	9.72	32
	50	1	18	3.17	5.69	7.34	16
		2	18	2.76	5.11	10.16	16
		All	36	2.96	5.69	8.32	32
	A50%	1	32	2.12	3.31	8.53	2
		2	32	1.86	3.28	9.30	2
		All	64	1.91	3.31	8.90	4
	A70%	1	34	0.26	0.58	1.41	0
		2	34	0.26	0.48	1.78	0
		All	68	0.26	0.55	1.43	0
^{18}F-FLT	A41%	1	11	2.06	3.54	12.00	9
		2	11	2.15	2.64	11.05	9
		All	22	2.07	3.44	11.18	18
	50%	1	13	2.70	2.83	10.90	7
		2	12	2.10	3.28	9.70	8
		All	25	2.25	2.83	10.90	15
	A50%	1	20	1.67	2.42	6.62	0
		2	19	1.35	1.86	8.20	1
		All	39	1.54	2.19	7.27	1
	A70%	1	20	0.26	0.45	0.64	0
		2	20	0.26	0.39	0.93	0
		All	40	0.26	0.42	0.66	0

Q1 and Q3 are interquartile ranges.

lesion VOIs of less and more than 4.2 mL. With ^{18}F-FDG, for lesions with VOIs greater than 4.2 mL, the percentage RC was substantially lower than that for lesions with VOIs less than 4.2 mL, for all VOI methods. Similar trends were observed with ^{18}F-FLT. Even though there were fewer ^{18}F-FLT than ^{18}F-FDG lesions, the data suggested that the improvement in RC was smaller for ^{18}F-FLT than for ^{18}F-FDG (Table 4). This is in line with Figure 3, indicating that percentage difference depended less on metabolic volume for ^{18}F-FLT than for ^{18}F-FDG. A possible explanation could be that for small lesions, partial-volume effects may be more pronounced and thus affect the overall accuracy and precision more for ^{18}F-FDG than for ^{18}F-FLT (e.g., because of the higher image contrasts usually seen for ^{18}F-FDG). Yet this finding needs to be further explored. In the case of ^{18}F-FDG studies, for lesions with VOIs less than 4.2 mL absolute RCs equaled 1.3, 1.4, 1.0, and 0.8 mL for the VOIs A41%, 50%, A50%, and A70%, respectively.

In the ^{18}F-FLT group, 1 lesion showed heterogeneous uptake that differed largely between the 2 scans. A subanalysis excluding this 1 heterogeneous lesion (>4.2 mL) improved the metabolic volume reproducibility obtained with an A50% threshold from an RC of 73% to an RC of 16%.

Metabolic volumes derived from VOIs that were based on a relative threshold of the maximum standardized uptake value (SUV) could depend on maximum SUV or mean SUV itself. However, no correlation between metabolic volume

TABLE 4
Absolute Mean, RC, and Percentage Difference with RC for ^{18}F-FDG and ^{18}F-FLT

Radiotracer	VOI	*n* (pairs)	Mean absolute difference (mL)	RC (1.96 × SD) (mL)	Mean percentage difference	RC (1.96 × SD) (%)
^{18}F-FDG						
For total lesions	A41%	18	1.8	5.9	22.6	44.4
	50%	18	2.0	6.0	24.3	52.5
	A50%	32	1.8	4.2	35.5	62.4
	A70%	34	0.4	1.1	43.7	71.1
For lesions < 4.2 mL	A41%	7	0.7	1.3	28.0	47.7
	50%	8	0.6	1.4	25.8	57.7
	A50%	13	0.7	1.0	39.8	64.5
	A70%	31	0.3	0.8	46.2	72.0
For lesions > 4.2 mL	A41%	9	2.5	7.8	12.1	21.9
	50%	9	2.8	7.9	16.2	28.9
	A50%	12	2.9	5.4	21.3	37.2
	A70%	3	1.4	2.2	17.2	36.5
^{18}F-FLT						
For total lesions	A41%	10	1.2	2.2	24.1	34.9
	50%	12	1.0	2.1	21.0	39.4
	A50%	19	1.4	6.3	19.7	50.2
	A70%	20	0.5	0.9	56.8	94.0
For lesions < 4.2 mL	A41%	6	0.7	0.9	29.5	30.9
	50%	7	0.5	0.9	24.4	33.2
	A50%	12	0.3	0.9	16.0	33.2
	A70%	20	0.5	1.5	56.8	94.0
For lesions > 4.2 mL	A41%	4	2.0	2.8	16.1	39.0
	50%	5	1.6	2.8	16.2	49.2
	A50%	7	3.2	9.6	25.9	72.7
	A50%*	6	1.4	2.0	12.2	16.0
	A70%	0	—	—	—	—

*Data, with exclusion of heterogeneous lesion that highly affected outcome.

None of mean percentage differences were significantly different from 0. Percentage differences were calculated by the following formula: |volume scan 1 − volume scan 2|/(0.5 [volume scan 1 + volume scan 2]) × 100.

test–retest and maximum SUV test–retest variabilities was observed (R^2 = 0.0002 and 0.13 for ^{18}F-FDG and ^{18}F-FLT data, respectively). Moreover, no clear correlation was found between absolute or relative metabolic active volume differences and SUV A50% (Fig. 4).

Finally, testing for potential dependency of multiple lesions within patients (compared with treating them as independent observations) did not affect the result of any of the RCs at the level of 3 digits (data not shown).

DISCUSSION

For clinical implementation of any parameter of response assessment, test–retest repeatability has to be known. In this study, we explored 4 currently often-used VOI methodologies in lung cancer. From the array of VOI methods, we prefer semiautomatic delineation for reasons of consistency, lack of observer variability, and practical standpoints (*8*). Many sophisticated and sometimes complicated VOI methods are being developed, in part driven by the demand for radiotherapy planning. Results observed in the phantom studies closely corresponded to those found elsewhere—that is, for lung cancer (^{18}F-FDG) PET studies a threshold close to 41%–50% has been reported to provide accurate tumor volumes (*14*). Obviously, higher thresholds and SBRs will provide smaller measured volumes, as seen both in the phantom and the patient studies.

Change in COV and SD of observed VOIs in the phantom study seem to follow the same trends as seen in the clinical studies. Higher-threshold VOIs provided smaller volumes and larger COVs, as was seen for both the ^{18}F-FDG and the ^{18}F-FLT studies. Moreover, COV seems to worsen for smaller spheres and for lower SBRs. Yet, COV results seemed to be better—that is, a lower COV—than those seen in clinical studies. A possible explanation could be that the uptake in the sphere and background is homogeneous, whereas this is clearly not the case in patient studies. Moreover, day-to-day physiologic variation in uptake may further have affected the clinical repeatabilities.

In the clinical studies, we identified high-percentage RCs for volumetric test–retest variability of all VOI methods when considering all lesions. For lesions greater than 4.2 mL (i.e., >2-cm diameter), a true metabolic volume effect may be identified if the metabolic volume obtained with an A50% threshold changes more than 37% for ^{18}F-FDG and 73% for ^{18}F-FLT. However, a volume change of 37% corresponded to a change in diameter from 2 to 1.7 cm—that

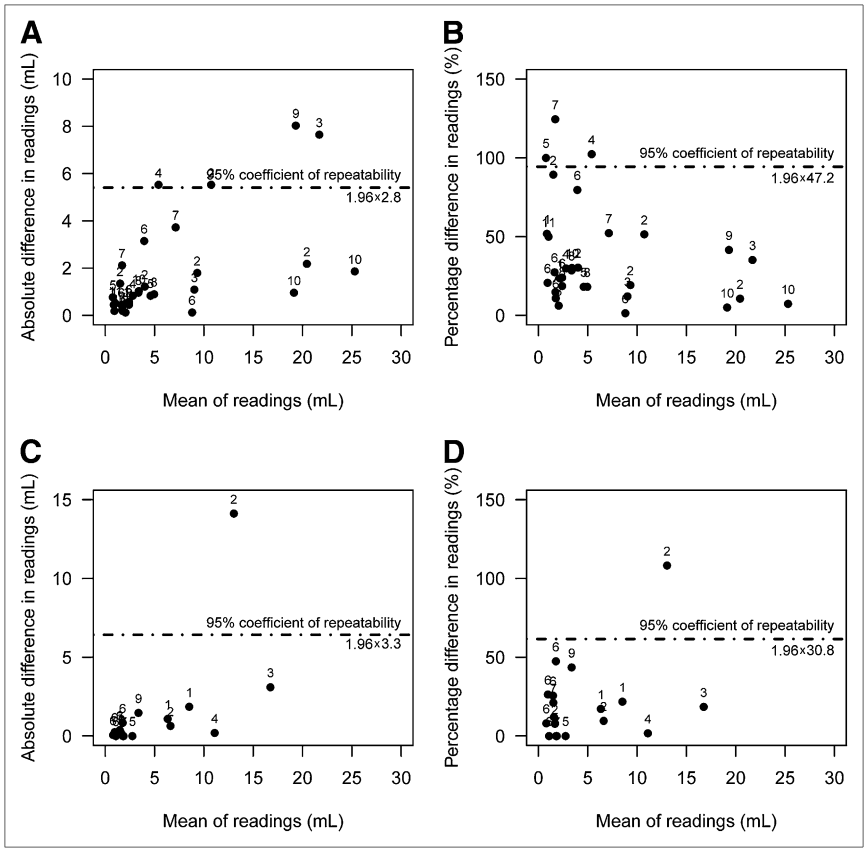

FIGURE 3. Plot of absolute difference between 2 scans against their mean for ^{18}F-FDG (A) and ^{18}F-FLT (C), respectively, and of percentage difference between scans against their mean for ^{18}F-FDG (B) and ^{18}F-FLT (D). Difference is proportional to SD of repeated measurements in each individual. The 95% RC is shown. Numbers near dots indicate patient number. One lesion of patient 7, with mean value of 96 for ^{18}F-FDG, is not shown. Absolute difference for this particular lesion was 0.9 mL or 0.93%.

is, a 15% reduction in diameter for spheric lesions. For metabolic volumes smaller than 4.2 mL, use of absolute rather than percentage difference may be considered. Figure 3 and the phantom data show that high-percentage differences are associated with small (average) metabolic volumes. For lesions less than 4.2 mL, an absolute change in metabolic volume (A50%) of 1.0 mL may reflect a true metabolic volume effect.

Figures 3A and 3C and the phantom data also suggest that the absolute difference between metabolic volumes of test and retest scans increases with (the mean) of the observed metabolic volumes. When more data are available, a (linear or even a nonlinear) relationship between test–retest variability and metabolic volume may be derived. An important issue for response assessment is whether to use absolute or relative (i.e., percentage) changes. Some studies suggest (*3*) the classification of a 30% and 0.8 SUV decrease as a partial response. Moreover, a minimal metabolic volume, lesion size, or SUV at baseline may need to be defined (i.e., when a tumor has minimal volume or uptake, there is not much that can change) to reliably measure response. A minimal metabolic volume threshold at, for example, 4.2 mL or 2-cm diameter, may also be required because small lesions are affected by partial-volume effects, influencing volume and SUV measurement precision.

Several issues mandate further exploration. First, we delineated VOIs over a summed image of 15 min, starting 45 min after injection. Delayed scanning, for example, up to 90 min, could give different results if metabolic volume is a function of the postinjection interval.

Second, we collected data using a PET system, and image characteristics may be different on modern PET/CT systems. These systems can be operated at higher resolutions and sensitivities, which could improve both metabolic volume accuracy and repeatability.

Third, the range of tumor-to-background ratios can differ for different tumor locations and types and radiotracers. For example, the spatial distribution of ^{18}F-fluoromisonidazole may change considerably from day to day (*15*). Although metabolic volume test–retest variability for 2 different radiotracers has now been investigated, test–retest repeatability needs to be determined for each combination of radiotracer, tumor location, and type.

Fourth, the small number of lesions is a limitation of this pilot study. Unfortunately, several issues limit the collection of large datasets to date. First, the burden to patients and oncology and imaging departments usually limits the collection of test–retest studies to small sample sizes. Our observations, therefore, require external validation, especially because the number of observations was limited in some parts of the volume spectrum (e.g., the >4.2-mL group). We suggest that multicenter clinical trials incorporate baseline test–retest studies, not only to provide quality control but also to provide the additional data on precision

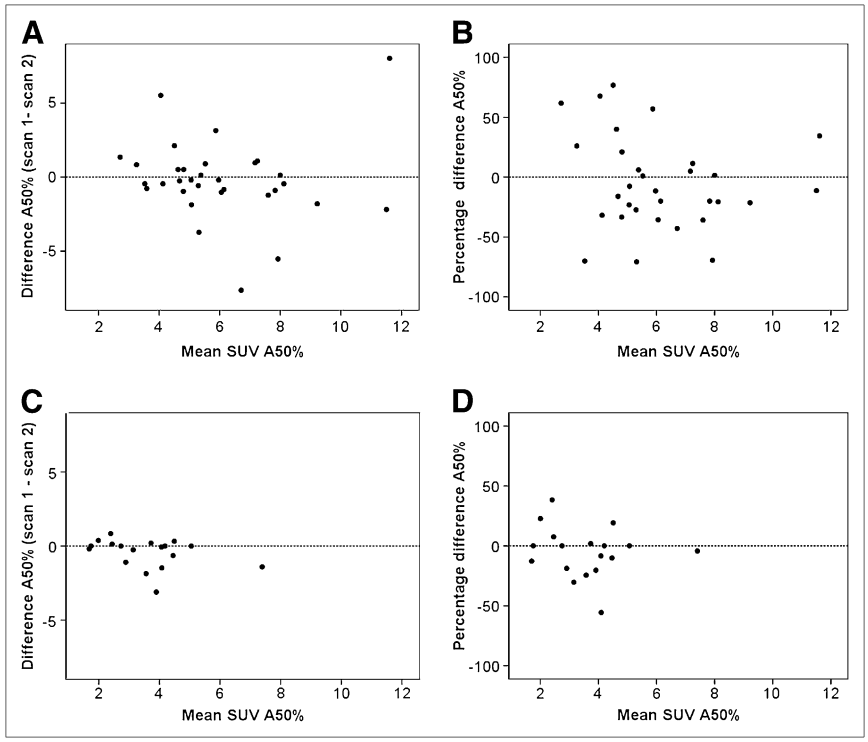

FIGURE 4. Difference in absolute volume measurement for ^{18}F-FDG (A) and ^{18}F-FLT (C), respectively, and percentage difference for ^{18}F-FDG (B) and ^{18}F-FLT (D) against mean SUV A50%.

of metabolic volume quantification required for further qualification of this potentially valuable biomarker. This would nicely fit in the current worldwide attempt to standardize quantitative PET procedures.

More work is also needed to optimize PET volume measurements. The effect of different image characteristics (image resolution and noise) and use of other VOI methods (e.g., gradient-based and iterative) on the accuracy and precision of metabolic volume assessments need to be evaluated. The performance of many VOI methods likely depends on or requires optimization of PET image acquisition. Therefore, it is also important to strive for standardized PET measurements (*6,16*).

Our VOI methods use a relative threshold of the maximum SUV and capture the metabolically most active part of the tumor only. This may be justified when PET is used to assess response to chemotherapy, assuming that the metabolically most active part of the tumor is the most relevant one. In the case of heterogeneous uptake, parts of the tumors will be missed (or oversegmented) using threshold-based methods. Figure 5 shows a lesion with variably heterogeneous uptake, resulting in different VOIs. Therefore, further development of VOI methods that account for radiotracer uptake heterogeneity, along with the development of methods that can describe or quantify intratumoral heterogeneous responses, is needed.

CONCLUSION

This study investigated the test–retest variability of metabolic volume for 2 different radiotracers. For lesions with a metabolic volume (A50%) greater than 4.2 mL, volumetric (3-dimensional) changes of more than 37% for ^{18}F-FDG and more than 73% for ^{18}F-FLT (1.96 × SD) seem necessary to represent a true effect. For smaller lesions (<4.2 mL), an absolute change of 1.0 and 0.9 mL is needed for ^{18}F-FDG and ^{18}F-FLT, respectively. For evaluating the tested VOIs in oncologic response monitoring, these test–retest boundaries should be taken into account. Considering the balance between success rate and repeatability of true tumor volume, using a VOI A50% threshold seems the most optimal and widely available or applicable of the tested VOI methods.

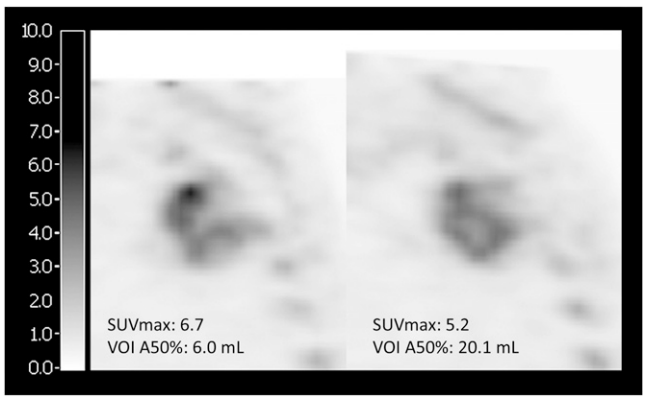

FIGURE 5. Test and retest image of heterogeneous lesion showing variation in uptake pattern resulting in highly different VOIs, implicating limitation of VOI methodology in (variation in uptake in) heterogeneous lesions. SUVmax = maximum SUV.

ACKNOWLEDGMENTS

We thank the patients and their families for participating in this study. In addition, we acknowledge the staff of the Department of Nuclear Medicine and PET Research of the VU University Medical Centre, Amsterdam, The Netherlands, for their help with tracer production and data collection.

REFERENCES

1. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;45: 228–247.
2. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst.* 2000;92:205–216.
3. 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.
4. Humm JL, Lee J, O'Donoghue JA, et al. Changes in FDG tumor uptake during and after fractionated radiation therapy in a rodent tumor xenograft. *Clin Positron Imaging.* 1999;2:289–296.
5. 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.
6. Boellaard R, Oyen WJ, Hoekstra CJ, et al. The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *Eur J Nucl Med Mol Imaging.* 2008;35:2320–2333.
7. de Langen AJ, Klabbers B, Lubberink M, et al. Reproducibility of quantitative ^{18}F-3′-deoxy-3′-fluorothymidine measurements using positron emission tomography. *Eur J Nucl Med Mol Imaging.* 2009;36:389–395.
8. 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.
9. Boellaard R. van LA, van Balen SC, Lammertsma AA. Optimization of attenuation correction for positron emission tomography studies of thorax and pelvis using count-based transmission scans. *Phys Med Biol.* 2004;49:N31–N38.
10. Erdi YE, Mawlawi O, Larson SM, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. *Cancer.* 1997;80(12, suppl):2505–2509.
11. British Standards Institution (BSI). *Precision of Test Methods 1: Guide for the Determination of Repeatability and Reproducibility for a Standard Test Method*. BSI 5497, part 1. London, England: BSI; 1979.
12. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat.* 2007;17:571–582.
13. Daly LE, Bourke GJ. Bias and measurement error. In: *Interpretation and Uses of Medical Statistics*. Oxford, U.K.: Blackwell Science Ltd.; 2000:381–421.
14. Wu K, Ung YC, Hornby J, et al. PET CT thresholds for radiotherapy target definition in non-small-cell lung cancer: how close are we to the pathologic findings? *Int J Radiat Oncol Biol Phys.* 2010;77:699–706.
15. Nehmeh SA, Lee NY, Schroder H, et al. Reproducibility of intratumor distribution of ^{18}F-fluoromisonidazole in head and neck cancer. *Int J Radiat Oncol Biol Phys.* 2008;70:235–242.
16. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med.* 2009;50(suppl 1):11S–20S.