Journal of Nuclear Medicine, published on April 10, 2012 as doi:10.2967/jnumed.111.095299

Repeatability of ^{18}F-FDG Uptake Measurements in Tumors: A Metaanalysis

Adrianus J. de Langen[1], Andrew Vincent[2], Linda M. Velasquez[3], Harm van Tinteren[2], Ronald Boellaard[4], Lalitha K. Shankar[5], Maarten Boers[6], Egbert F. Smit[1], Sigrid Stroobants[7], Wolfgang A. Weber[8], and Otto S. Hoekstra[4]

[1]Department of Pulmonary Diseases, VU University Medical Center, Amsterdam, The Netherlands; [2]Department of Biostatistics, The Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; [3]Bristol-Myers Squibb Co., Princeton, New Jersey; [4]Department of Nuclear Medicine and PET Research, VU University Medical Center, Amsterdam, The Netherlands; [5]Cancer Imaging Program, National Cancer Institute, Bethesda, Maryland; [6]Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands; [7]Department of Nuclear Medicine, University Hospital Antwerpen, Antwerpen, Belgium; and [8]Nuklearmedizinische Klinik, Universitätsklinikum Freiburg, Freiburg, Germany

PET with the glucose analog ^{18}F-FDG is increasingly used to monitor tumor response to therapy. To use quantitative measurements of tumor ^{18}F-FDG uptake for assessment of tumor response, the repeatability of this quantitative metabolic imaging method needs to be established. Therefore, we determined the repeatability of different standardized uptake value (SUV) measurements using the available data. **Methods:** A systematic literature search was performed to identify studies addressing ^{18}F-FDG repeatability in malignant tumors. The level of agreement between test and retest values of 2 PET uptake measures, maximum SUV (SUV_{max}) and mean SUV (SUV_{mean}), was assessed with the coefficient of repeatability using generalized linear mixed-effects models. In addition, the influence of tumor volume on repeatability was assessed. Principal component transformation was used to compare the reproducibility of the 2 different uptake measures. **Results:** Five cohorts were identified for this metaanalysis. For SUV_{max} and SUV_{mean}, datasets of 86 and 102 patients, respectively, were available. Percentage repeatability is a function of the level of uptake. SUV_{mean} had the best repeatability characteristics; for serial PET scans, a threshold of a combination of 20% as well as 1.2 SUV_{mean} units was most appropriate. After adjusting for uptake rate, tumor volume had minimal influence on repeatability. **Conclusion:** SUV_{mean} had better repeatability performance than SUV_{max}. Both measures showed poor repeatability for lesions with low ^{18}F-FDG uptake. We recommend the evaluation of biologic effects in PET by reporting a combination of minimal relative and absolute changes to account for test–retest variability.

Key Words: ^{18}F-FDG, PET; repeatability; cancer

J Nucl Med 2012; 53:1–8
DOI: 10.2967/jnumed.111.095299

PET with ^{18}F-FDG has gained an important role in the clinical setting to detect and stage malignancies and assess treatment response (*1–6*). In the research setting, PET is increasingly being used to study early changes of biologic effects during and after anticancer treatment (*7–10*). The noninvasive nature of PET allows multiple serial measurements without interfering with biologic processes within the tumor and might obviate more invasive procedures, such as biopsy.

Even though PET clinical practice is still dominated by qualitative (visual) image analysis, several potential indications require quantification, for example, when prognostic and predictive information is required beyond the level of TNM staging. A decrease of ^{18}F-FDG uptake after therapy is associated with favorable clinical outcome (*1–6,11*). However, to date only qualitative, and not quantitative, PET measures have been incorporated in response classification systems, in solid tumors and lymphoma (*12,13*).

The European Organization for Research and Treatment of Cancer PET Study Group published recommendations, as far back as 1999, for response monitoring using quantitative PET data to promote consistency in the reporting of studies (*14*). The proposed system was based on the results of both drug evaluation and a few repeatability studies.

To discriminate true signal change from noise and to be able to stratify patients on the basis of changes in ^{18}F-FDG uptake values, the repeatability of the measurement and the error of the determination need to be known. The present metaanalysis determined the repeatability of different SUV measurements using the available data and evaluated potential sources of heterogeneity.

MATERIALS AND METHODS

Study Design

We performed a systematic literature search of Medline and Embase databases to identify studies addressing ^{18}F-FDG repeatability in malignant tumors using the following search terms: PET, FDG, repeatability, and test-retest. Additionally, extensive cross-

Received Jul. 6, 2011; revision accepted Jan. 3, 2012.
For correspondence or reprints contact: Adrianus J. de Langen, Department of Pulmonary Diseases, VU University Medical Center, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands.
E-mail: j.delangen@vumc.nl
Published online ■■■■.

referencing was done, review articles were screened, and experts in the field were consulted. Studies were included when the following criteria were met: the study assessed the repeatability with ^{18}F-FDG PET in malignant tumors, it used standardized uptake values (SUVs), it used uniform acquisition and reconstruction protocols, and it applied the same scanner for the test and retest scan for each patient (i.e., no within-patient scanner variation).

For dynamic studies, SUVs were calculated using the last frame of the dynamic acquisition. Because the method of tumor delineation (e.g., maximum pixel value or threshold-based or fixed-diameter volume of interest [VOI]) can affect ^{18}F-FDG uptake measures, we attempted to obtain uniformly defined tumor volumes between studies for each uptake measure. This volume was defined by a 3-dimensional threshold-based volume (isocontour-defined, with a cutoff of 50% of the maximum ^{18}F-FDG concentration within the tumor). If we were unable to extract the required data from the original publications, authors were asked to provide the required data or to reanalyze their data with the isocontour technique (using software developed in-house and provided by us).

Statistical Analysis

For both PET uptake measures, maximum SUV (SUV_{max}) and mean SUV (SUV_{mean}), the level of agreement between test and retest measurements was assessed using the intraclass correlation coefficient. The intraclass correlation coefficient was calculated using a random-effects model with random intercepts for published study, patient, and tumor location. Kruskal–Wallis tests were applied to the uptake measures to assess systematic bias between studies. One-sample Anderson–Darling tests were used to assess the distribution of the means of the test and retest observations.

Variance–Mean and Variance–Volume Relations

To assess repeatability, we determined the relation between the mean and variance of the test and retest scans, where the variance in the test–retest measurements is assumed to be due solely to measurement error. To account for differences in this relation between published studies, we assessed this association using generalized linear mixed-effects models (with published study as a random factor) and generalized linear models (with published study as a fixed effect). The outcome variable was the square of the difference in test–retest measurements, with the log-transformed test–retest mean as a fixed effect. Given the assumption of normality, the square of the difference is χ^2-distributed, and therefore γ-error distributions were used (*15*). The log-link function was used to relate the estimated variance to the test–retest mean, resulting in an allometric mean–variance relation. Differences between published studies were assessed using the generalized linear models with published study as a fixed effect. The influence of tumor volume on ^{18}F-FDG uptake test–retest repeatability was assessed by including the log-transformed tumor volumes as fixed effects. To avoid extrapolation, we limited the estimated variance–mean associations to values between the fifth smallest and fifth largest observed PET measurement values.

Coefficient of 95% (CR95) Repeatability

Once the variance–mean relation had been estimated, the relation between the CR95 repeatability and the mean was calculated as 1.96 times the SD (*16,17*). The CR95 is the variation solely due to measurement error. If the difference of 2 measurements exceeds the CR95, then this difference is 95% likely to be due to a true change in tumor ^{18}F-FDG uptake rate and not measurement error.

Single PET Observations

The test–retest CR95 corresponds to a 2-observation setting, such as serial baseline and posttreatment assessments. In some clinical trials, however, a single PET scan result is used, for example, for patient stratification during randomization. The equivalent single-observation CR95 can be calculated by dividing the test–retest CR95 by $\sqrt{2}$ (i.e., half the variance), with the assumption that all measurements are performed on the same PET scanner using the same acquisition and reconstruction protocols.

Comparison of Different Quantitative ^{18}F-FDG Measures

To compare the repeatability of the two ^{18}F-FDG uptake measurements, a principal component analysis was performed using data from the 4 studies that analyzed both PET measures. This analysis resulted in a transformation of each of the measures onto the first principal component and thus allowed their CR95s to be compared on the same scale. The significance level of all tests was set at 0.05. The generalized linear mixed-effects models were fitted using the GLIMMIX procedure in SAS (SAS Institute Inc.). All other analyses were performed in R (version 2.9.2; The R Foundation for Statistical Computing, 2009).

RESULTS

Eight repeatability studies were identified (*16,18–24*). One study was excluded because of the use of different scanners for the test and retest scan (*24*). All authors were contacted to provide patient-based data on tumor volume, location, and ^{18}F-FDG uptake. The study characteristics are summarized in Table 1. Nakamoto et al. (*22*) and Krak et al. [**Table 1**] (*19*) were extended analyses of cohorts that were originally described and published by Minn et al. (*20*) and Hoekstra et al. (*18*), respectively. Thus, 5 cohorts were available for this metaanalysis. All repeated scans were obtained using the same scanner as previously for each patient. Three studies used a dynamic scan protocol (*16,18,20*), and the remaining 2 applied a static protocol (*21,23*). All but 1 (*23*) were single-center studies. Tumor types were gastrointestinal (*23*), lung (*18,20*), and miscellaneous primary, predominantly located in the chest (*16,21*). Two studies (*21,23*) used integrated PET/CT scanners; the others a PET-only system. Nahmias et al. (*21*) applied a 90-min time interval between radiotracer injection and scanning, whereas all other studies applied a 60-min time interval. The multicenter study by Velasquez et al. (*23*) reported on 2 datasets, 1 before and 1 after quality assurance assessment. The latter dataset was used for the present study. Finally, Minn et al. (*20*) excluded lesions with a diameter lower than 2 cm, whereas the other studies had no clear restrictions regarding lesion size.

Repeatability of SUVs

SUV_{max}. The 4 available datasets for SUV_{max} (*18,20,21,23*) included 163 tumor lesions in 86 patients. The intraclass correlation coefficient of SUV_{max} was 0.90. The goodness-of-fit tests for normality indicated that SUV_{max} on the original

TABLE 1
Individual Study Characteristics

Characteristic	Study					
			Nahmias			
	Hoekstra	Minn	50% isocontour	Manual	Velasquez	Weber
Patients (*n*)	10	10	21	21	45	16
Tumors (*n*)	27	10	21	21	105	50
Tumors per patient						
Median	2	1	1	1	3	2.5
Range	1–7	1–1	1–1	1–1	1–4	1–8
Time* (min)	60	60	90	90	60	70
Bed position time (min)	Dynamic scan	Dynamic scan	3	3	Not available	Dynamic scan
Time (days)†						
Mean	1	1.8	2.8	2.8	4.1	3
Maximum	1	7	5	5	7	10
Scanner (no. of patients)						
PET alone	10	10	0	0	7	16
PET/CT	0	0	21	21	38	0
Location (no. of patients)						
Thoracic	10	10	19	19	0	13
Abdominal	0	0	2	2	45	3
Tumor volume (cm^3)						
Median	6.2	42.6	4.9	4.9	6.4	5.1
Range	0.7–111.4	18.6–231.0	1.0–79.4	1.0–79.4	0.4–491.4	0.6–86.9
Threshold technique						
4 × 4 voxels around the maximum		X				
50% of the maximum voxel	X		X			X
70% of the maximum voxel					X	
Manual delineation				X		
SUV_{mean}‡						
Median	5.5	8.1	6.6	5.1	6.8	4.5
Range	2.3–11.3	3.9–16.4	2.6–17.3	1.6–17.4	2.5–24.4	1.3–10.5
SUV_{max}§						
Median	8.3	9.2	10	8.9	7.3	
Range	3.8–18.5	4.6–19.5	4.1–24.1	4.0–23.8	2.5–29.7	

*Time between ^{18}F-FDG injection and start of scan for static scans and total scan time for dynamic scans.
†Time between test and retest scan.
‡SUV_{mean} was lower in Weber study than in other 4 studies ($P = 0.0006$).
§SUV_{max} was lower in Velasquez study than in other 3 studies ($P = 0.04$).

scale was significantly different from normal ($P < 0.0001$). After log-transformation, the distribution was normalized [**Fig. 1**] ($P = 0.37$). Figure 1A presents the overall and study-specific relations between the mean and the SD (i.e., variance) of the test and retest scans. The difference in test–retest variability between studies was not significant ($P = 0.25$). Test–retest variability in absolute terms increases with higher SUV_{max} (Figs. 1A and 1B), whereas the percentage change decreases with higher SUV_{max} (Fig. 1C). As a consequence, it is not possible to generate just 1 value that defines the limits of test–retest variability for the full range of SUV_{max}; however, this dilemma can be overcome by combining an absolute and a percentage change of SUV_{max}. From Figure 1C, these values can be extracted. The required absolute difference is calculated as the relative difference multiplied by the mean value where the CR95 curve crosses the relative difference value. For SUV_{max}, the 30% relative difference crosses the CR95 at an SUV_{max} of 6.7. Thus, the combination of an absolute change in SUV_{max} of more than 2 units (0.3 × 6.7) and a relative change of more than 30% is more than 95% likely to be a true change rather than a measurement error, for any SUV_{max} that was encountered in this study. From the same figure, it can be calculated that a change of more than 3.1 units and 25% relative change also exceeds 95% test–retest variability. The combinations of the required minimal absolute and percentage changes that are needed to exceed 95% test–retest variability for the SUV parameters are presented in Table 2. [**Table 2**]

SUV_{mean}. The 5 available datasets for SUV_{mean} (*16,18,20,21,23*) included 213 tumor lesions in 102 patients. The tumor delineation methods differed between the 5 available studies. Hoekstra et al. (*18*) and Weber et al.

jnm095299-pm ■ 4/10/12

RGB

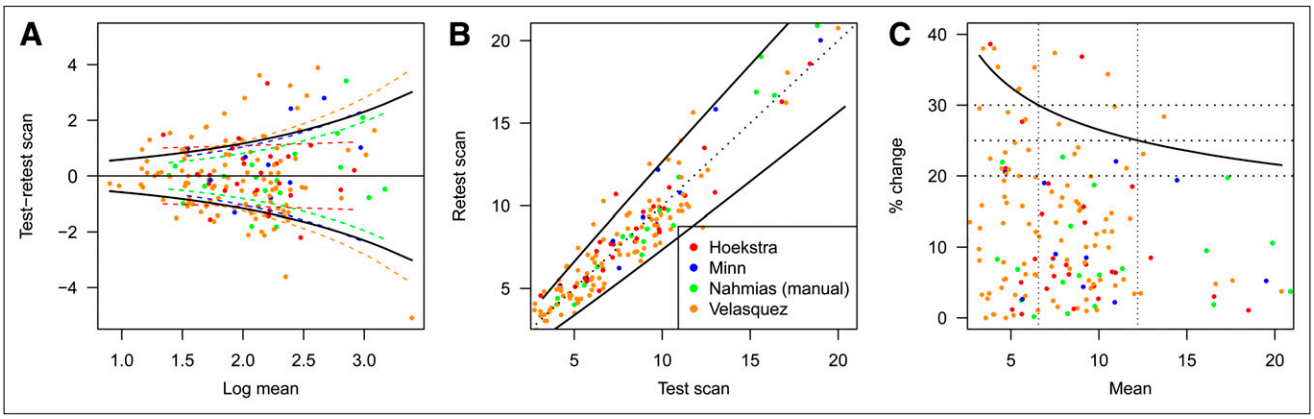

FIGURE 1. Repeatability of SUV_{max}. (A) Estimated study-specific SD (colored dashed lines; study as fixed effect) and overall SD (black solid lines; study as random effect). (B) Test and retest scan values plotted on original scale. Solid line is coefficient of repeatability (CR95). (C) Relation between CR95, as percentage change, with level of SUV_{max}. Dotted lines indicate absolute and relative differences, as presented in Table 2.

(*16*) delineated isocontour-defined volumes with a 50% threshold of the maximum voxel value. Nahmias et al. (*21*) originally published manually defined volumes but reanalyzed their data on our request using the 50% isocontour technique. Velasquez et al. (*23*) used a 70% threshold technique because it was not possible to define 50% volumes without contaminating the tumor VOI with background tissue (because of ^{18}F-FDG uptake heterogeneity and a relatively low tumor-to-background contrast). Minn et al. (*20*) delineated tumor volumes as 4×4 voxels around the voxel with maximum ^{18}F-FDG uptake.

The intraclass correlation coefficient of SUV_{mean} was 0.91. The goodness-of-fit tests for normality indicated that SUV_{mean} on the original scale was significantly different from normal ($P = 0.002$). After log-transformation this was normalized ($P = 0.77$).

Figure 2A presents the overall and study-specific relations between the mean and the SD (i.e., variance) of the test and retest scans. These associations differed significantly across studies ($P < 0.0001$), with the greatest variance observed in the study of Velasquez et al. (*23*) and the lowest in the study of Nahmias et al. (*21*) (manual VOI delineation). When only studies applying the 50% isocontour technique were included (Hoekstra et al. (*18*), Weber et al. (*16*), and 50% isocontour data of Nahmias et al. (*21*)), no difference in test–retest variability was apparent between the studies ($P = 0.13$; Fig. 3). In the dataset of Nahmias et al. (*21*), the manual delineation method performed better in terms of repeatability than the automated method ($P = 0.0001$). [**Fig. 2**] [**Fig. 3**]

Figure 2 for SUV_{mean} shows the same trend as was observed for SUV_{max}. The absolute change in test–retest variability increases with higher SUV_{mean} (Figs. 2A and 2B), whereas the percentage change decreases with higher SUV_{mean} (Fig. 2C). As a consequence, a combination of an absolute and a percentage change is necessary to cover the limits of test–retest variability for the full range of SUV_{mean}. From Figure 2, it can be extracted that the combined change in SUV_{mean} of more than 1.2 units and 20% exceeded 95% test–retest variability for any SUV_{mean} that was encountered in this study, irrespective of the tumor delineation method (values are presented in Table 2). Homogeneous tumor delineation (50% threshold technique) resulted in lower test–retest variability (Fig. 3; Table 2).

We found no statistically significant difference between studies using static (Nahmias et al. (*21*) and Velasquez et al. (*23*)) as compared with dynamic PET techniques (Hoekstra et al. (*18*), Minn et al. (*20*), and Weber et al. (*16*)) ($P = 0.90$). The principal component analysis indicated that the test–retest variability of SUV_{max} was larger than that of SUV_{mean} (Supplemental Figs. 1 and 2; supplemental materials are available online only at http://jnm.snmjournals.org).

TABLE 2
Required Relative and Absolute Differences to Exceed Test–Retest Variability

Parameter	Relative difference	Absolute difference
SUV_{max}	—	—
	25%	3.1
	30%	2.0
SUV_{mean} (H, N [manual analysis], M, W, and V)	20%	1.17
	25%	0.96
	30%	0.75
SUV_{mean} (H, N [50% isocontour analysis], and W)	20%	0.96
	25%	0.78
	30%	0.66

Thresholds can be extracted from Figures 1–3 as explained for SUV_{max} in text. Differences have to exceed both relative and absolute thresholds.

H = Hoekstra et al.; N = Nahmias et al.; M = Minn et al.; V = Velasquez et al.; W = Weber et al.

jnm095299-pm ■ 4/10/12

RGB

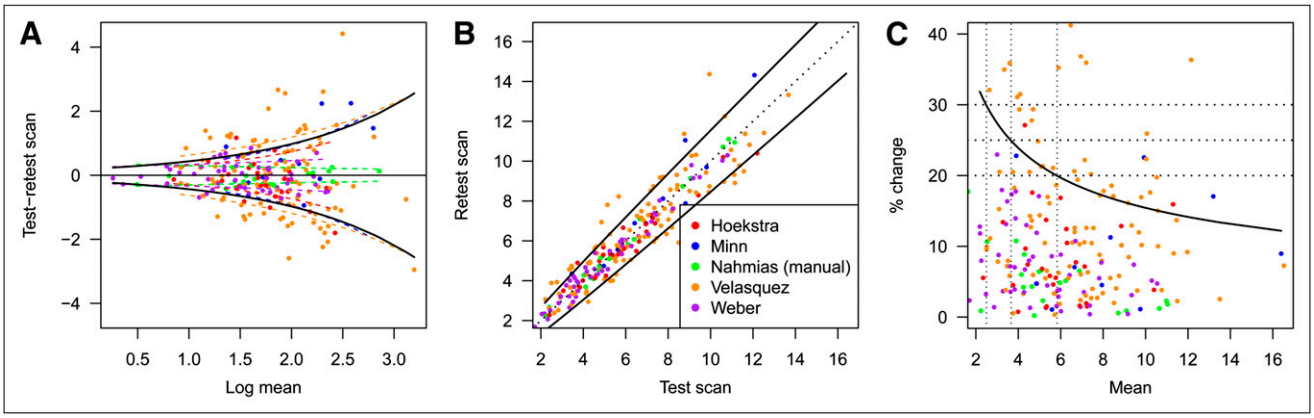

FIGURE 2. Repeatability of SUV_{mean}. (A) Estimated study-specific SD (colored dashed lines; study as fixed effect) and overall SD (black solid lines; study as random effect). (B) Test and retest scan values plotted on original scale. Solid line is coefficient of repeatability (CR95). (C) Relation between CR95, as percentage change, with level of SUV_{mean}. Dotted lines indicate absolute and relative differences, as presented in Table 2.

Single-Assessment Setting

By definition, the CR95 for 1 observation is smaller than the equivalent 2 observation values because of fewer sources of measurement error. The single-assessment CR95s were within 25% change for SUV_{max} and within 20% for SUV_{mean} (Supplemental Table 1; Supplemental Fig. 2).

Impact of Tumor Volume on Repeatability

Tumor volume was moderately correlated with the ^{18}F-FDG uptake level (Table 3; Supplemental Fig. 3A). However, uptake level was a better indicator for changes in test–retest variability than tumor volume for both parameters (Table 3). Repeatability of SUV_{mean} seems to be relatively unaffected by tumor volume, irrespective of the delineation method (Table 3; Supplemental Fig. 3B).

[Table 3]

DISCUSSION

This metaanalysis summarizes the published evidence on the repeatability of commonly used quantitative ^{18}F-FDG measurements in oncology. Our results apply to the use of PET in both serial- and single-assessment settings. Compared with SUV_{max}, SUV_{mean} had the better repeatability.

For both measures, the percentage change in test–retest variability was not constant across the range of parameter values and negatively related to the level of ^{18}F-FDG uptake. In a clinical setting, combining a minimal relative and absolute change is sufficient to define a (biologic) effect that cannot be explained by measurement error only (Table 2 shows thresholds). In the context of analyzing serial ^{18}F-FDG uptake changes, the PET response criteria in solid tumors (PERCIST) classification system assumes a biologic change with peak SUV (SUV_{peak}) changes greater than 30% in combination with 0.8 unit change of absolute SUV_{peak} (*25*). In the present metaanalysis, there were insufficient data to explore SUV_{peak}. Numeric values of SUV_{peak} will typically vary between SUV_{max} and SUV_{mean} (if VOIs of SUV_{mean} exceed the SUV_{peak} VOIs). Figure 4 corroborates the PERCIST assumption of combining 30% with 0.8 unit SUV_{peak} (for SUV_{mean}, we found 30% and 0.75). In addition, Figure 4 shows that, when using SUV_{mean}, a minimal relative change of 20% in combination with 1.2 unit change will also represent a biologic change.

[Fig. 4]

The first combination is advantageous for tumors with low uptake values in which small differences in absolute values correspond to large differences in percentage change. The latter combination is advantageous for high uptake values for which relatively large differences in absolute values correspond to relatively small differences in percentage change. It is probable that for tumors with extremely high uptake values, an even smaller relative change threshold would be appropriate. We are unable to suggest corresponding absolute change values because these uptake values are outside the range of our data.

Alternatively, to select lesions for evaluation with ^{18}F-FDG PET, a minimal SUV could be applied (see figures for trade-off values). This selection simplifies the situation by excluding low uptake values with large (relative change) test–retest variability, leaving the remaining values with a consistent relative test–retest change that can be used as a (1 value) cutoff for response assessment. The disadvantage of this approach is that increases in ^{18}F-FDG uptake (e.g., due to disease progression) can be measured reliably, but decreases to values below the threshold (e.g., due to treatment response) cannot.

Even though methodologies were not fully consistent throughout the studies, we found no important interstudy differences in parameter repeatability. For SUV_{max}, no difference was found at all, whereas for SUV_{mean} a difference was present only when all studies were included, irrespective of the delineation method used. This difference was caused by a lower performance of the Velasquez data and good performance of the manually delineated data of Nahmias et al. (*21*). Velasquez et al. (*23*) applied a multicen-

RGB

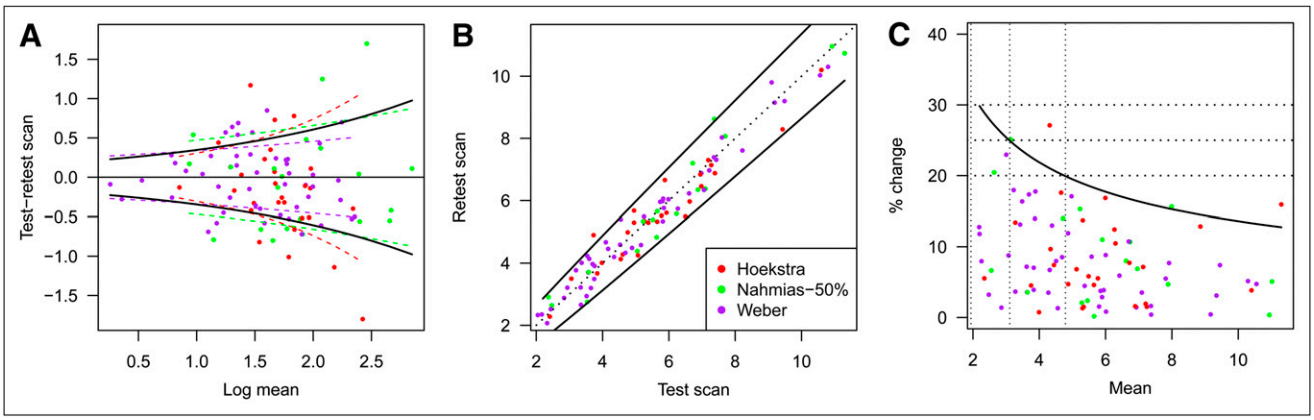

FIGURE 3. Repeatability of SUV_{mean} using 50% isocontouring. (A) Estimated study-specific SD (colored dashed lines; study as fixed effect) and overall SD (black solid lines; study as random effect). (B) Test and retest scan values plotted on original scale. Solid line is coefficient of repeatability (CR95). (C) Relation between CR95, as percentage change, with level of SUV_{mean} using 50% isocontouring. Dotted lines indicate absolute and relative differences, as presented in Table 2.

ter design with a large number of centers. Also, the patient population was substantially different from that of the other studies, because only patients with advanced gastrointestinal malignancies were included by Velasquez et al., whereas the other studies predominantly evaluated lung lesions (Table 1). Tumor delineation in the liver and gastrointestinal tract may have been challenged by physiologic uptake, which is higher and more variable there than in the lung. Furthermore, spillover of activity from normal organs with variable ^{18}F-FDG uptake (e.g., kidneys, bowel loops) may have affected the test–retest variability. The authors reported that they were forced to apply a higher threshold of the maximum voxel value (70%) to enable tumor tissue discrimination from the normal background. Earlier work from our center showed that test–retest variability increases with higher thresholds of the maximum tumor volume (*19*). Possibly these factors resulted in lower repeatability.

Therefore, we performed a subset analysis of homogeneously delineated lesions using the 50% isocontour technique. This inherently excluded the data of Velasquez et al. (*23*), resulting in interstudy homogeneity. Better repeatability was found for this subset, as presented in Figure 3 and Table 2. It remains elusive whether the worse repeatability in the overall dataset is due to the delineation technique (70% threshold) or the location of the lesions (abdominal). Therefore, the cutoff values found for the 50% threshold subset hold only for 50% isocontoured extraabdominal lesions, whereas the overall results can be used for all organ sites and histology and irrespective of the SUV_{mean} tumor delineation method.

Interestingly, the manually delineated data of Nahmias et al. (*21*) performed better than the threshold analysis of the same dataset. This better performance was also reported earlier by our group (*19*). However, both studies were single-observer, whereas the strength of the semiautomated technique is the high intra- and interobserver repeatability for lesion delineation. Therefore, superiority can be evaluated only by performing a head-to-head analysis in a multiobserver setting.

In previous studies, it was suggested that repeatability is a function of ^{18}F-FDG uptake value and tumor volume. Weber et al. (*16*) reported that the absolute difference between the test and retest values remained constant across the range of parameter value but that the relative change increased with decreasing uptake value. Nahmias et al. (*21*)

TABLE 3
Correlations Between ^{18}F-FDG Uptake and Tumor Volume and Their Influence on Test–Retest Variability

			Log likelihood ratio test *P*			
			Uptake		Volume	
Parameter	Studies	Correlations, uptake vs. volume	Alone	Adjacent volume	Alone	Adjacent mean
SUV_{max}	H, M, N (manual analysis), and V	0.38	<0.0001	<0.0001	0.03	0.26
SUV_{mean}	H, M, N, W, and V	0.39	<0.0001	<0.0001	0.03	0.71
SUV_{mean}	H, N (50% isocontour analysis), and W	0.45	0.002	0.003	0.32	0.78

H = Hoekstra et al.; N = Nahmias et al.; M = Minn et al.; V = Velasquez et al.; W = Weber et al.

jnm095299-pm ■ 4/10/12

RGB

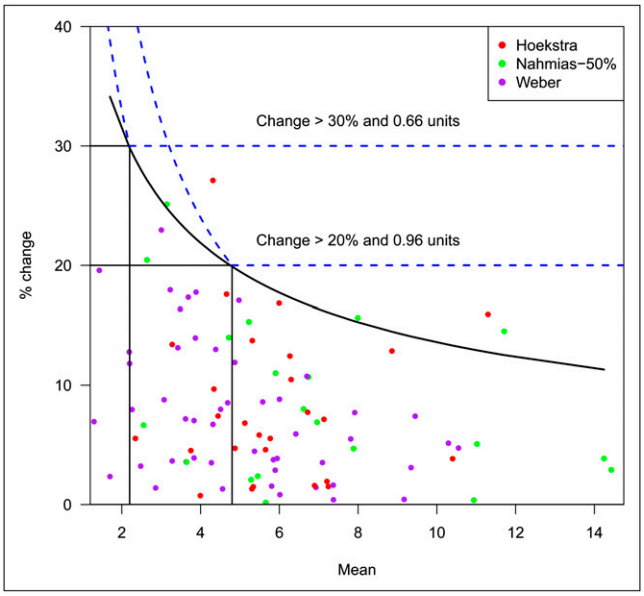

FIGURE 4. Relation between CR95, as percentage change, with level of SUV_{mean} using 50% isocontouring. Cutoff rules were at 20% and 30% relative change, with associated absolute changes for 3 studies. Required absolute difference is calculated as relative difference multiplied by mean value, where CR95 curve crosses relative difference value. For example, CR95 equals 30% at 2.2; hence, associated absolute difference is 0.66 (2.2 × 0.3). Therefore, likelihood is less than 5% that change of at least 30% and more than 0.66 units is due to measurement error in ^{18}F-FDG uptake rate. Blue dashed lines represent areas of combined coverage for each rule.

also reported that the absolute difference of the mean uptake value was constant across the range of SUV_{mean}, with increasing relative change for lower SUV_{mean}. However, this group reported that for SUV_{max} the absolute difference increased with increasing SUV_{max}, whereas the relative change remained constant.

In the pooled dataset of the current study, mean uptake was a better indicator for changes in test–retest variability than tumor volume (Table 3). SUV repeatability seemed to be relatively unaffected by VOI size.

We observed a moderately positive correlation between the uptake value and tumor volume (Table 3; Supplemental Fig. 3A). Hypothetically this can, at least in part, be explained by the partial-volume phenomenon, which results in an underestimation of ^{18}F-FDG uptake in smaller lesions, in turn resulting in more difficult tumor delineation because of lower tumor-to-background contrast (*26*).

Our results apply to the setting in which the same scanner is being used for both the baseline and the posttreatment scans, with the same acquisition, reconstruction, data processing, and analytic protocol. These inclusion criteria were mandatory because variation in these parameters can result in parameter differences of up to 50%, representing only noise and not a true biologic change (*27,28*). Although interstudy differences in image contrast and resolution remained, their effects on repeatability were negligible (*29*). Whether this is also true for newer systems with better spatial resolution cannot be predicted on the basis of the current data.

The study by Kamibayashi et al. (*24*) was excluded because of the use of a different PET scanner for the test and retest scans. In this study, tumor SUV test–retest variability was evaluated in patients who were scanned once on a PET-only scanner and once on a PET/CT scanner without treatment intervention between the 2 scans. The scanners were from the same vendor, and the same acquisition and analytic protocol was used for both scans. The authors applied a manually drawn 2-dimenstional region of interest for tumor delineation. The SD was only slightly worse than that of the other single-center repeatability studies and comparable to the multicenter study of Velasquez et al. (*23*), with an SD of 12% ± 10.2% for SUV_{mean} and 16.1% ± 10.5% for SUV_{max}. This result might indicate that different scanners can be used at baseline and after treatment, provided that standardization of acquisition, reconstruction, data processing, and analysis is applied. Whether this possibility also holds for semiautomatic VOI definition and the use of PET scanners from different vendors remains to be further examined.

Although no statistical interstudy difference was present after the data were pooled and VOI definition heterogeneity was corrected, the cutoff values for definition of test–retest variability should be assessed with care because some methodologic and patient spectrum heterogeneity between studies prevailed.

However, we believe that this metaanalysis provides the most critical and robust view of ^{18}F-FDG PET repeatability in the oncologic setting to date. It enabled the evaluation of dependency on mean uptake value and tumor volume, issues that were open for discussion since the publication of Weber's study in 1999 (*16*). Ideally, these results should be confirmed prospectively, preferably in a large multicenter study. Possibly, studies such as the "FDG PET/CT as a Predictive Marker of Tumor Response and Patient Outcome: Prospective Validation in Non-Small Cell Lung Carcinoma" (American College of Radiology Imaging Network study 6678), incorporating a test–retest study evaluating ^{18}F-FDG PET/CT with a static protocol at 60 min after injection, can provide the opportunity for this confirmation.

CONCLUSION

This metaanalysis shows the repeatability of different ^{18}F-FDG uptake measurements using the available data. SUV_{mean} performed better than SUV_{max}. Importantly, both parameters showed worse repeatability for lesions with low ^{18}F-FDG uptake, which can be accounted for by combining relative and absolute differences. For SUV_{mean}, a 30% and 0.75 unit change or 20% and 1.2 unit change exceed 95% test–retest variability. Homogeneous delineation and exclusion of abdominal lesions improved repeatability. For SUV_{max}, the equivalent values were 30% and 2 units change. These cutoff values can serve as a guide for future clinical trials. Given the limited data and because some

interstudy heterogeneity prevailed, our results should preferably be confirmed in a prospective repeatability study, preferentially including SUV_{peak}. For serial (e.g., baseline and postintervention scans) and single PET measurements, different thresholds should be used because each scan is affected by test–retest noise.

DISCLOSURE STATEMENT

The costs of publication of this article were defrayed in part by the payment of page charges. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

ACKNOWLEDGMENTS

We are grateful to Drs. C. Nahmias and L.M. Wahl for performing reanalyses of their published study results. No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Avril N, Sassen S, Roylance R. Response to therapy in breast cancer. *J Nucl Med.* 2009;50:55S–63S.
2. de Geus-Oei LF, Vriens D, van Laarhoven HW, van der Graaf WT, Oyen WJ. Monitoring and predicting response to therapy with ^{18}F-FDG PET in colorectal cancer: a systematic review. *J Nucl Med.* 2009;50:43S–54S.
3. Hicks RJ. Role of ^{18}F-FDG PET in assessment of response in non-small cell lung cancer. *J Nucl Med.* 2009;50:31S–42S.
4. Hutchings M, Barrington SF. PET/CT for therapy response assessment in lymphoma. *J Nucl Med.* 2009;50:21S–30S.
5. Krause BJ, Herrmann K, Wieder H, zum Buschenfelde CM. ^{18}F-FDG PET and ^{18}F-FDG PET/CT for assessing response to therapy in esophageal cancer. *J Nucl Med.* 2009;50:89S–96S.
6. Schöder H, Fury M, Lee N, Kraus D. PET monitoring of therapy response in head and neck squamous cell carcinoma. *J Nucl Med.* 2009;50:74S–88S.
7. Martoni AA, Zamagni C, Quercia S, et al. Early ^{18}F-2-fluoro-2-deoxy-D-glucose positron emission tomography may identify a subset of patients with estrogen receptor-positive breast cancer who will not respond optimally to preoperative chemotherapy. *Cancer.* 2010;116:805–813.
8. McLarty K, Fasih A, Scollard DA, et al. ^{18}F-FDG small-animal PET/CT differentiates trastuzumab-responsive from unresponsive human breast cancer xenografts in athymic mice. *J Nucl Med.* 2009;50:1848–1856.
9. Prior JO, Montemurro M, Orcurto MV, et al. Early prediction of response to sunitinib after imatinib failure by ^{18}F-fluorodeoxyglucose positron emission tomography in patients with gastrointestinal stromal tumor. *J Clin Oncol.* 2009; 27:439–445.
10. Storto G, De RA, Pellegrino T, et al. Assessment of metabolic response to radioimmunotherapy with ^{90}Y-ibritumomab tiuxetan in patients with relapsed or refractory B-cell non-Hodgkin lymphoma. *Radiology.* 2010;254:245–252.
11. Herrmann K, Krause BJ, Bundschuh RA, Dechow T, Schwaiger M. Monitoring response to therapeutic interventions in patients with cancer. *Semin Nucl Med.* 2009;39:210–232.
12. 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.
13. Juweid ME, Stroobants S, Hoekstra OS, et al. Use of positron emission tomography for response assessment of lymphoma: consensus of the Imaging Subcommittee of International Harmonization Project in Lymphoma. *J Clin Oncol.* 2007;25:571–578.
14. Young H, Baum R, Cremerius U, et al. Measurement of clinical and subclinical tumour response using [^{18}F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. *Eur J Cancer.* 1999;35:1773–1782.
15. Nevill AM, Copas JB. Using generalized linear models (GLMs) to model errors in motor performance. *J Mot Behav.* 1991;23:241–250.
16. 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.
17. Daly LE, Bourke GJ. *Interpretation and Uses of Medical Statistics*. Oxford, U.K.: Blackwell Science; 2000.
18. Hoekstra CJ, Hoekstra OS, Stroobants SG, et al. Methods to monitor response to chemotherapy in non-small cell lung cancer with ^{18}F-FDG PET. *J Nucl Med.* 2002;43:1304–1309.
19. 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.
20. Minn H, Zasadny KR, Quint LE, Wahl RL. Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET. *Radiology.* 1995;196:167–173.
21. 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.
22. Nakamoto Y, Zasadny KR, Minn H, Wahl RL. Reproducibility of common semiquantitative parameters for evaluating lung cancer glucose metabolism with positron emission tomography using 2-deoxy-2-[^{18}F]fluoro-D-glucose. *Mol Imaging Biol.* 2002;4:171–178.
23. 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.
24. Kamibayashi T, Tsuchida T, Demura Y, et al. Reproducibility of semi-quantitative parameters in FDG-PET using two different PET scanners: influence of attenuation correction method and examination interval. *Mol Imaging Biol.* 2008;10:162–166.
25. 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;50:122S–150S.
26. de Langen AJ, van den Boogaart VE, Marcus JT, Lubberink M. Use of H2^{15}O-PET and DCE-MRI to measure tumor blood flow. *Oncologist.* 2008;13:631–644.
27. Boellaard R, Krak NC, Hoekstra OS, Lammertsma AA. Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study. *J Nucl Med.* 2004;45:1519–1527.
28. Westerterp M, Pruim J, Oyen W, et al. Quantification of FDG PET studies using standardised uptake values in multi-centre trials: effects of image reconstruction, resolution and ROI definition parameters. *Eur J Nucl Med Mol Imaging.* 2007;34:392–404.
29. Cheebsumon P, van Velden FH, Yaqub M, et al. Effects of image characteristics on performance of tumor delineation methods: a test-retest assessment. *J Nucl Med.* 2011;52:1550–1558.