Assessment of Interobserver Reproducibility in Quantitative ^{18}F-FDG PET and CT Measurements of Tumor Response to Therapy

Heather A. Jacene[1], Sophie Leboulleux[2], Shingo Baba[3], Daniel Chatzifotiadis[1], Behnaz Goudarzi[1], Oleg Teytelbaum[1], Karen M. Horton[1], Ihab Kamel[1], Katarzyna J. Macura[1], Hua-Ling Tsai[4], Jeanne Kowalski[4], and Richard L. Wahl[1,5]

[1]Divisions of Nuclear Medicine and Body CT, Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University, Baltimore, Maryland; [2]Department of Nuclear Medicine and Endocrine Oncology, Institut Gustave Roussy, University Paris Sud-XI, Villejuif, France; [3]Department of Clinical Radiology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan; [4]Division of Oncology Biostatistics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland; and [5]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland

Our goal was to estimate and compare across different readers the reproducibility of the ^{18}F-FDG PET standardized uptake value (SUV) and CT size measurements, and changes in those measurements, in malignant tumors before and after therapy. **Methods:** Fifty-two tumors in 25 patients were evaluated on ^{18}F-FDG PET/CT scans. Maximum SUVs (SUV_{bw} max) and CT size measurements were determined for each tumor independently on pre- and posttreatment scans by 8 different readers (4 PET, 4 CT) using routine nonautomated clinical methods. Percentage changes in SUV_{bw} max and CT size between pre- and posttreatment scans were calculated. Interobserver reproducibility of SUV_{bw} max, CT size, and changes in these values were described by intraclass correlation coefficients (ICCs) and estimates of variance. **Results:** The ICC was higher for the pretreatment, posttreatment, and percentage change in SUV_{bw} max than the ICC for the longest CT size and the 2-dimensional CT size (before treatment, 0.93, 0.72, and 0.61, respectively; after treatment, 0.91, 0.85, and 0.45, respectively; and percentage change, 0.94, 0.70, and 0.33, respectively). The variability of SUV_{bw} max was significantly lower than the variability of the longest CT size and the 2-dimensional CT size (mean ± SD before treatment, 6.3% ± 14.2%, 16.2% ± 17.8%, and 27.5% ± 26.7%, respectively, $P \leq 0.001$; and after treatment, 18.4% ± 26.8%, 35.1% ± 47.5%, and 50.9% ± 51.4%, respectively, $P \leq 0.02$). The variability of percentage change in SUV_{bw} max (16.7% ± 36.2%) was significantly lower than that for percentage change in the longest CT size (156.3% ± 157.3%, $P \leq 0.0001$) and the 2-dimensional CT size (178.4% ± 546.5%, $P < 0.0001$). **Conclusion:** The interobserver reproducibility of SUV_{bw} max for both untreated and treated tumors and percentage change in SUV_{bw} max are substantially higher than measurements of CT size and percentage change in CT size. Measurements of tumor metabolism by PET should be included in trials to assess response to therapy. Although PET reproducibility was high, the variability observed in analyses of identical image sets by 4 readers indicates that automated analytic tools to assess response might be helpful to further enhance reproducibility.

Key Words: reproducibility; SUV; ^{18}F-FDG PET; CT; variability

J Nucl Med 2009; 50:1760–1769
DOI: 10.2967/jnumed.109.063321

Measurement of tumor response is an essential component of most anticancer therapy clinical trials. Standardized and reproducible assessments of response are required for meaningful comparisons and conclusions across multiple trials. Presently, the major response criteria for solid tumors, Response Evaluation Criteria in Solid Tumors (RECIST) (*1*) and the World Health Organization criteria (*2*), primarily rely on changes in tumor size on anatomic imaging. Although several studies have demonstrated reasonably good intra- and interobserver reproducibility of tumor size measurements (*3,4*), others have demonstrated that inconsistent tumor size measurements can lead to incorrect interpretations of tumor response (*5,6*).

Functional imaging with ^{18}F-FDG PET is being applied with growing frequency in cancer treatment response trials because of the ability of this modality to predict the response of a tumor to therapy and outcomes (*7,8*). As a result, ^{18}F-FDG PET is being applied with growing frequency in cancer treatment response trials. A visual assessment of ^{18}F-FDG PET images has been incorporated into the revised International Workshop Criteria for monitoring response of lymphoma to therapy (*9*), and ^{18}F-FDG PET has also been included in a recent update of RECIST (RECIST 1.1) (*10*).

Investigators have begun to focus on developing standardized metabolic response criteria, and the methods used

Received Feb. 17, 2009; revision accepted Jul. 13, 2009.
For correspondence or reprints contact: Richard L. Wahl, Division of Nuclear Medicine/PET, Russell H. Morgan Department of Radiology and Radiological Science, 601 N. Caroline St., JHOC 3223, Baltimore, MD 21287.
E-mail: rwahl@jhmi.edu

to quantify ^{18}F-FDG uptake in tumors are being carefully scrutinized. PET is intrinsically quantitative, and a commonly used parameter is the standardized uptake value (SUV), defined as concentration of radioactivity in tissue normalized to injected dose and body mass, lean body mass, or surface area (*11*). Although a considerable range of approaches to SUV determination has been used, maximum SUV in a single voxel is widely used because of its simplicity (*12*).

Despite the relative ease of SUV determination, compared with other quantitative parameters, such as Patlak analysis or full kinetic approaches, numerous patient- and technique-related factors can affect SUV (*12–18*). For repeated tumor measurements, the technique for obtaining the ^{18}F-FDG PET study and SUV should be the same in all institutions and SUV must be highly reproducible so that data obtained in multicenter trials can be compiled for the evaluation of large patient populations. This is particularly true for trials of therapy assessment when small changes in tumor metabolism are being evaluated.

A limited number of test–retest studies have shown that SUV is a reproducible parameter, with intrasubject variability of tumor SUV ranging from 3% to 14% (*12,15, 18–22*). However, most of these studies were performed in carefully controlled settings to best optimize the precision of the SUV measurement, and untreated tumors of substantial size and high tumor metabolism only were evaluated. Data on interobserver reproducibility of SUV are more limited but reported to be high in untreated tumors (*20,21,23*). To the best of our knowledge, just 1 recent study has evaluated the interobserver reproducibility of SUV in the posttherapy setting (*23*).

The reproducibility of SUV measurements has not been directly compared with the reproducibility of quantitative measurements of CT size, the current standard for assessing tumor response to therapy (*1,2*), in the same patient. We hypothesized that SUV is more reproducible for pre- versus posttherapy studies and more reproducible than measurements of CT size on anatomic imaging. Our purpose was to estimate and compare the interobserver reproducibility of SUV and CT size measurements, and changes in those measurements, in malignant tumors before and after treatment using readily available clinical methodologies.

MATERIALS AND METHODS

Patients

Retrospective data compilation and image review were approved by the Johns Hopkins Institutional Review Board. Between April 2003 and April 2005, 25 patients (6 men, 19 women; mean age, 51 ± 14 y; 16 with primary breast carcinoma, 9 with primary lung carcinoma) were identified as having a pretreatment ^{18}F-FDG PET/CT scan and a posttreatment scan soon after treatment was begun. Nineteen patients had untreated primary disease, and 6 patients had untreated recurrent malignancy. Interval therapy consisted of chemotherapy (n = 21), hormonal therapy (n = 1), chemotherapy and hormonal therapy (n = 1), chemotherapy and biologic therapy (n = 1), or chemotherapy and radiation therapy (n = 1).

^{18}F-FDG PET/CT Scans

Patients fasted for a minimum of 4 h and had blood glucose levels less than or equal to 200 mg/dL just before the intravenous injection of ^{18}F-FDG (8.14 MBq/kg [0.22 mCi/kg]). Oral, but not intravenous, contrast was administered for the CT portion of the study.

After an approximately 60-min uptake phase, combined whole-body PET/CT (Discovery LS; GE Healthcare) was performed. Whole-body CT was performed first with a 4-slice multidetector helical scanner and the following parameters: 140 kV, weight-based amperage (range, 80–160 mA), 0.8 s per CT rotation, pitch of 6, table speed of 22.5 mm/s, 722.5-mm coverage, and 31.9-s acquisition time. A CT transmission map was generated for image fusion. PET emission data were acquired for 5 min at each bed position, with the patient in the same position as for the CT portion of the study. PET images were reconstructed using the ordered-subset expectation maximization algorithm (2 iterations, 28 subsets), an 8-mm gaussian filter with a 128 × 128 matrix, and non–contrast-enhanced CT attenuation correction.

Image Analysis

Fifty-two tumors (up to 3 per patient) were identified for analysis by 1 author and indicated to the readers as the reference tumors by location and transaxial image number on the pretreatment scan. For the posttreatment study, readers independently identified the tumors to be analyzed by comparing the posttreatment with the pretreatment scan. In patients with multiple tumors, the largest tumors, greater than 10 mm in at least 1 dimension, were chosen for analysis. Median tumor size was 22 mm (range, 10–58 mm) by 15 mm (range, 7–41 mm).

PET images were reviewed on a Xeleris workstation (GE Healthcare) by 4 nuclear medicine physicians or nuclear radiologists with experience in ^{18}F-FDG PET/CT. Images were viewed on a single split screen displaying PET, CT, and fused PET/CT images. Readers were asked to manually determine the single voxel SUV_{bw} maximum (SUV_{bw} max) of each tumor on the pre- and posttreatment PET scans using the SUV tools on the Xeleris software. If a tumor completely resolved on the posttreatment scan, readers were instructed to record the single voxel value SUV_{bw} max of background tissues in the area of the previous tumor. No further instructions regarding the exact method to determine SUV_{bw} max were specified.

CT images were reviewed with Emageon UltraVisual software (UltraVisual Medical Systems Inc.) by 4 board-certified radiologists with extensive CT experience. The longest and perpendicular sizes of each tumor were determined with the UltraVisual measuring tool, with the PET images available for comparison. Readers were instructed to record the CT size as zero (0) millimeters if the tumor completely resolved on the posttreatment scan. The 2-dimensional size of each tumor was determined by multiplying the longest and perpendicular dimensions.

The longest and 2-dimensional CT sizes chosen for evaluation as response assessments were based on changes in these parameters by the major response criteria in clinical trials, RECIST and World Health Organization criteria, respectively (*1,2*).

Statistical Analyses

For all analyses, the individual tumors were considered independently. Generalized estimating equations (*24*) were used to

model mean SUV_{bw} max and CT size results. A minimum of 3 readers was required to measure an individual lesion for the lesion to be included in the analyses for both SUV_{bw} max and CT size. A χ^2 statistic was used to test the hypotheses of mean differences among the readers and between pre- and posttreatment scans for SUV_{bw} max and both CT size parameters.

Percentage declines in SUV_{bw} max and CT size for each lesion between the pre- and posttreatment scans were calculated using the following equation:

$$\text{Percentage decline} = [(\text{pretreatment} - \text{posttreatment})/\text{pretreatment}] \times 100.$$

Differences in percentage change between PET and CT parameters were tested with a χ^2 statistic.

We used 2 methods to assess interobserver agreement for the various parameters. Intraclass correlation coefficients (ICCs) were estimated as a direct measure of agreement among the raters (reproducibility) and were calculated using variance estimates obtained through ANOVA (*25*). The ICC ranges between 0.00 and 1.00, with values closer to 1.00 representing better reproducibility. Interpretation of ICC was categorized according to Landis and Koch (*26*) (<0, no reproducibility; 0.0–0.20, slight reproducibility; 0.21–0.40, fair reproducibility; 0.41–0.60, moderate reproducibility; 0.61–0.80, substantial reproducibility; and 0.81–1.00, almost-perfect reproducibility). The reproducibility of the ICC estimates based on their precision (half the width of the 95% confidence interval [CI] × 100%) was also determined.

Coefficients of variation (CV) were estimated to assess the percentage variability between SUV_{bw} max and CT size parameters among the readers. CV was calculated for each tumor by dividing the SD of 4 readers by the mean of 4 readers. The mean CV and SD across all tumors was then determined. Differences between CV of the various parameters were compared using *t* tests.

It is possible that interobserver reproducibility is dependent on the level of ^{18}F-FDG uptake in a tumor or tumor size. To test this, the same analyses described above were repeated for the tumors with the highest and lowest average SUV_{bw} max ($n = 20$, each) and the largest and smallest average CT size ($n = 20$, each) on the pretreatment scan. The tumors with the highest and lowest average SUV_{bw} max were not necessarily the same tumors with the largest and smallest average CT size.

Statistical analyses were performed using R 2.6 online software and SAS 9.1 statistical software (SAS Institute). *P* values less than or equal to 0.05 were considered statistically significant.

RESULTS

The median time between the pretreatment and posttreatment ^{18}F-FDG PET/CT scans was 52 d (range, 8–175 d). On the basis of 2-tailed paired *t* tests, parameters known to affect SUV were not significantly different when comparing pre- versus posttreatment scans (serum glucose levels, 100 ± 16 mg/dL vs. 102 ± 18 mg/dL, $P = 0.62$; patient body weight, 76.3 ± 20.7 kg vs. 77.0 ± 20.6 kg, $P = 0.57$; injected activity of ^{18}F-FDG, 609 ± 145 MBq vs. 630 ± 164 MBq, $P = 0.24$; and ^{18}F-FDG uptake time, 64.4 ± 11.5 min vs. 62.2 ± 16.6 min, $P = 0.62$).

For pretreatment scans, SUV_{bw} max was determined by all 4 PET readers for all 52 tumors. SUV_{bw} max was also determined by all 4 PET readers for all 52 tumors (persistent or residual tumor or background tissue in the location of previous tumor) for posttreatment scans. In addition to absolute SUV_{bw} max determination, PET readers were asked to indicate whether the measurement was obtained in persistent or residual tumor or background tissues after treatment by visual assessment. For 39 of 52 tumors, all 4 PET readers agreed and determined SUV_{bw} max in 32 persistent or residual tumors and 7 background tissues. For the remaining 13 original tumor foci, there was not complete consensus and SUV_{bw} max was determined in residual or persistent tumor or background tissues in the region of the previously visualized tumor, depending on individual readers' assessments.

Before treatment, CT size was determined by all 4 CT readers in 46 tumors, by 3 readers in 4 tumors, and by 2 and 1 readers for 1 tumor each. After therapy, CT size was determined by all 4 readers in 43 tumors, by 3 readers in 6 tumors, and by 2 readers in 3 tumors. The tumors not measured by all readers were not confidently seen or were considered to be unmeasurable because their edge was not clearly defined.

Table 1 summarizes the hypothesis-testing results of mean differences based on our model for the SUV_{bw} max and CT size parameters. No significant difference in mean SUV_{bw} max was found among the 4 readers on the pretreatment scan (reader 1, 9.4 ± 6.3; reader 2, 9.7 ± 6.5; reader 3, 9.8 ± 6.2; and reader 4, 9.3 ± 6.3, $P = 0.98$) or the posttreatment scan (reader 1, 4.4 ± 4.0; reader 2, 4.8 ± 4.2; reader 3, 4.7 ± 4.0; and reader 4, 4.5 ± 3.8, $P = 0.96$). On average, SUV_{bw} max was significantly higher on the pretreatment scan than on the posttreatment scan (9.6 ± 6.3 vs. 4.6 ± 4.0, $P < 0.001$).

Mean CT size measurements were not significantly different among the 4 readers before treatment for the longest CT size (reader 1, 25.5 ± 11.6 mm; reader 2, 25.9 ± 11.7 mm; reader 3, 23.5 ± 10.4 mm; and reader 4, 25.6 ± 14.8 mm, $P = 0.75$) but were significantly different after treatment (reader 1, 22.5 ± 13.4 mm; reader 2, 23.0 ± 12.9 mm; reader 3, 17.7 ± 11.6 mm; and reader 4, 17.4 ± 11.8 mm, $P = 0.04$). The 2-dimensional CT size was not significantly different among readers before treatment (reader 1, 564.6 ± 568.3 mm^2; reader 2, 541.1 ± 448.3 mm^2; reader 3, 453.1 ± 411.5 mm^2; and reader 4, 614.1 ± 894.6 mm^2, $P = 0.60$) or after treatment (reader 1, 471.2 ± 568.7 mm^2; reader 2, 440.3 ± 459.4 mm^2; reader 3, 316.6 ± 366.4 mm^2; and reader 4, 422.3 ± 990.0 mm^2, $P = 0.65$). CT size was, on average, significantly larger on the pretreatment scan than on the posttreatment scan (longest dimension, 25.1 ± 12.2 mm vs. 20.1 ± 12.6 mm, $P < 0.001$; 2-dimensional size, 541.8 ± 607.8 mm^2 vs. 410.7 ± 637.3 mm^2, $P = 0.009$).

The average percentage decline in SUV_{bw} max between the pretreatment and the posttreatment scans was 45% ±

TABLE 1. Summary of Hypothesis Testing of Mean Differences for PET and CT (Semi)Quantitative Parameters

Parameter	Pretreatment scan	Posttreatment scan	*P*
SUV_{bw} max			0.98/0.96*, <0.001†
Reader 1	9.4 ± 6.3 (7.6–11.1)	4.4 ± 4.0 (3.3–5.5)	
Reader 2	9.7 ± 6.5 (7.9–11.5)	4.8 ± 4.2 (3.6–5.9)	
Reader 3	9.8 ± 6.2 (8.0–11.5)	4.7 ± 4.0 (3.6–5.8)	
Reader 4	9.3 ± 6.3 (7.6–11.1)	4.5 ± 3.8 (3.5–5.6)	
All readers	9.6 ± 6.3 (8.7–10.4)	4.6 ± 4.0 (4.1–5.1)	
Longest CT dimension (RECIST, mm)			0.75/0.04*, <0.001†
Reader 1	25.5 ± 11.6 (22.0–28.9)	22.5 ± 13.4 (18.5–26.5)	
Reader 2	25.9 ± 11.7 (22.7–29.2)	23.0 ± 12.9 (19.3–26.6)	
Reader 3	23.5 ± 10.4 (20.6–26.4)	17.7 ± 11.6 (14.5–21.0)	
Reader 4	25.6 ± 14.8 (21.4–29.8)	17.4 ± 11.8 (14.0–20.8)	
All readers	25.1 ± 12.2 (23.4–26.8)	20.1 ± 12.6 (18.3–21.9)	
Two-dimensional CT size (WHO, mm^2)			0.60/0.65*, 0.009†
Reader 1	564.6 ± 568.3 (395.8–733.3)	471.2 ± 568.7 (300.3–642.0)	
Reader 2	541.1 ± 448.3 (415.0–667.2)	440.3 ± 459.4 (311.1–569.5)	
Reader 3	453.1 ± 411.5 (338.5–567.6)	316.6 ± 366.4 (213.6–419.7)	
Reader 4	614.1 ± 894.6 (359.8–868.3)	422.3 ± 990.0 (137.9–706.6)	
All readers	541.8 ± 607.8 (456.9–626.1)	410.7 ± 637.3 (320.9–500.5)	

*Mean SUV_{bw} max was not significantly different among PET readers (reader 1 vs. reader 2 vs. reader 3 vs. reader 4), irrespective of pre- or posttreatment scans. Means for longest CT size were not different among CT readers before treatment but were after treatment. Means for 2-dimensional CT size were not significantly different for 4 CT readers, irrespective of pre- or posttreatment scans.
†Mean differences between pre- and posttreatment scans were tested based on generalizing estimating equations model.
WHO = World Health Organization.
Data are mean ± SD, with 95% CIs in parentheses.

35% (Table 2). This decline was significantly higher than both declines in the longest CT dimension (20% ± 33%, $P < 0.001$) and the 2-dimensional CT size (24% ± 56%, $P = 0.003$).

A summary of the results for the ICC estimates and the reproducibility of the ICC estimates based on their precision are shown in Table 3. ICC was 0.93 (95% CI, 0.90–0.96; precision, ±3%) for pretreatment SUV_{bw} max and 0.91 (95% CI, 0.86–0.94; precision, ±4%) for posttreatment SUV_{bw} max, indicating almost-perfect reproducibility. Pretreatment CT size measurements were substantially reproducible, with an ICC of 0.72 (95% CI, 0.61–82; precision, ±11%) for the longest CT size and an ICC of 0.61 (95% CI, 0.48–0.74; precision, ±13%) for the 2-dimensional CT size. There was almost-perfect reproducibility for the longest CT size after treatment (ICC, 0.85; 95% CI, 0.77–0.91; precision, ±7%), but the 2-dimensional CT size was just moderately reproducible after treatment (ICC, 0.45; 95% CI, 0.30–0.61; precision, ±16%). The precision of the ICC estimate was highest for measurements of SUV_{bw} max before and after treatment (Table 3). Individual SUV_{bw} max and CT size measurement data points for each tumor and reader before and after treatment are shown in Figures 1 and 2.

TABLE 2. Summary of Percentage Change in PET and CT Parameters Between Pretreatment and Posttreatment Scans

Parameter	Percentage decline (observed mean ± SD)	*P* vs. SUV_{bw} max
SUV_{bw} max	45% ± 35%	
Longest CT dimension (RECIST)	20% ± 33%	<0.001
Two-dimensional CT size (WHO)	24% ± 56%	0.003

WHO = World Health Organization.

There was almost-perfect reproducibility for percentage decline in SUV_{bw} max among the 4 PET readers, with an ICC of 0.94 (95% CI, 0.90–0.96; precision, ±3%) (Table 3). Reproducibility among the 4 CT readers was substantial for percentage decline in the longest CT dimension (ICC, 0.70; 95% CI, 0.57–0.81; precision, ±12%) but just fair for the percentage decline in the 2-dimensional CT size (ICC, 0.33; 95% CI, 0.18–0.50; precision, ±16%). Reproducibility of percentage decline in SUV_{bw} max was higher than that for percentage change in CT size measurements (Fig. 3).

A summary of the results for the percentage variability between SUV_{bw} max and CT size parameters among the readers using CV is shown in Table 4. For the pretreatment scan, the CV of SUV_{bw} max (mean ± SD, 6.3% ± 14.2%) was significantly lower than the CV of the longest CT size (16.2% ± 17.8%, $P = 0.001$) and the 2-dimensional CT size (27.5% ± 26.7%, $P < 0.0001$). The CV of SUV_{bw} max (18.4% ± 26.8%) was also significantly lower than the CV of the longest CT dimension (35.1% ± 47.5%, $P = 0.02$) and the 2-dimensional CT size (50.9% ± 51.4%, $P < 0.001$) after treatment. The CV of the longest CT size was significantly lower than the CV of the 2-dimensional CT

TABLE 3. Interobserver Reproducibility Using ICCs

Parameter	Pretreatment scan			Posttreatment scan			Percentage decline between pre- and posttreatment scans		
	ICC	95% CI	Precision*	ICC	95% CI	Precision*	ICC	95% CI	Precision*
SUV_{bw} max	0.93	0.90−0.96	±3%	0.91	0.86−0.94	±4%	0.94	0.90−0.96	±3%
Longest CT size (RECIST)	0.72	0.61−0.82	±11%	0.85	0.77−0.91	±7%	0.70	0.57−0.81	±12%
Two-dimensional CT size (WHO)	0.61	0.48−0.74	±13%	0.45	0.30−0.61	±16%	0.33	0.18−0.50	±16%

*Precision of ICC estimate is defined as one-half length of 95% CI (expressed as percentage) and is measure of reproducibility of ICC estimate.
WHO = World Health Organization.

size before treatment ($P = 0.007$) but not after treatment ($P = 0.06$). The CVs of SUV_{bw} max and both CT size parameters were less on the pre- than on the posttreatment scans (Table 4).

Mean CV for percentage decline in SUV_{bw} max (16.7% ± 36.2%) was significantly lower than mean CV for percentage decline in the longest CT size (156.3% ± 157.3%, $P < 0.0001$) and the 2-dimensional CT size (178.4% ± 546.5%, $P < 0.0001$) (Table 4).

For the 20 tumors with the highest and lowest SUV_{bw} max on the pre- and posttreatment scans, there was almost-perfect reproducibility (ICC > 0.81) for both pre- and posttreatment scans (Supplemental Table 1; supplemental materials are available online only at http://jnm.snmjournals.org). No significant differences in CV of SUV_{bw} max for the 20 lesions with the highest versus the lowest metabolic rates were found within the pretreatment group.

For the 20 largest and 20 smallest tumors, pretreatment reproducibility of both CT size parameters was fair to moderate (ICC range, 0.36–0.43). After treatment, there was substantial reproducibility for the longest CT size (ICC range, 0.73–0.81), but reproducibility was fair to moderate for the 2-dimensional CT size (ICC range, 0.32–0.62). For both CT size parameters, no significant differences were found comparing CV for the largest 20 versus the smallest 20 lesions within pretreatment and posttreatment scans (Supplemental Table 2).

DISCUSSION

This study was designed to expand on previous studies evaluating SUV_{bw} max interobserver reproducibility (*12,15,19–23*) and to compare the interobserver reproducibility of SUV and CT size measurements using clinically available software including posttreatment assessments. Regardless of the statistical method used (ICC and the precision of its estimate or CV), the reproducibility of SUV_{bw} max measurements before and after therapy and percentage change were higher than the reproducibility for CT size measurements. The interobserver variability of SUV_{bw} max was approximately 6% ± 14%—compared with approximately 16%−28% ± 18%−27%—for CT size parameters before therapy, and approximately 18% ± 27%—compared with approximately 35%−51% ± 48%−51%—for CT size parameters after therapy. Perhaps even more importantly, the percentage decline in SUV_{bw} max was much less variable than the percentage decline in tumor size measurements on CT.

The higher variability observed for measurements of tumor size on CT agrees with our hypothesis. Although we applied a routine, vendor-supplied manual technique of region-of-interest (ROI) selection for SUV_{bw} max determination in this study, obtaining tumor measurements on CT requires an even more manual and subjective approach because the reader has to precisely and accurately identify the edges of the tumor. Slightly different angles of measurement could result in greater variability of tumor measurements. The effect is magnified with the 2-dimensional CT size because of increasing error associated with multiplication of 2 uncertain numbers. Studies have demonstrated both good reproducibility and considerable variability of linear CT size measurements (*3,4,27*), and the results are difficult to compare with themselves and our study because of differing methodologies. Interobserver reproducibility of linear size measurements may be improved with semiautomated techniques and volumetric measurements of tumor size (*28–32*).

We found slightly lower interobserver reproducibility of SUV_{bw} max than did previous studies of untreated and treated (*20,21,23*) tumors, 2 of which reported 100% agreement in SUV_{bw} max between 2 readers (*20,23*). Prior studies focused on highly ^{18}F-FDG–avid, typically solitary, tumors. Our tumor population was probably more variable in its characteristics. Our goal was to emulate SUV determination in the clinical setting and across various expertise levels as much as possible. Location and transaxial image number may not have been sufficient for the correct identification of reference tumors, particularly in patients with multiple tumors located in close proximity. On post-analysis rereview of our data, it was determined that at least 1 of the 4 PET readers likely measured a different lesion than the others for 4 tumors with substantial interobserver variability in pretreatment SUV_{bw} max.

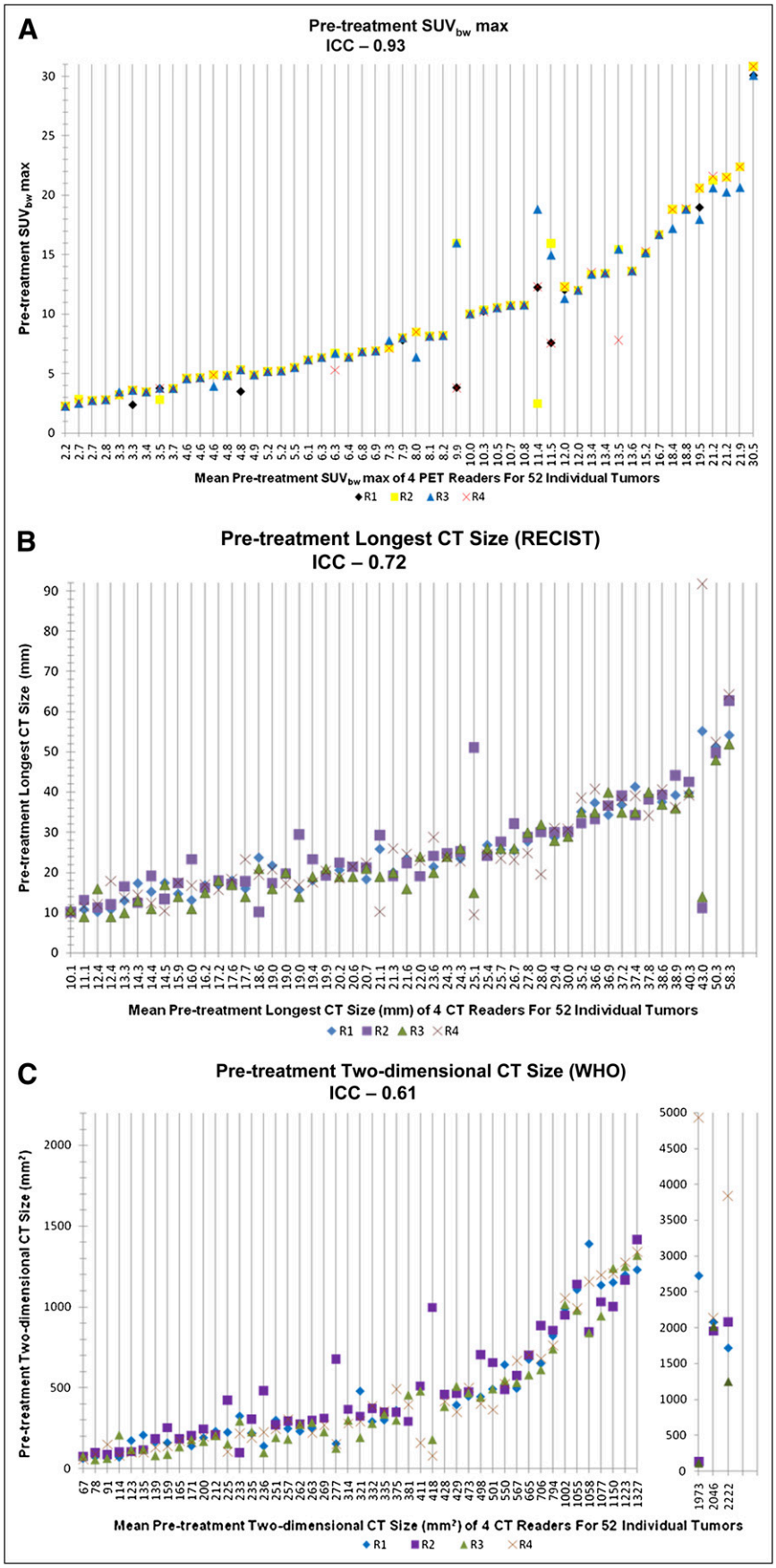

FIGURE 1. Data points for each individual tumor and reader before treatment are shown: SUV_{bw} max (A), longest CT size (B), and 2-dimensional CT size (C). Before treatment, almost-perfect reproducibility of SUV_{bw} max (ICC, 0.93) was better than substantial reproducibility of longest CT size (ICC, 0.72) and 2-dimensional CT size (ICC, 0.61). CT size was not measured by at least 3 readers for 2 tumors, and these tumors were excluded from ICC analysis (last 2 vertical lines). R1 = reader 1; R2 = reader 2; R3 = reader 3; R4 = reader 4.

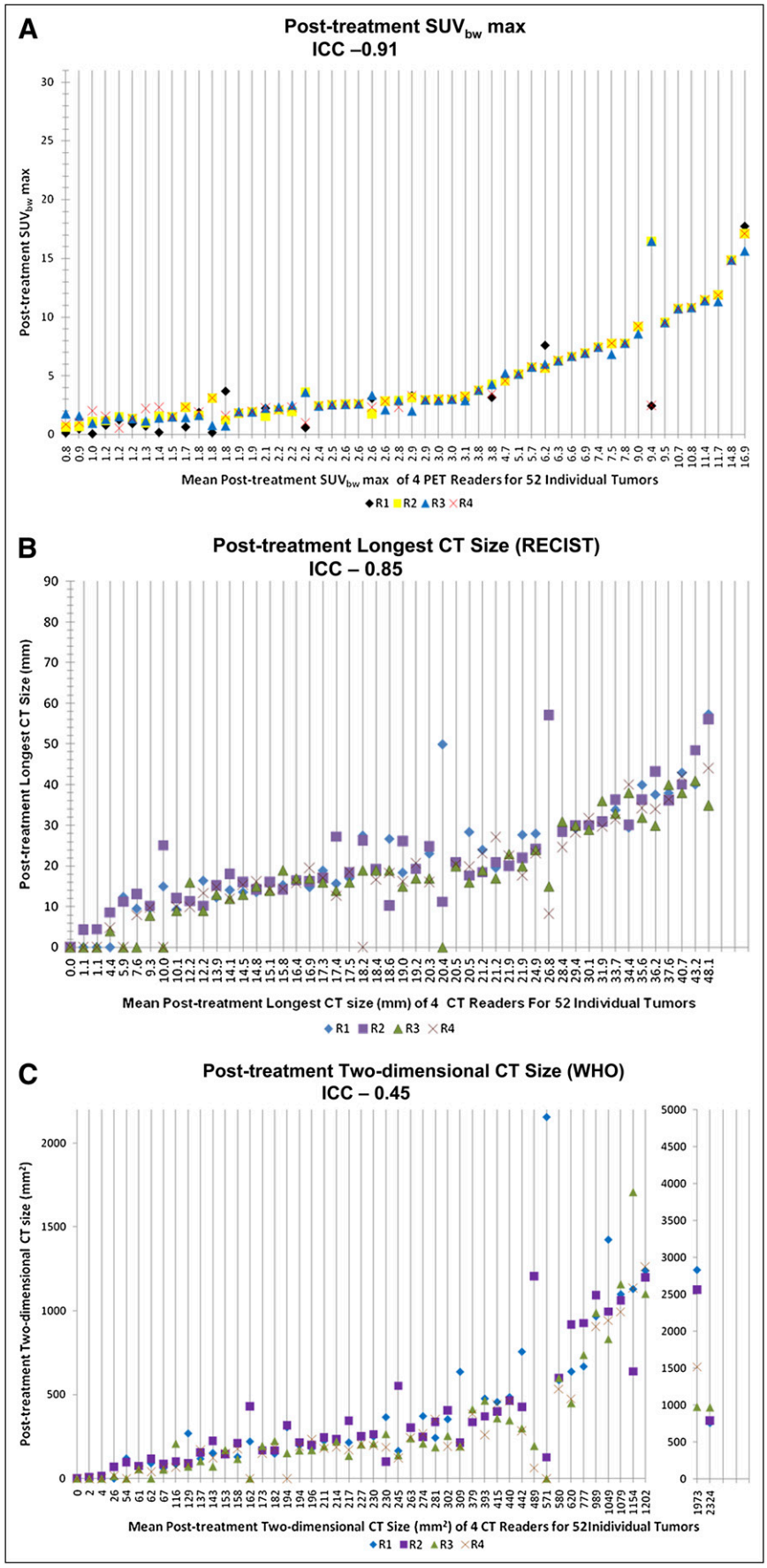

FIGURE 2. Data points for each individual tumor and reader after treatment are shown: SUV_{bw} max (A), longest CT size (B), and 2-dimensional CT size (C). After treatment, SUV_{bw} max (ICC, 0.91) and longest CT size (ICC, 0.85) had almost-perfect reproducibility, although that for SUV_{bw} max was higher. Reproducibility of 2-dimensional CT size was moderate (ICC, 0.45). CT size was not measured by at least 3 readers for 3 tumors, and these tumors were excluded from ICC analysis (last 3 vertical lines). R1 = reader 1; R2 = reader 2; R3 = reader 3; R4 = reader 4.

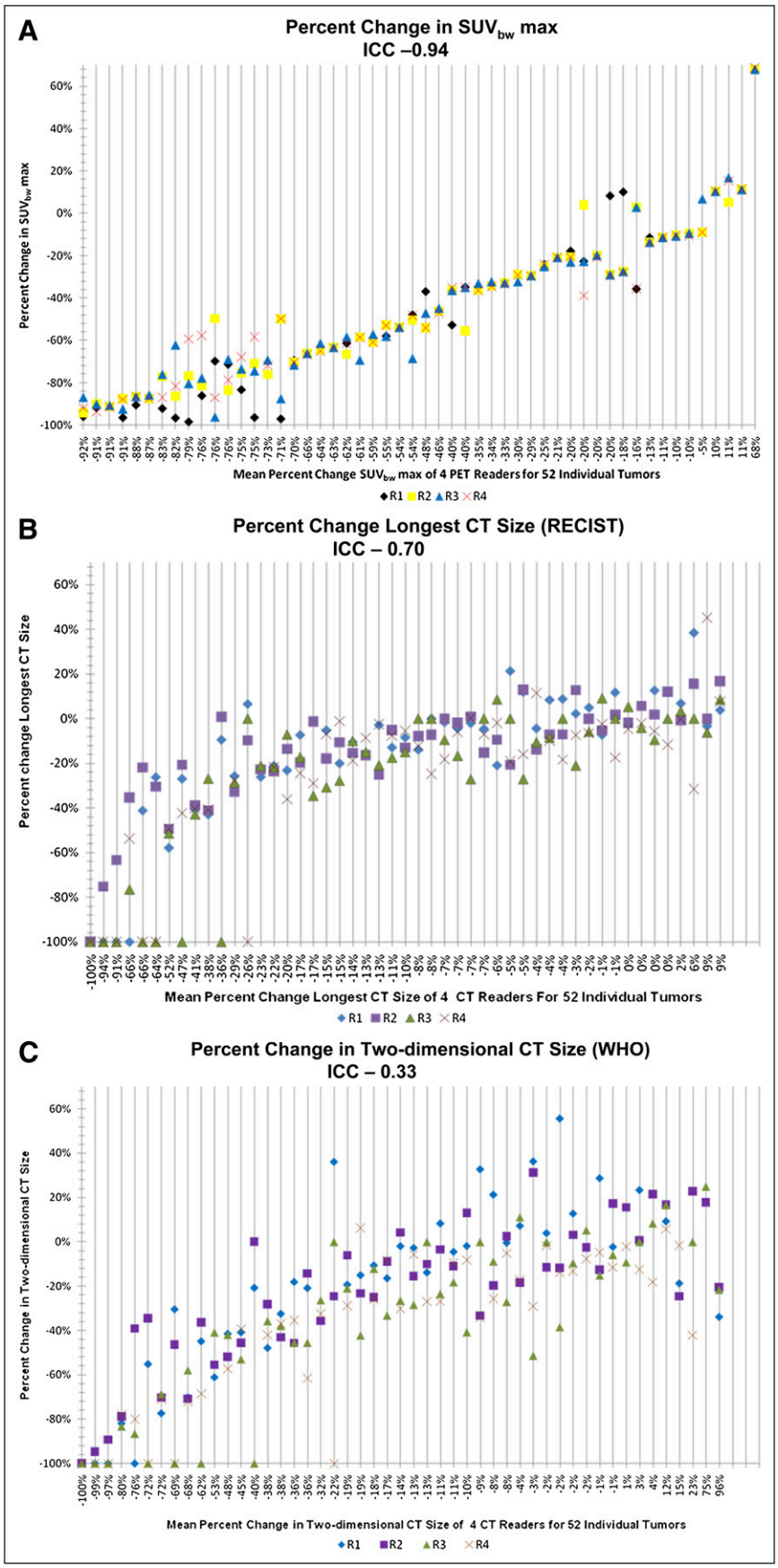

FIGURE 3. Data points for percentage change in each individual tumor for each reader are shown: SUV_{bw} max (A), longest CT size (B), and 2-dimensional CT size (C). Almost-perfect reproducibility of percentage change in SUV_{bw} max (ICC, 0.94) was better than substantial reproducibility of longest CT size (ICC, 0.70) and fair reproducibility of 2-dimensional CT size (ICC, 0.33). Percentage change in CT size was not determined by at least 3 readers for 3 tumors, and these tumors were excluded from ICC analysis (last 3 vertical lines). R1 = reader 1; R2 = reader 2; R3 = reader 3; R4 = reader 4.

TABLE 4. Percentage Variability Between SUV_{bw} max and CT Size Parameters Among Readers

Parameter	Pretreatment scan CV	Posttreatment scan CV	Percentage decline between pre- and posttreatment scans
SUV_{bw} max	6.3% ± 14.2%*†	18.4% ± 26.8%‡	16.7% ± 36.2%§
Longest CT size (RECIST, mm)	16.2% ± 17.8%†‖	35.1% ± 47.5%	156.3% ± 157.3%
Two-dimensional CT size (WHO, mm^2)	27.5% ± 26.7%†	50.9% ± 51.4%	178.4% ± 546.5%

*$P = 0.001$, compared with longest CT size, and $P < 0.0001$, compared with 2-dimensional CT size.
†$P \leq 0.001$, compared with posttreatment scans.
‡$P = 0.02$, compared with longest CT size, and $P < 0.001$, compared with 2-dimensional CT size.
§$P \leq 0.0001$, compared with longest CT size and 2-dimensional CT size.
‖$P = 0.007$, compared with 2-dimensional CT size.
WHO = World Health Organization.
Data are mean ± SD.

We purposely did not specify that to determine whole-tumor SUV_{bw} max, an ROI must be placed around the whole tumor and not on selected slices. This seems obvious to a skilled PET reader, but might elude a less experienced one and result in operator-dependent variations in SUV. Additionally, 1 of our readers rounded some tumor SUVs to 1 decimal place whereas the others always recorded 2. Display of SUVs to 1 or 2 significant figures to the right of the decimal by different software packages could have an effect when small differences are being evaluated, as in this study.

We hypothesized that the interobserver reproducibility of SUV_{bw} max in treated tumors would be lower than that in untreated tumors because of the challenge of selecting ROIs in tumors with lower levels of ^{18}F-FDG uptake and in background tissues (if there was a visual complete response) and more statistical noise (*33*). This effect was not observed based on the analysis using the ICC. The posttreatment tumor mean SUV_{bw} max (4.6 ± 4.0) for all tumors might have been too high to demonstrate such an effect. The higher CV of SUV_{bw} max in treated versus untreated tumors must be viewed with caution because the CV is sensitive to small changes in the mean when it is closer to zero, limiting its usefulness. After treatment, the mean SUV_{bw} max of 15 tumors evaluated in this study was less than 2 (compared with none before treatment). Similar explanations probably apply to the subgroup analyses of the tumors with the highest and least metabolic activity.

In therapy response trials, the percentage change in tumor metabolism or size between 2 scans is potentially more important than the absolute quantitation. Interobserver variability of CT size measurements and, therefore, percentage change in CT size can result in the misclassification of tumor response based on standard anatomic imaging criteria (*5,6*). The probability of misclassifying the response of a lesion has been estimated to be between 8% and 43% because of errors in CT size measurements obtained by different readers (*6*). In our study, the interobserver reproducibility of percentage change in SUV_{bw} max between pre- and posttreatment scans was substantially higher than that for the percentage change in the CT size parameters. On the basis of our data, it appears that percentage change in SUV_{bw} max is a more robust parameter for monitoring treatment response on therapy trials, particularly when assessing for small changes, than changes on CT.

Despite the high interobserver reproducibility of SUV_{bw} max determination, the fact that some variability exists must be considered when defining thresholds for response criteria. Previous recommendations for threshold values to define response and progressive disease using changes in tumor metabolism were based on retrospective studies with limited patient numbers and various tumor and treatment types (*34*). Weber et al. (*35*) prospectively showed that in patients with non–small cell lung cancer the ultimate responders could be separated from the nonresponders using a 20% reduction in tumor SUV after 1 cycle of chemotherapy. Whether a 20% change is large enough to truly indicate a clinically significant response needs further study and may be treatment- and time-after-treatment–dependent.

The retrospective nature of the study is a possible limitation. We attempted to pick easily visualized tumors on the pretreatment scans for evaluation. The lack of intravenous contrast on CT is another potential limitation, particularly for the primary breast tumors, but these accounted for a minority of lesions (16/52). Increasingly, data suggest that there may be no benefit of obtaining a diagnostic CT scan in addition to an ^{18}F-FDG PET/CT scan (*36,37*). Follow-up studies in differing tumor types and with differing uses of intravenous contrast may be helpful to better refine the precision of CT.

CONCLUSION

SUV_{bw} max is highly reproducible when determined by multiple readers with clinically available software from routine ^{18}F-FDG PET scans before and after therapy and is more reproducible than CT size measurements, particularly 2-dimensional CT measurements. Percentage changes in tumor SUV were more highly reproducible than percentage changes in tumor size on CT and should be seriously considered for inclusion in the future establishment of criteria for trials to assess response to therapy. In fact, the

recently proposed PET Response Criteria in Solid Tumors (PERCIST), version 1.0, uses percentage change in peak SUV_{lean} for assessing response after treatment (*38*). That some variability was seen in analyses of the identical image sets by 4 PET readers using clinical software also points to a need for standardized and specific SUV determinations performed by experienced PET readers for clinical trials and automated analytic tools to assess response and improve reproducibility.

ACKNOWLEDGMENT

This study was supported by a grant to the Imaging Response Assessment Teams in Cancer Center Supplement at Johns Hopkins University from the National Cancer Institute (P30 CA006973-43S2).

REFERENCES

1. 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.
2. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer.* 1981;47:207–214.
3. Wormanns D, Diederich S, Lentschig MG, Winter F, Heindel W. Spiral CT of pulmonary nodules: interobserver variation in assessment of lesion size. *Eur Radiol.* 2000;10:710–713.
4. Monsky WL, Raptopoulos W, Keogan MT, et al. Reproducibility of linear tumor measurements using PACS: comparison of caliper method with edge-tracing method. *Eur Radiol.* 2004;14:519–525.
5. Warr D, McKinney S, Tannock I. Influence of measurement error on assessment of response to anticancer chemotherapy: proposal for new criteria of tumor response. *J Clin Oncol.* 1984;2:1040–1046.
6. Erasmus JJ, Gladish GW, Broemeling L, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J Clin Oncol.* 2003;21:2574–2582.
7. Wahl RL, Zasadny K, Helvie M, Hutchins GD, Weber B, Cody R. Metabolic monitoring of breast cancer chemohormonotherapy using positron emission tomography: initial evaluation. *J Clin Oncol.* 1993;11:2101–2111.
8. Gallamini A, Hutchings M, Rigacci L, et al. Early interim 2-[^{18}F]fluoro-2-deoxy-D-glucose positron emission tomography is prognostically superior to international prognostic score in advanced-stage Hodgkin's lymphoma: a report from a joint Italian-Danish study. *J Clin Oncol.* 2007;25:3746–3752.
9. Cheson BD, Pfistner B, Juweid ME, et al. Revised response criteria for malignant lymphoma. *J Clin Oncol.* 2007;25:579–586.
10. 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.
11. Zasadny KR, Wahl RL. Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose: variations with body weight and a method for correction. *Radiology.* 1993;189:847–850.
12. 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.
13. Sugawara Y, Zasadny KR, Neuhoff AW, Wahl RL. Reevaluation of the standardized uptake value for FDG: variations with body weight and methods for correction. *Radiology.* 1999;213:521–525.
14. Hamberg LM, Hunter GJ, Alpert NM, Choi NC, Babich JW, Fischman AJ. The dose uptake ratio as an index of glucose metabolism: useful parameter or oversimplification? *J Nucl Med.* 1994;35:1308–1312.
15. 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.
16. Torizuka T, Fisher SJ, Wahl RL. Insulin-induced hypoglycemia decreases uptake of 2-[F-18]fluoro-2-deoxy-D-glucose into experimental mammary carcinoma. *Radiology.* 1997;203:169–172.
17. Jaskowiak CJ, Bianco JA, Perlman SB, Fine JP. Influence of reconstruction iterations on ^{18}F-FDG PET/CT standardized uptake values. *J Nucl Med.* 2005; 46:424–428.
18. Schoder H, Erdi YE, Chao K, Gonen M, Larson SM, Yeung HW. Clinical implications of different image reconstruction parameters for interpretation of whole-body PET studies in cancer patients. *J Nucl Med.* 2004;45:559–566.
19. 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.
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. Marom EM, Munden RF, Truong MT, et al. Interobserver and intraobserver variability of standardized uptake value measurements in non-small-cell carcinoma lung cancer. *J Thorac Imaging.* 2006;21:205–212.
22. 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.
23. Benz MR, Evilevitch W, Allen-Auerbach MS, et al. Treatment monitoring by ^{18}F-FDG PET/CT in patients with sarcomas: interobserver variability of quantitative parameters in treatment-induced changes in histopathologically responding and nonresponding tumors. *J Nucl Med.* 2008;49:1038–1046.
24. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986;73:13–22.
25. Scheffé H. *The Analysis of Variance.* New York, NY: Wiley; 1959:221–260.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
27. Revel M, Bissery A, Bienvenu M, Aycard L, Lefort C, Frija G. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? *Radiology.* 2004;231:453–458.
28. Schwartz LH, Ginsberg MS, DeCorato D, et al. Evaluation of tumor measurements in oncology: use of film-based and electronic techniques. *J Clin Oncol.* 2000;18:2179–2184.
29. Goodman LR, Gulsun M, Washington L, Nagy PG, Piacsek KL. Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements. *AJR.* 2006;186:989–994.
30. Gietema HA, Wang Y, Xu D, et al. Pulmonary nodules detected at lung cancer screening: interobserver variability of semiautomated volume measurements. *Radiology.* 2006;241:251–257.
31. Fraioli F, Bertoletti L, Napoli A, et al. Volumetric evaluation of therapy response in patients with lung metastases: preliminary results with a computer system (CAD) and comparison with unidimensional measurements. *Radiol Med (Torino).* 2006;111:365–375.
32. Mukherji SK, Toledano AY, Beldon C, et al. Interobserver reliability of computed tomography-derived primary tumor volume measurement in patients with supraglottic carcinoma. *Cancer.* 2005;103:2616–2622.
33. Cherry S, Sorenson J, Phelps M, eds. Nuclear counting statistics. In: *Physics in Nuclear Medicine.* 3rd ed. Philadelphia, PA: W.B. Saunders Company; 2003: 131–140.
34. 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.
35. Weber WA, Petersen V, Schmidt B, et al. Positron emission tomography in non-small-cell lung cancer: prediction of response to chemotherapy by quantitative assessment of glucose use. *J Clin Oncol.* 2003;21:2651–2657.
36. Schaefer NG, Hany TF, Taverna C, et al. Non-Hodgkin lymphoma and Hodgkin disease: coregistered FDG PET and CT at staging and restaging—do we need contrast-enhanced CT? *Radiology.* 2004;232:823–829.
37. Rodriguez-Vigil B, Gomez-Leon N, Pinilla I, et al. PET/CT in lymphoma: prospective study of enhanced full-dose PET/CT versus unenhanced low-dose PET/CT. *J Nucl Med.* 2006;47:1643–1648.
38. 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.