Repeatability of ^{18}F-FDG PET in a Multicenter Phase I Study of Patients with Advanced Gastrointestinal Malignancies

Linda M. Velasquez[1], Ronald Boellaard[2], Georgia Kollia[1], Wendy Hayes[1], Otto S. Hoekstra[2], Adriaan A. Lammertsma[2], and Susan M. Galbraith[1]

[1]Bristol-Myers Squibb Co., Princeton, New Jersey; and [2]Department of Nuclear Medicine and PET Research, VU University Medical Center, Amsterdam, The Netherlands

^{18}F-FDG PET is often used to monitor tumor response in multicenter oncology clinical trials. This study assessed the repeatability of several semiquantitative standardized uptake values (mean SUV [SUV_{mean}], maximum SUV [SUV_{max}], peak SUV [SUV_{peak}], and the 3-dimensional isocontour at 70% of the maximum pixel value [$SUV_{70\%}$]) as measured by repeated baseline ^{18}F-FDG PET studies in a multicenter phase I oncology trial. **Methods:** Double-baseline ^{18}F-FDG PET studies were acquired for 62 sequentially enrolled patients. Tumor metabolic activity was assessed by SUV_{mean}, SUV_{max}, SUV_{peak}, and $SUV_{70\%}$. The effect on SUV repeatability of compliance with recommended image-acquisition guidelines and quality assurance (QA) standards was assessed. Summary statistics for absolute differences relative to the average of baseline values and repeatability analysis were performed for all patients and for a subgroup that passed QA, in both a multi- and a single-observer setting. Intrasubject precision of baseline measurements was assessed by repeatability coefficients, intrasubject coefficients of variation (CV), and confidence intervals on mean baseline differences for all SUV parameters. **Results:** The mean differences between the 2 SUV baseline measurements were small, varying from −2.1% to 1.9%, and the 95% confidence intervals for these mean differences had a maximum half-width of about 5.6% across the SUV parameters assessed. For SUV_{max}, the intrasubject CV varied from 10.7% to 12.8% for the QA multi- and single-observer datasets and was 16% for the full dataset. The 95% repeatability coefficients ranged from −28.4% to 39.6% for the QA datasets and up to −34.3% to 52.3% for the full dataset. **Conclusion:** Repeatability results of double-baseline ^{18}F-FDG PET scans were similar for all SUV parameters assessed, for both the full and the QA datasets, in both the multi- and the single-observer settings. Centralized quality assurance and analysis of data improved intrasubject CV from 15.9% to 10.7% for averaged SUV_{max}. Thresholds for metabolic response in the multicenter multiobserver non-QA settings were −34% and 52% and in the range of −26% to 39% with centralized QA. These results support the use of ^{18}F-FDG PET for tumor assessment in multicenter oncology clinical trials.

Received Mar. 8, 2009; revision accepted Jul. 10, 2009.
For correspondence or reprints contact: Linda M. Velasquez, Bristol-Myers Squibb Co., Room E1335, P.O. Box 4000, Princeton, NJ 08543.
E-mail: linda.velasquez@bms.com

Key Words: SUV; repeatability; ^{18}F-FDG PET; multicenter clinical trial

J Nucl Med 2009; 50:1646–1654
DOI: 10.2967/jnumed.109.063347

PET, with the tracer ^{18}F-FDG, is used for tumor detection, staging, and follow-up studies for multiple neoplasms (*1*) and is increasingly becoming an integral part of multicenter clinical trials in oncology for the assessment of treatment effect. Accurate quantitative assessment of response as measured by changes in standardized uptake value (SUV) parameters over the course of treatment serves as an early surrogate for clinical benefit and facilitates drug development in oncology (*2*).

For the accurate assessment of tumor response using ^{18}F-FDG PET, it is crucial to know the intrasubject variation in the measurement of semiquantitative parameters before the initiation of treatment (*3*). This study focused on the repeatability of ^{18}F-FDG PET in a multicenter phase I study. For this study, repeatability is defined by the British Institution of Standards as the variation of repeated measurements in an experiment performed under the same conditions (*4*).

Repeatability results of quantitative parameters derived from ^{18}F-FDG PET studies have been well published (*5–9*). Two single-center studies, focusing on double-baseline ^{18}F-FDG PET studies, have reported up to 12% variation in relative absolute percentage difference (*5,6*) and a 15%−20% repeatability coefficient (RC) (*6*). Weber et al. (*6*) reviewed double-baseline ^{18}F-FDG PET studies performed in a single setting with 16 patients and 50 separate tumor lesions including the primary tumor and liver, lung, and lymph node metastasis. Similarly, repeated baseline measurements of SUV showed an SD of the mean percentage difference of approximately 10%. In the review by Weber et al. (*6*)—although RCs (reference ranges) were calculated for SUV measurements with and without glucose correction—mean SUV (SUV_{mean}), maximum SUV (SUV_{max}), peak SUV (SUV_{peak}), and the 3-dimensional isocontour at 70% of the

maximum pixel value ($SUV_{70\%}$) parameters (SUV measurements derived using different region-of-interest [ROI] methods) were not evaluated individually for repeatability. Kamibayashi (*5*) reviewed double-baseline ^{18}F-FDG PET studies in 45 patients with tumors of the lung on 2 different scanners in the same institution. SUV_{mean} and SUV_{max}, tumor-to-mediastinum and tumor-to-liver ratios, and the relative absolute baseline difference in parameter values between the 2 PET images were calculated. No statistically significant differences between the 2 PET images were observed, except for SUV_{max} in the liver and tumor-related parameters, tumor to mediastinum and tumor to liver.

Hoekstra et al. (*10*) published data on SUV variability in a multicenter setting; however, in the study by Hoekstra et al., data were collected at 2 sites only. Studies assessing the repeatability of the SUV parameters SUV_{mean}, SUV_{max}, SUV_{peak}, and $SUV_{70\%}$ on double-baseline studies for ^{18}F-FDG PET in a larger multicenter setting have not been previously reported.

The goal of this study was to assess the repeatability of select SUV measurements on double-baseline ^{18}F-FDG PET studies and to assess the effect of site compliance with recommended methodologic guidelines, overall data quality, and reader setting on scan data collected in a multicenter setting. Different approaches to explore the variability of baseline SUV changes will be presented, to allow for a comparison with results in similar publications (*6,8*).

MATERIALS AND METHODS

Patient Population

Sixty-two patients (38 men, 24 women; mean age, 58 ± 11 y; range, 28–78 y) with advanced gastrointestinal malignancies (60 patients with colorectal carcinoma, 1 patient with esophageal carcinoma, and 1 patient with hepatocellular carcinoma), who failed prior therapy and had evaluable metastatic lesions, were included. A single patient was excluded from the dataset because of a limited field of view and the inability to identify suitable lesions for longitudinal assessment. The lesions selected for the remainder of the patients ($n = 145$) for repeatability assessment and longitudinal follow-up were primarily hepatic (65%) and lung (26%) metastases. The remaining 9% of lesions included lymph node, bone, gastric, intestinal, and kidney metastases. A total of 8 academic sites (5 in the United States, 2 in Canada, and 1 in The Netherlands) performed the ^{18}F-FDG PET studies. At each site, for the 2 wk before the baseline ^{18}F-FDG PET scan, no therapy (chemotherapy, radiotherapy, or surgical treatment) was administered to any of the patients. After patients signed the appropriate informed consent form, ^{18}F-FDG PET was scheduled to be performed on all patients enrolled in the clinical trial. The study was approved by the medical ethics review board of each participating institution.

The patient ^{18}F-FDG PET scans were grouped into 3 datasets for this study, defined as follows: full dataset (multiobserver), patients with double-baseline ^{18}F-FDG PET studies analyzed with local software at each imaging site; quality assurance (QA) dataset (multiobserver), patients with double-baseline ^{18}F-FDG PET studies analyzed with local software at each imaging site that passed a QA assessment on central review; and QA dataset (single-observer), patients with double-baseline ^{18}F-FDG PET studies that passed a QA assessment and were analyzed at the central image-analysis laboratory using a single software platform on central review.

^{18}F-FDG PET

Double-baseline ^{18}F-FDG PET studies were performed within 7 d (4.1 ± 2.6 d) of each other and within 14 d of the start of therapy.

Protocol-specified ^{18}F-FDG PET procedures were established from published recommendations for the use of ^{18}F-FDG PET in the assessment of response to therapy in oncology trials (*11–14*) in conjunction with local institutional procedures and standards. The specifications included that the ^{18}F-FDG PET studies should be performed at the same facility, with the same equipment and personnel and be processed with the same attenuation and reconstruction methods.

Patients were instructed to fast for a minimum of 4 h before the ^{18}F-FDG PET study and refrain from strenuous activity. Serum glucose measurements were recorded before ^{18}F-FDG administration. The time of the last insulin or hypoglycemic agent dose for diabetic patients was recorded. Acceptable serum glucose concentration levels were defined as less than 11.1 mmol/L.

The dose of administered ^{18}F-FDG ranged from 185 to 740 MBq. The tracer dose, tracer dose assay time, and exact time of injection were recorded. Static emission images covering the area of tumor involvement were to be acquired between 50 and 70 min after ^{18}F-FDG administration. The period between tracer injection and the start of the scan was documented, and subsequent studies were to be performed within a 30-min window (±15 min). In addition to the emission scan, a (low-dose) CT scan or a transmission scan was acquired for attenuation-correction purposes. Apart from the guidelines specified in the study protocol, PET or PET/CT studies were collected and reconstructed according to local guidelines.

PET Data Analysis

ROIs were drawn on up to 3 target lesions from a subset of lesions selected for anatomic measurement on the basis of modified World Health Organization criteria, based on a baseline CT scan. The recommended minimum tumor size was at least 2 times the spatial resolution of the PET scanner and was determined locally. The number of pixels in each of the ROIs was reported and reviewed to ensure selection of comparable areas of tumor and to assess variation in the ROI selection within a patient.

SUV measurements were corrected for lean body mass (*15,16*) based on the Hume method (*17*).

$$\text{SUV} = \frac{\text{measured activity concentration (Bq/g)} \times \text{lean body mass (g)}}{\text{injected activity (Bq)}}$$

SUV_{mean}, SUV_{peak}, and SUV_{max} were calculated by each site using their respective software analysis packages. These SUV parameters, along with $SUV_{70\%}$, were also analyzed centrally by the VU University Medical Center. Specific SUV parameter definitions are outlined in Table 1.

Statistical Methods

SUV_{max}, SUV_{mean}, SUV_{peak}, and $SUV_{70\%}$ were measured in up to 3 lesions per patient on the 2 baseline studies. The same lesions were analyzed and compared for both studies. Analysis of repeatability of these parameters was performed on a patient-by-patient basis. Each patient's individual SUV parameters from the selected lesions were summarized across lesions using 2 derived

TABLE 1. SUV Parameter Definitions

Parameter	Definition
SUV_{mean}	SUV for activity in largest diameter of tumor and 2 adjacent slices, representing largest cross-section of tumor
SUV_{peak}	SUV of 1-cm ROI (0.75–1.25 cm) placed in region of highest ^{18}F-FDG uptake
SUV_{max}	SUV for single pixel with highest activity in tumor
$SUV_{70\%}$	SUV generated using 70% threshold of maximum tumor SUV and isocontour adapted for local background

measurements (average value defined as the average of the SUV parameter values across lesions, and maximum value defined as the lesion with the maximum SUV value).

For each SUV parameter and patient (i), the differences (di) between the 2 baseline scan (average or maximum) values were calculated. An initial assessment of variability of SUV percentage changes at baseline was based on the patient's absolute differences |di|, relative to the patient's average (μi) of the 2 baseline values, expressed as a percentage:

$$\text{RelAb\_d} = 100 \cdot |\text{di}|/\mu\text{i}. \qquad \text{Eq. 1}$$

As SUV is known to have a log normal distribution (*18*), the data was log-transformed before most analysis, and the results were expressed as percentage changes. To confirm the appropriateness of using percentage changes in this study, Kendall τ correlation statistic and diagnostic plots were used in the original and log-transformed (or percentage) scales.

For each parameter, to estimate the mean difference in 2 measurements from a sample of size n, point estimates and 95% confidence intervals (CIs) were calculated on log-transformed data. Exponentiation was applied to these results to express the differences as ratios on the original scale and report them as percentage differences:

$$\text{CI}_d = 100 \cdot (\exp(đ_{ln} \pm 1.96\text{SD}_{dln}/\sqrt{(n)}) - 1), \qquad \text{Eq. 2}$$

where $đ_{ln}$ is the mean difference, and SD_{dln} is the SD of the difference on the log scale.

To calculate the RC for each parameter, the within-subject SD, wSD_{ln}, of the log-transformed measurements was determined. wSD_{ln} can be obtained from the SD of the differences, di_{ln}, assuming the repeated measurements are from a distribution with common variance (as described in the supplemental materials, which are available online only at http://jnm.snmjournals.org):

$$\text{wSD}_{ln} = \text{SD}_{dln}/\sqrt{2}. \qquad \text{Eq. 3}$$

Exponentiation was applied to the results on the log-transformed scale to calculate the within-subject coefficient of variation (wCV) (%), and the results were expressed as a percentage:

$$\text{wCV}(\%) = 100 \cdot (\exp(\text{wSD}_{ln}) - 1). \qquad \text{Eq. 4}$$

The 95% RC for each parameter was then calculated as described by Bland and Altman (*19*); it was first obtained on the log-transformed data (RC_{ln}). Using the expression $RC_{ln} = \pm 1.96 \cdot SD_{dln} = \pm 2.77 \cdot wSD_{ln}$, we applied exponentiation and multiplied by 100 to express it as a percentage:

$$\text{RC} = 100 \cdot (\exp(\pm 1.96 \cdot \text{SD}_{dln}) - 1). \qquad \text{Eq. 5}$$

RCs from log-transformed data are nonsymmetric and presented as lower and upper RCs (LRC and URC, respectively). The precision of the RCs was also assessed by 95% CIs using the χ^2 distribution (supplemental materials).

The results were visualized graphically for the parameters averaged across lesions by Bland–Altman plots on individual patients' percentage differences versus their average μi overlaid with the RC (LRC, URC) reference lines and with the 95% CIs for the mean percentage difference.

In the full dataset, the effect of clinical site, scan time relative to the dose (50–70 min), between-scan difference in relative time of scan (±15 min), and diabetic status on the SUV_{max} differences were explored by a general linear model 4-way ANOVA. The model estimated the effect of these parameters on the magnitude of the SUV_{max} differences.

In addition, for the QA multiobserver dataset, the mean (±SD) for absolute baseline percentage differences in each SUV parameter was tabulated by compliance status for the required scan time parameters. Distribution plots of absolute values of percentage differences were also presented by site for each of the SUV parameters using the average across lesions.

RESULTS

Compliance and QA

The patient-preparation procedures, such as length of fast, blood glucose concentration, and hypoglycemic control, are summarized as follows: the mean (±SD) blood glucose concentrations for each of the 2 baseline ^{18}F-FDG PET scans were 5.7 ± 1.2 mmol/L (range, 3.2–8.6 mmol/L) and 5.7 ± 1.4 mmol/L (range, 2.8–11.6 mmol/L). One of the 8 diabetic patients had poor glycemic control on scan 2 (scan 1, 2.9 mmol/L; and scan 2, 11.6 mmol/L). Glucose values were not reported for 2 patients. All patients fasted for at least 4 h before scanning. The 3 patients with missing or elevated glucose values were considered QA failures. Tracer extravasation occurred in a single patient, resulting in the removal of this patient from the QA dataset.

In addition to the assessment of compliance with requested acquisition and patient-preparation parameters, a technical QA assessment was performed centrally (VU University Medical Center). Two patients did not have scans submitted for this analysis. Three patients had blank or unreadable compact disks. Seven patients had irresolvable issues resulting from changes in technology or Digital Imaging and Communications in Medicine inconsistencies during the trial. On the basis of compliance and technical quality, a set of 45 patients comprises the QA dataset (Fig. 1).

Table 2 shows summary statistics and frequency of the scan acquisition parameters, ^{18}F-FDG dose, scan start time relative to ^{18}F-FDG dose administration (50–70 min),

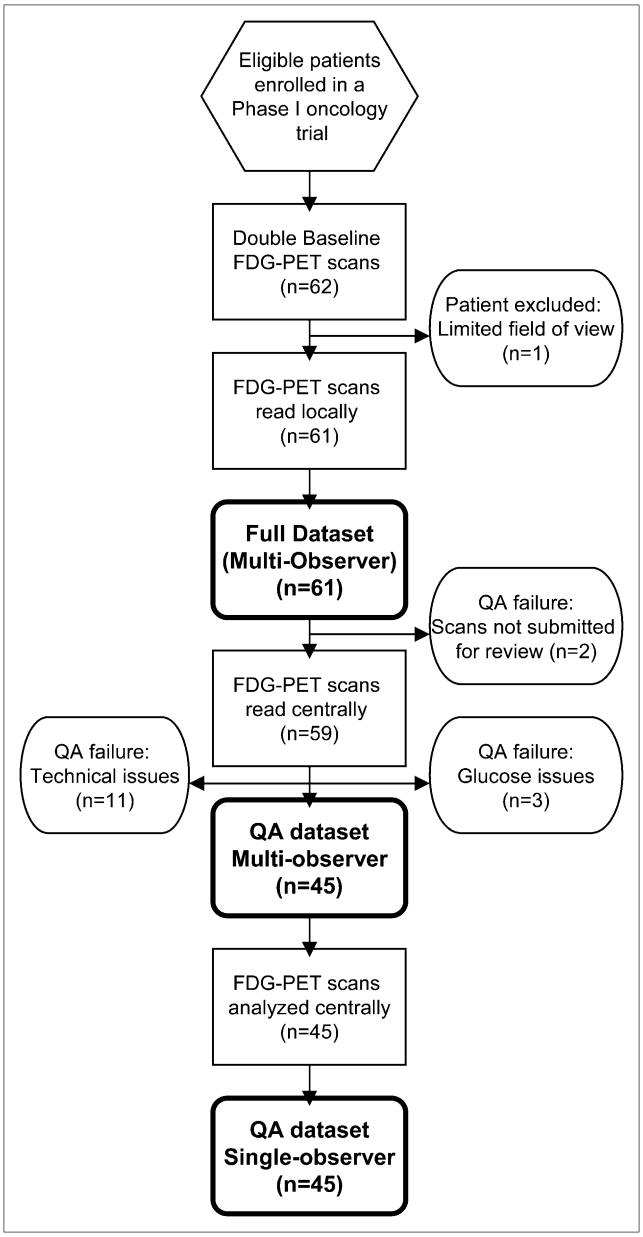

FIGURE 1. Standards for Reporting of Diagnostic Accuracy Studies–style diagram outlining patient image flow through acquisition and analysis process and resultant datasets based on compliance and QA standards.

between-scan time difference in the relative scan times (required within ±15 min), number of days between the baseline scans (required within 7 d), and acceptable data passing QA assessment, by study site and overall.

Visual inspection of the baseline differences on the log scale, for example, by normal probability and distribution plots, indicated approximately normal distributions for the baseline differences in SUV parameters.

Results of the statistical analysis on SUV differences, assessing the effects of site, scan time relative to ^{18}F-FDG dose, between-scan time difference, and diabetic status, demonstrated that the average size of SUV_{max} differences across sites varied from 8% to 24%. Patients without glucose control had SUV differences of 14%, versus 4% for patients with glucose control. This analysis excluded a patient who had an out-of-range glucose value in 1 scan. Overall, site, diabetic status, and scan time parameters did not appear to affect average SUV changes in this study.

SUV Parameters

Absolute baseline percentage differences were summarized by scan time relative to dose and between-scan-time relative differences for SUV parameters averaged across lesions (Table 3). In the QA and full datasets, for patients whose scans were not compliant with the timing recommendations, either outside the 50- to 70-min window (47% and 51%, respectively) or exceeding 15 min in relative time between the 2 scans (24% and 30%, respectively), the differences in SUV_{mean} and SUV_{peak} were similar to those for scans meeting both criteria. Absolute percentage differences were larger in baseline SUV_{max} for patients outside the 50- to 70-min window and exceeding the 15-min relative time between scan recommendations, particularly in the full dataset.

Figure 2 shows the distribution of the absolute values of percentage differences in the 2 baseline scans presented by study site, using averages across lesions. Some variability was noted across sites but was comparable among the 3 parameters in the QA multiobserver dataset (Figs. 2A and 2C) and only somewhat higher for SUV_{mean} (Fig. 2B).

Repeatability Assessment

To assess the effect of the QA procedures, repeatability analysis was performed for SUV_{max} for the full dataset ($n = 61$) and for the datasets that passed the QA assessment ($n = 45$) in both the multi- and the single-observer settings. Summary statistics (means and SD) for absolute differences relative to the average of baseline values as in Equation 1, based on average and maximum across lesions, are presented in Table 4. These results reflect a reduction in both the absolute differences and the variability on central QA assessment (QA multiobserver) and a further subtle decrease in variability on central data analysis (QA single-observer).

The intrasubject precision of baseline measurements was assessed by RCs for the individual patient differences, by intrasubject CVs and by CIs on the mean differences. Analysis of SUV_{max} was performed for the full and QA datasets, in the multi- and single-observer settings. SUV_{mean}, SUV_{peak}, and $SUV_{70\%}$ were assessed for the QA multi- and single-observer datasets only (Table 5).

A test of association using the Kendall τ rank correlation statistic for the absolute differences $|di|$ and averages μi on the original scale showed statistically significant results for all parameters. This analysis and the diagnostic plots (Fig. 3A) indicated a dependence of the size of the SUV differences on the size of the parameter value. In contrast, Kendall τ statistic on log-transformed data showed a lack of

TABLE 2. Site Compliance with Select Image Acquisition Parameters and Overall Quality

Site no. (*n*)	^{18}F-FDG dose (MBq) (mean [SD])		Scan start time relative to ^{18}F-FDG dose (min) (mean [SD])		Patients within 50–70 min (%)		Between-scan time difference (min) (mean [SD])	Patients within 15 min (%)	Days between baseline scans (mean [SD])	Patients with acceptable QA data (%)
	Scan 1	Scan 2	Scan 1	Scan 2	Scan 1	Scan 2				
1 (2)	570 (4)	551 (11)	52 (1)	55 (3)	100	100	3 (4)	100	4.3 (0.4)	0
2 (8)	470 (107)	470* (96)	55 (0)	55 (0)	100	100	0 (0)	100	5.3 (1.0)	88
3 (3)	596 (30)	640 (70)	77 (14)	67 (14)	33	67	−10 (25)	67	7.0 (1.3)	100
4 (18)	448 (48)	448 (56)	109 (31)	109 (25)	0	0	−1 (37)	39	3.5 (2.1)	61
5 (5)	426 (104)	418 (89)	63 (27)	80 (11)	60	40	17 (35)	40	8.0 (2.4)	80
6 (11)	574 (33)	577 (30)	61 (6)	53 (16)	91	73	−8 (20)	82	2.9 (1.7)	82
7 (13)	418 (85)	426 (85)	64 (6)	68 (7)	92	77	4 (8)	92	2.8 (3.0)	85
8 (1)	307 (—)	229 (—)	54 (—)	66 (—)	100	100	12 (—)	100	7.0 (—)	0
All (61)	470 (93)	474 (100)	76 (29)	76 (28)	61	54	0.4 (25)	70	4.1 (2.6)	74

*^{18}F-FDG dose for could not be confirmed for single patient.

statistically significant correlation of differences $|di_{ln}|$ with the means, and scatter plots on percentage changes showed less dependence on the size of the measurements (Fig. 3B). This supports the selection of percentage changes in this study as a more appropriate measurement for assessing repeatability.

The mean percentage differences between baseline measurements ranged from −2.1% to 1.9% across the parameters, and the 95% CIs had a maximum half-width of 5.6% (Table 5). The intrasubject CV for SUV_{max} was approximately 16% for the full dataset and 10%−12% for patients in the QA datasets. Repeatability was similar for all SUV parameters across settings, with lower RCs for SUV_{max} for the QA datasets (up to −26.8% and 36.7% [single-observer] and −26.2% and 35.6% [multiobserver]) and for the full multiobserver dataset (up to −34.3% and 52.3%). There was somewhat smaller variability with the performance of a centralized single-observer QA assessment for the SUV_{max} calculated as mean of parameter values across lesions.

The individual patient percentage changes in the SUV_{max} parameter for the full multiobserver, QA multiobserver, and QA single-observer datasets, with the 95% RCs and CIs, are presented by Bland–Altman (*19*) plots based on averages across lesions (Figs. 4A−4C). Centralized QA has the largest impact, with some further, but smaller, improvement with single-observer data analysis.

DISCUSSION

^{18}F-FDG PET studies are increasingly implemented as an objective method for response assessment in drug development. For accurate and reproducible quantitative assessment,

TABLE 3. Summary of Baseline Absolute Percentage Difference for SUV Parameters by Scan Time

Acquisition parameter (*n*)	SUV_{mean}		SUV_{max}		SUV_{peak}	
	Mean	SD	Mean	SD	Mean	SD
QA dataset (*n* = 45)						
Scan time relative to ^{18}F-FDG dose						
50–70 min* (*n* = 24)	11.0	10.0	10.9	7.2	11.3	8.2
<50 or >70 min† (*n* = 21)	10.0	8.3	13.8	11.5	13.4	9.3
Between-scan relative difference						
≤15 min (*n* = 34)	11.1	9.2	12.7	9.9	12.6	8.8
>15 min (*n* = 11)	8.7	9.4	10.9	8.2	11.4	
Full dataset (*n* = 61)						
Scan time relative to ^{18}F-FDG dose						
50–70 min* (*n* = 30)	14.0	18.1	12.5	12.9	15.8	20.7
<50 or >70 min† (*n* = 31)	13.9	13.4	18.1	18.6	17.4	15.7
Between-scan relative difference						
≤15 min (*n* = 43)	13.0	15.6	13.1	13.0	15.6	18.1
>15 min (*n* = 18)	16.4	16.5	20.7	21.4	19.0	18.8

*Patients meeting criteria for both scans.
†Patients missing 50- to 70-min criterion for at least 1 scan.

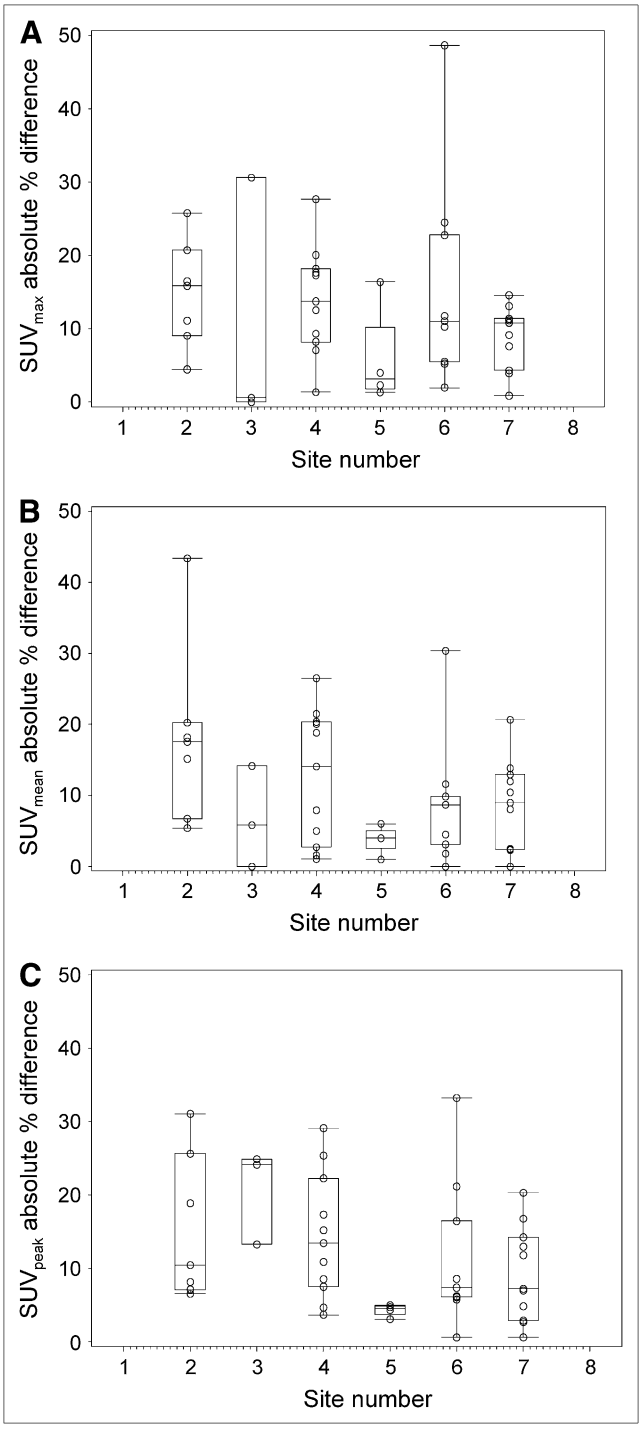

FIGURE 2. Distribution of absolute percentage differences at baseline by site and SUV parameter: (A) SUV_{max}, (B) SUV_{mean}, and (C) SUV_{peak}. Boxes represent values between 25th and 75th percentiles; horizontal lines (within boxes) indicate median; and box plot whiskers (above and below boxes) represent values at 10th and 90th percentiles.

standardization of ^{18}F-FDG PET methodology, including patient preparation, image scan acquisition guidelines, and image analysis, is essential, particularly in the multicenter setting.

Site compliance with several common parameters used in the acquisition of ^{18}F-FDG PET and basic QA were evaluated. Overall compliance with successful scan acquisition (123/124 expected scans) was excellent. Site-reported data for 61 patients revealed acceptable repeatability. An effort to corroborate the site-reported data by central review resulted in a smaller dataset, predominately because of QA issues.

Patient-Related Parameters

To account for changes in blood glucose concentration (*20*), which may affect SUV, it is recommended that patients fast for at least 4 h before the ^{18}F-FDG PET study, that accurate blood glucose concentration be measured before the scan, and that the patient's diabetic status be documented. Fasting blood glucose concentration was within the ^{18}F-FDG PET guideline recommendation, defined in this study as less than 11.1 mmol/L for both diabetic and nondiabetic patients with the exception of a single diabetic patient. High serum glucose concentration can diminish the accuracy of the SUV determination, and the single patient outlier with an elevated glucose value (11.6 mmol/L) did show large baseline differences in all SUV parameters. On the basis of the data from 5 diabetic patients included in this study, repeatability was not affected by a patient's diabetic status, as long as glucose concentration was controlled (within acceptable range for this study) at the time of the ^{18}F-FDG PET scan.

Image-Acquisition Parameters

The consensus recommendation (*11*) for the collection of a static scan at 60 min after the intravenous injection of ^{18}F-FDG and a ±15-min window between scans of a patient was used in this study. The lack of compliance with the study-recommended timing for scan performance had the greatest effect on SUV_{max} in the full dataset. In addition, deviation from consensus guidelines resulted in increased baseline absolute differences for all SUV parameters with a greater than 15-min scan-to-scan time. Because ^{18}F-FDG continues to accumulate for 150 min, ^{18}F-FDG uptake values can be variable at different times in the uptake period (*21*), thus ensuring scan performance within the recommended 50- to 70-min window; the interscan time frame of ±15 min is good practice.

QA Assessment

In an effort to corroborate site-reported SUV data, submission of scan data for central review was requested. In this study, overall quality was acceptable; however, to improve quality in a multicenter setting, the rescheduling of patients in specific instances is recommended (i.e., unacceptable blood glucose elevation or tracer extravasation). Assessing quality in real time, following stringent guidelines regarding the format of the image submission, and ensuring local system back-up of the data may prevent loss of data due to resolvable technical issues.

TABLE 4. SUV_{max} Relative Absolute Baseline Differences (RelAb_d*) (%)

SUV_{max} value across lesions	Full dataset (n = 61) Multiobserver Mean	Full dataset (n = 61) Multiobserver SD	QA dataset (n = 45) Multiobserver Mean	QA dataset (n = 45) Multiobserver SD	QA dataset (n = 45) Single-observer Mean	QA dataset (n = 45) Single-observer SD
Average SUV_{max}	14.7	14.3	12.2	9.3	11.7	8.4
Maximum SUV_{max}	14.7	15.0	12.1	9.3	12.7	9.3

*RelAb_d = absolute percentage difference at baseline relative to average of 2 baseline values.

Image-Analysis Parameters

Ideally, a method for ROI definition should be simple, reproducible, generally applicable, and user-independent (*7*). In this study, the different SUV parameters (SUV_{mean}, SUV_{peak}, and SUV_{max}) resulted in similar levels for repeatability. An additional parameter, $SUV_{70\%}$, generated using a 70% threshold of the maximum tumor SUV and isocontour-adapted for local background, was also assessed. The repeatability was similar for all studied SUV parameters, evaluated either by the lesion with the highest SUV or by the average SUV across lesions, showing only slight variation among the RCs. Use of a single software platform for defining ROIs and SUV calculation may further enhance test–retest variability, as suggested by the somewhat better test–retest data (Table 4) of single- (central) versus multiobserver analysis. This may be important in a response-monitoring setting and in avoiding incorrect SUV response assessments because of technical, data entry, or human error.

Approaches for Assessing Variability of SUV

In this study, various approaches for assessing the variability of SUV differences are presented, including RCs, intrasubject CV, and absolute percentage or relative

TABLE 5. Summary of Repeatability Analysis Results* and Other Descriptive Statistics for SUV Parameters for All Datasets

Measured parameter values across lesions	Mean	Mean % difference: Point estimate	Mean % difference: 95% CI_d	wCV (%)	95% RC LRC (%)	95% RC URC (%)	95% CI for LRC (%)	95% CI for URC (%)
Multiobserver mean†								
$SUV_{max\text{-}full}$	5.76	1.9	−3.3 to 7.3	15.9	−33.6	50.6	−39.2 to −29.3	41.5 to 64.6
SUV_{max}	5.73	0.0	−4.4 to 4.7	11.6	−26.2	35.6	−31.9 to −22.3	28.7 to 46.9
SUV_{mean}	3.16	−2.1	−6.1 to 2.1	10.6	−24.3	32.1	−29.7 to −20.6	25.9 to 42.2
SUV_{peak}	4.85	−0.5	−4.9 to 4.1	11.5	−26.1	35.2	−31.7 to −22.1	28.9 to 46.4
Multiobserver maximum‡								
$SUV_{max\text{-}full}$	7.08	1.4	−3.9 to 7.0	16.4	−34.3	52.3	−40.1 to −30.0	42.9 to 66.9
SUV_{max}	7.03	−0.2	−4.6 to 4.4	11.5	−26.2	35.5	−31.9 to −22.2	28.6 to 46.8
SUV_{mean}	3.82	−1.1	−5.4 to 3.2	11.1	−25.4	34.0	−30.9 to −21.5	27.4 to 44.7
SUV_{peak}	6.00	−0.4	−4.9 to 4.3	11.9	−26.8	36.5	−32.5 to −22.7	29.4 to 48.2
Single-observer mean§								
SUV_{max}	7.45	0.4	−3.8 to 4.8	10.7	−24.9	33.1	−30.3 to −21.1	26.7 to 43.5
SUV_{mean}	4.70	0.9	−3.7 to 5.8	12.0	−27.0	36.9	−32.8 to −22.9	29.7 to 48.7
SUV_{peak}	6.45	−1.4	−5.7 to 3.2	11.6	−26.2	35.6	−31.9 to −22.3	28.6 to 46.9
$SUV_{70\%}$	5.86	0.3	−5.2 to 4.3	10.6	−24.4	32.3	−29.8 to −20.7	26.1 to 42.5
Single-observer maximum‖								
SUV_{max}	8.87	0.9	−3.7 to 5.7	11.9	−26.8	36.7	−32.6 to −22.8	29.5 to 48.4
SUV_{mean}	5.41	1.1	−3.8 to 6.3	12.8	−28.4	39.6	−34.5 to −24.1	31.8 to 52.4
SUV_{peak}	7.75	0.6	−5.4 to 4.4	12.8	−28.3	39.4	−34.3 to −24.1	31.7 to 52.2
$SUV_{70\%}$	6.91	1.1	−3.4 to 5.9	11.8	−26.6	36.2	−32.3 to −22.5	29.1 to 47.7

*Exponentiation was applied to results from analyses on log scale, and results were expressed as percentages.
†Full, n = 61; QA, n = 45.
‡Full, n = 61; QA, n = 45.
§QA, n = 45.
‖QA, n = 45.
CI_d = 95% confidence interval for mean difference; wCV = within-subject coefficient of variation.

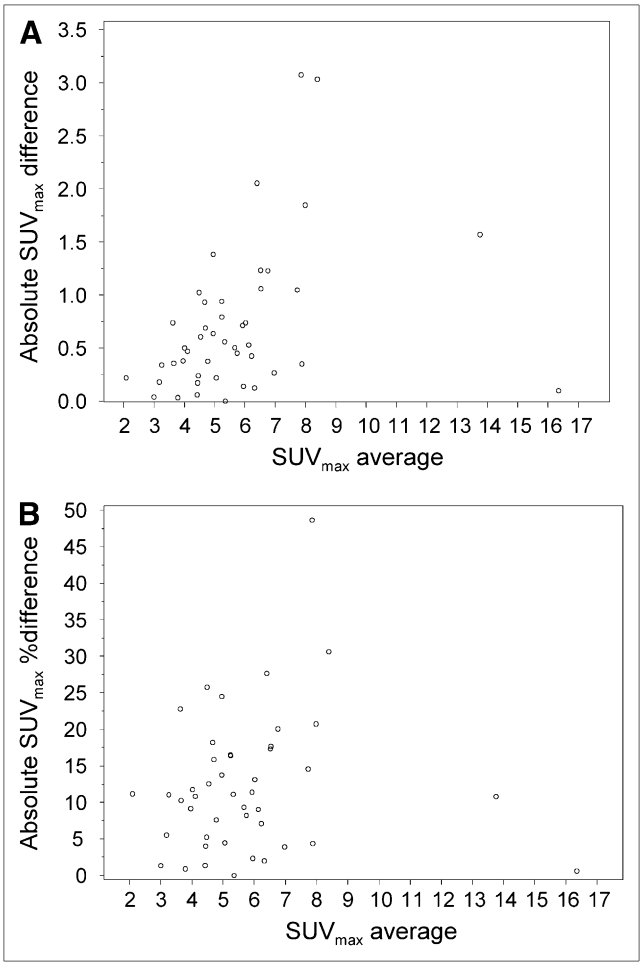

FIGURE 3. Scatter plot of absolute baseline SUV_{max} differences vs. average SUV_{max} based on QA dataset.

changes allowing for interpretability with published results (*6,8*).

SUV percentage change, rather than absolute change, was used to assess repeatability, as this is appropriate in settings in which SUV differences increase with SUVs (Fig. 3) and was used broadly for assessing response (*13*). Clinical applications in which absolute SUV is used, that is, assessing residual SUV during or after treatment or when SUV is used as prognostic factor (*22*), and studies that have addressed assessment of an absolute SUV floor (*23*) are reported. Optimal measurements to assess response may depend on the tumors in combination with therapies being investigated or a combination of assessments, such as a defined relative change along with an absolute SUV change, as suggested by Wahl et al. (*24*).

The results of this study demonstrate variability to be somewhat larger for the non-QA multiobserver analysis (15.9%) than what was seen in single-center studies (10%−12%) (*5,6*), though still within a reasonable range, as single-center test–retest variability ranges from 6% to 10% to up to 42% (*6,8,9,25*). Performing centralized QA to assess protocol compliance resulted in variability (10.7%−12.8%).

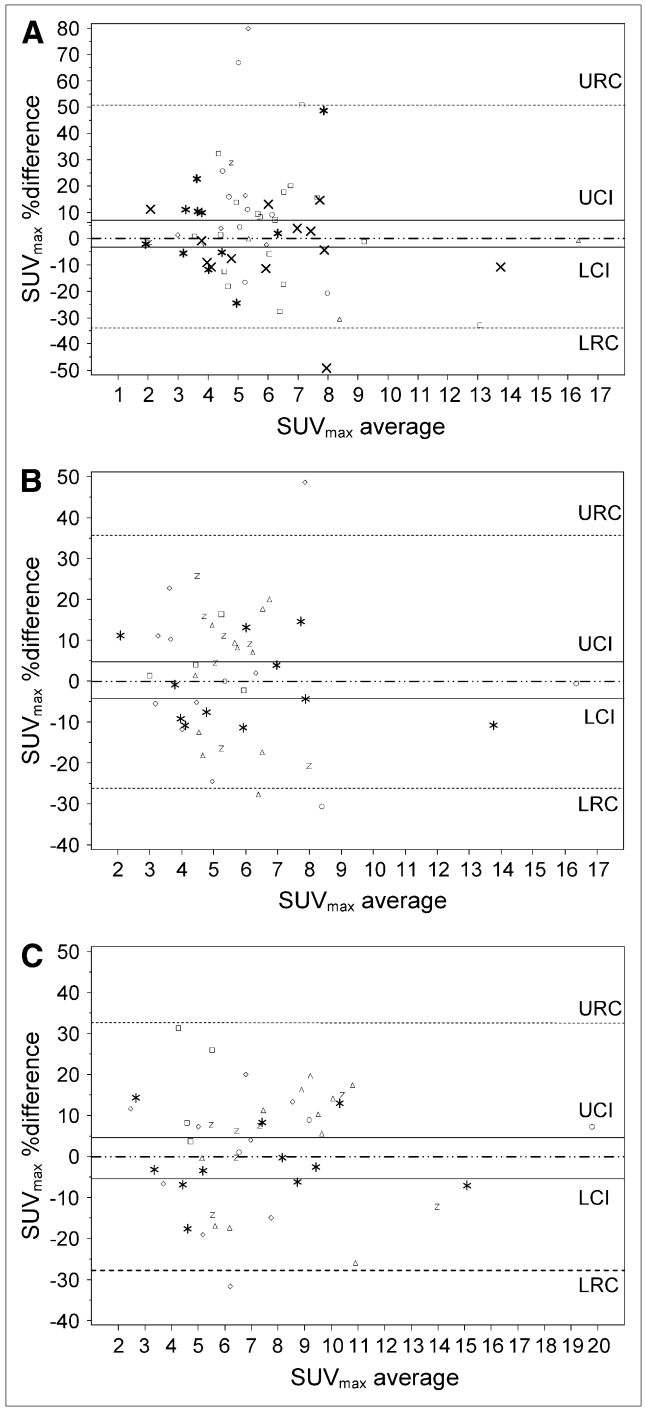

FIGURE 4. Bland–Altman Plots of ^{18}F-FDG PET SUV_{max} using average across lesions for full (n = 61) (A), QA multiobserver (n = 45) (B), and QA single-observer (n = 45) (C) datasets. Horizontal lines denote no-change line, 95% CIs for mean differences (LCI, UCI), and 95% RCs (LRC, URC), both expressed as percentages. Site symbols (A): 1 = z; 2 = ○; 3 = Δ; 4 = □; 5 = ◇; 6 = ⋆; 7 = x; and 8 = Y. Site symbols (B and C): 2 = z; 3 = ○; 4 = Δ; 5 = □; 6 = ◇; and 7 = ⋆.

True response versus statistical fluctuation can be delineated, and standardized criteria for response assessment can be defined on the basis of test–retest repeatability and an

accurate ROI definition and the SUV parameter in carefully selected lesions. Current European Organization for Research and Treatment of Cancer guidelines (*13*) for ^{18}F-FDG PET response assessment delineate progressors and responders based on a ±25% deviation from baseline values. On the basis of the repeatability results of this study, the threshold for determining metabolic response may be on the order of up to −34% in a multicenter multiobserver non-QA setting and up to −25% to −27% in a multicenter centralized QA setting, allowing for increased confidence that a true change from baseline has occurred. In addition, these RCs show that increases in the ranges of 40%−50% in SUV from baseline values after treatment (39% for QA datasets to 52% for non-QA datasets) may be indicative of lack of treatment effect and therefore be deemed progression from baseline (Table 5).

CONCLUSION

Repeatability, defined as the variation of repeated measurements in an experiment performed under the same conditions, was similar for the studied SUV parameters (SUV_{mean}, SUV_{max}, SUV_{peak}, and $SUV_{70\%}$) assessed with double-baseline ^{18}F-FDG PET scans in a variety of analysis settings for this multicenter study. Although the variability in the absence of close compliance with consensus recommendations was comparable with reported single-center test–retest studies, centralized QA of data improved intrasubject CV from 15.9% to 10.7% for SUV_{max} and the threshold for determining metabolic changes from approximately −34% to −27% and from +52% to +37%, respectively. This study supports the use of ^{18}F-FDG PET as a tumor-assessment tool in multicenter oncology clinical trials, provided a centralized QA assessment of the data is performed.

ACKNOWLEDGMENTS

We acknowledge the patients who participated in this study and their families. We also thank the clinical and imaging teams at the 8 study sites included in this study: John Marshall and David Earl-Graef, Georgetown University Hospital, Washington, DC; Pierre Major and Carol Dunne, Juravinski Cancer Centre, Hamilton, Ontario; Patricia LoRusso and Anthony Shields, Karmanos Cancer Center, Detroit, Michigan; Chris Garrett and Claudia Berman, H. Lee Moffitt Cancer Center, Tampa, Florida; Lillian Siu and Dave Wilson, Princess Margaret Hospital, Toronto, Ontario; Caio Rocha-Lima and Hilton Gomes/Sylvester Comprehensive Cancer Center, Miami, Florida; Anthony el-Khoueiry and Peter Conti, USC/Norris Comprehensive Cancer Center, Los Angeles, California; and Jan Buter and Otto Hoekstra, VU University Medical Center, Amsterdam, The Netherlands. A special thanks to Nikie Hoetjes and Reina Kloet for their continued efforts. Consultation and central QA and data analysis performed by VU University Medical Center was sponsored by Bristol-Myers Squibb Co. Linda Velasquez, Georgia Kollia, Wendy Hayes, and Susan Galbraith are employees of Bristol-Myers Squibb Co.

REFERENCES

1. Juweid ME, Cheson BD. Positron-emission tomography and assessment of cancer therapy. *N Engl J Med.* 2006;354:496–507.
2. Kelloff GJ, Hoffman JM, Johnson B, et al. Progress and promise of FDG-PET imaging for cancer patient management and oncologic drug development. *Clin Cancer.* 2005;11:2785–2808.
3. Nakamoto Y, Chang AE, Zasadny KR, Wahl RL. Comparison of attenuation-corrected and non-corrected FDG-PET images for axillary nodal staging in newly diagnosed breast cancer. *Mol Imaging Biol.* 2002;4:161–169.
4. Halligan S. Reproducibility, repeatability, correlation and measurement error. *Br J Radiol.* 2002;75:193–195.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. Shankar LK, Hoffman JM, Bacharch S, et al. Consensus recommendations for the use of ^{18}F-FDG-PET as an indicator of therapeutic response in patients in National Cancer Institute Trials. *J Nucl Med.* 2006;47:1059–1066.
12. Weber WA. Use of PET in monitoring cancer therapy and for predicting outcome. *J Nucl Med.* 2005;46:983–995.
13. 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.
14. Hoekstra CJ, Paglianiti I, Hoekstra OS, et al. Monitoring response to therapy in cancer using [^{18}F]-2-fluoro-2-deoxy-D-glucose and positron emission tomography: an overview of different analytical methods. *Eur J Nucl Med.* 2000;27:731–743.
15. 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.
16. 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.
17. Hume R. Prediction of lean body mass from height and weight. *J Clin Pathol.* 1966;19:389–391.
18. Thie JA, Hubner KF, Smith GT. The diagnostic utility of the lognormal behavior of PET standardized uptake values in tumors. *J Nucl Med.* 2000;41:1664–1672.
19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–310.
20. Lindholm P, Minn H, Leskinen-Kallio S, Bergman J, Routsalainen U, Joensuu H. Influence of the blood glucose concentration on FDG uptake in cancer: a PET study. *J Nucl Med.* 1993;34:1–6.
21. Lowe VJ, DeLong DM, Hoffman JM, Coleman RE. Optimum scanning protocol for FDG-PET evaluation of pulmonary malignancy. *J Nucl Med.* 1995;36:883–887.
22. Hoekstra CJ, Stroobants SG, Smit EF, et al. Prognostic relevance of response evaluation using [^{18}F]-2-fluoro-2-deoxy-D-glucose positron emission tomography in patients with locally advanced non-small-cell lung cancer. *J Clin Oncol.* 2005;23:8362–8370.
23. Lin C, Itti E, Haioun C, et al. Early ^{18}F-FDG PET for prediction of prognosis in patients with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis. *J Nucl Med.* 2007;48:1626–1632.
24. 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):122S–150S.
25. Krak NC, van der Hoeven JJ, Hoekstra OS, Twisk JW, van der Wall E, Lammertsma AA. Measuring [^{18}F]FDG uptake in breast cancer during chemotherapy: comparison of analytical methods. *Eur J Nucl Med Mol Imaging.* 2003;30:674–681.