Repeatability of ^{18}F-FDG PET/CT in Advanced Non–Small Cell Lung Cancer: Prospective Assessment in 2 Multicenter Trials

Wolfgang A. Weber[1], Constantine A. Gatsonis[2], P. David Mozley[3], Lucy G. Hanna[2], Anthony F. Shields[4], Denise R. Aberle[5], Ramaswamy Govindan[6], Drew A. Torigian[7], Joel S. Karp[7], Jian Q. (Michael) Yu[8], Rathan M. Subramaniam[9], Robert A. Halvorsen[10], and Barry A. Siegel[11] for the ACRIN 6678 and MK-0646-008 Research teams

[1]Memorial Sloan Kettering Cancer Center, New York, New York; [2]Department of Biostatistics and Center for Statistical Sciences, Brown University, Providence, Rhode Island; [3]Merck & Co., Whitehouse Station, New Jersey; [4]Karmanos Cancer Institute, Wayne State University, Detroit, Michigan; [5]University of California-Los Angeles, Los Angeles, California; [6]Division of Oncology and the Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri; [7]Department of Radiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania; [8]Fox Chase Cancer Center, Philadelphia, Pennsylvania; [9]Russell H. Morgan Department of Radiology and Radiological Science and Sidney Kimmel Cancer Center, Johns Hopkins University, Baltimore, Maryland; [10]Virginia Commonwealth University, Richmond, Virginia; and [11]Mallinckrodt Institute of Radiology and the Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri

PET/CT with the glucose analog ^{18}F-FDG has several potential applications for monitoring tumor response to therapy in patients with non–small cell lung cancer (NSCLC). A prerequisite for many of these applications is detailed knowledge of the repeatability of quantitative parameters derived from ^{18}F-FDG PET/CT studies. **Methods:** The repeatability of the ^{18}F-FDG signal was evaluated in 2 prospective multicenter trials. Patients with advanced NSCLC (tumor stage III–IV) underwent two ^{18}F-FDG PET/CT studies while not receiving therapy. Tumor ^{18}F-FDG uptake was quantified by measurement of the maximum standardized uptake value within a lesion (SUV_{max}) and the average SUV within a small volume of interest around the site of maximum uptake (SUV_{peak}). Analysis was performed for the lesion in the chest with the highest ^{18}F-FDG uptake and a size of at least 2 cm (target lesion) as well as for up to 6 additional lesions per patient. Repeatability was assessed by Bland–Altman plots and calculation of 95% repeatability coefficients (RCs) of the log-transformed SUV differences. **Results:** Test–retest repeatability was assessed in 74 patients (34 from the ACRIN 6678 trial and 40 from the Merck MK-0646-008 trial). SUV_{peak} was 11.57 ± 7.89 g/mL for the ACRIN trial and 6.89 ± 3.02 for the Merck trial. The lower and upper RCs were −28% (95% confidence interval [CI], −35% to −23%) and +39% (95% CI, 31% to 54%) in the ACRIN trial, indicating that a decrease of SUV_{peak} by more than 28% or an increase by more than 39% has a probability of less than 2.5%. The corresponding RCs from the Merck trial were −35% (95% CI, −42% to −29%) and +53% (95% CI, 41% to 72%). Repeatability was similar for SUV_{max} of the target lesion, averaged SUV_{max}, and averaged SUV_{peak} of up to 6 lesions per patient. **Conclusion:** The variability of repeated measurements of tumor ^{18}F-FDG uptake in patients with NSCLC is somewhat larger than previously reported in smaller single-center studies but comparable to that of gastrointestinal malignancies in a previous multicenter trial. The variability of measurements supports the definitions of tumor response according to PET Response Criteria in Solid Tumors.

Key Words: FDG PET/CT; quantification; repeatability; reproducibility

J Nucl Med 2015; 56:1137–1143
DOI: 10.2967/jnumed.114.147728

Lung cancer continues to be the leading cause of cancer deaths in the United States: more patients die of lung cancer than of breast cancer, prostate cancer, colorectal cancer, and lymphoma combined (*1*). Most patients present with advanced disease and undergo palliative chemotherapy. However, only about one third of patients respond to chemotherapy (*2*). Novel targeted therapies directed at the epidermal growth factor receptor do have higher tumor response rates but only in small subgroups of patients with activating mutations of the epidermal growth factor receptor kinase domain (*3*).

PET with the glucose analog ^{18}F-FDG has shown encouraging results for monitoring tumor response to treatment (*4*). Quantitative changes in tumor ^{18}F-FDG uptake a few weeks after the start of therapy have been shown to correlate well with subsequent tumor shrinkage and patient survival (*4*). Thus, ^{18}F-FDG PET has the potential to improve patient management by signaling the need for early therapeutic changes in nonresponders, thereby avoiding the side effects and costs of ineffective treatment. Analogously, early response biomarkers such as ^{18}F-FDG PET could also accelerate oncologic drug development by decreasing the length of time on trial per subject and reducing the number of subjects required to demonstrate a statistically significant difference between the arms of a randomized phase II trial (*5*).

Clinical use of ^{18}F-FDG PET as a biomarker for tumor response to therapy requires a high degree of test–retest reproducibility (repeatability). Six single-center studies have evaluated the test–retest repeatability of quantitative parameters derived from ^{18}F-FDG PET (*6–12*). The coefficient of variation for changes in tumor

Received Aug. 24, 2014; revision accepted Mar. 26, 2015.

For correspondence or reprints contact: Wolfgang A. Weber, Molecular Imaging and Therapy Service, Department of Radiology, Memorial Sloan Kettering Cancer Center, 1275 York Ave., New York, NY 10065.

E-mail: weberw@mskcc.org

Published online Apr. 23, 2015.

TABLE 1
Physiologic and Imaging Parameters for Patients in ACRIN and Merck Trials

	ACRIN (n = 34)						Merck (n = 40)					
	Scan 1			Scan 2			Scan 1			Scan 2		
Parameter	Mean ± SD	Median	Range	Mean ± SD	Median	Range	Mean ± SD	Median	Range	Mean ± SD	Median	Range
Age (y)	59.5 ± 10.0	59.0	40.0–83.0				59.6 ± 11.2	61.0	29.0–83.0			
Weight (kg)	75.5 ± 17.6	74.7	45.5–112.7				57.9 ± 10.5	58.0	32.0–87.0			
Glucose (mg/dL)	103.8 ± 21.7	98.0	74.0–174.0	105.8 ± 20.0	102.0	73.0–151.0	98.4 ± 15.9	95.0	76.0–149.0	97.4 ± 15.1	97.0	71.0–141.0
Activity (MBq)	500 ± 78 (13.5 ± 2.1)	514 (13.9)	333–714 (9.0–19.3)	503 ± 70 (13.6 ± 1.9)	533 (14.4)	344–611 (9.3–16.5)	359 ± 70 (9.7 ± 1.9)	366 (9.9)	215–581 (5.8–15.7)	359 ± 74 (9.7 ± 2.0)	352 (9.5)	215–588 (5.8–15.9)
Uptake time (min)	61.9 ± 9.3	60.0	52.0–99.0	62.4 ± 8.9	60.0	50.0–98.0	61.0 ± 7.0	60.0	49.0–97.0	62.1 ± 12.3	60.0	51.0–141.0

Data in parentheses are mCi.

^{18}F-FDG uptake was about 10%–15% when patients were scanned twice within 2–3 wk. A larger variability was reported when the baseline and follow-up scans were obtained on different scanners (*13*).

Velasquez et al. have reported the results of a multicenter trial evaluating the repeatability of ^{18}F-FDG PET in patients with metastatic cancers of the gastrointestinal tract (*14*). A dual-center study has also evaluated the repeatability of various quantitative indices derived from ^{18}F-FDG PET studies in patients with ovarian cancer (*11*). However, similar data from multicenter studies are still needed for non–small cell lung cancer (NSCLC). Therefore, repeatability of tumor ^{18}F-FDG uptake was assessed as part of a prospective multicenter trial (ACRIN 6678, NCT00424138) conducted by the American College of Radiology Imaging Network (ACRIN, now part of the Eastern Cooperative Oncology Group [ECOG]-ACRIN Cancer Research Group). In the present analysis, the data from ACRIN 6678 were analyzed together with unpublished data from a clinical trial performed by Merck & Co Inc. (MK-0646-008, NCT00729742) that addressed the same question in a similar patient population. A prespecified objective of both trials was to correlate changes in tumor ^{18}F-FDG uptake during chemotherapy with patient survival. In the Merck trial, characterizing the repeatability of measurement was the primary objective, whereas in the ACRIN trial, this was a secondary objective. Data on the correlation between tumor response to therapy on PET and patient outcomes will be reported separately.

TABLE 2
Summary of PET Quantitative Measures

	PET/CT 1		PET/CT 2		Pairwise difference (D)		Log difference (d)	
Parameter	Average	SD	Average	SD	Average	SD	Average	SD
Liver								
SUV_{mean}-A	2.24	0.46	2.16	0.43	−0.08	0.25	−0.035	0.12
SUV_{mean}-M	2.09	0.32	2.04	0.37	−0.05	0.31	−0.030	0.15
SUV_{mean}-P	2.15	0.39	2.09	0.40	−0.06	0.29	−0.032	0.13
Tumor								
SUV_{max}-A	14.93	10.05	14.46	10.15	−0.47	2.03	−0.038	0.16
SUV_{max}-M	9.06	4.04	9.39	3.28	−0.26	1.69	−0.010	0.18
SUV_{max}-P	11.62	7.81	11.72	7.67	−0.36	1.84	−0.023	0.17
SUV_{peak}-A	11.57	7.89	11.02	7.44	−0.56	1.66	−0.047	0.17
SUV_{peak}-M	6.89	3.02	7.32	2.66	0.00	1.44	0.021	0.22
SUV_{peak}-P	8.93	6.10	9.02	5.68	−0.26	1.56	−0.010	0.20
aSUVmax-A	13.00	9.09	12.52	8.95	−0.48	1.86	−0.048	0.16
aSUVmax-M	7.30	2.93	7.64	2.45	−0.07	1.32	0.010	0.17
aSUVmax-P	9.78	6.95	9.88	6.74	−0.26	1.59	−0.017	0.16
aSUVpeak-A	9.86	7.53	9.37	7.12	−0.49	1.32	−0.044	0.15
aSUVpeak-M	5.34	2.09	5.65	1.83	0.03	1.03	0.026	0.19
aSUVpeak-P	7.31	5.64	7.36	5.31	−0.21	1.19	−0.006	0.17

Results of ACRIN trial are denoted with -A, those from Merck trial with -M, and those from pooled ACRIN and Merck data with -P.

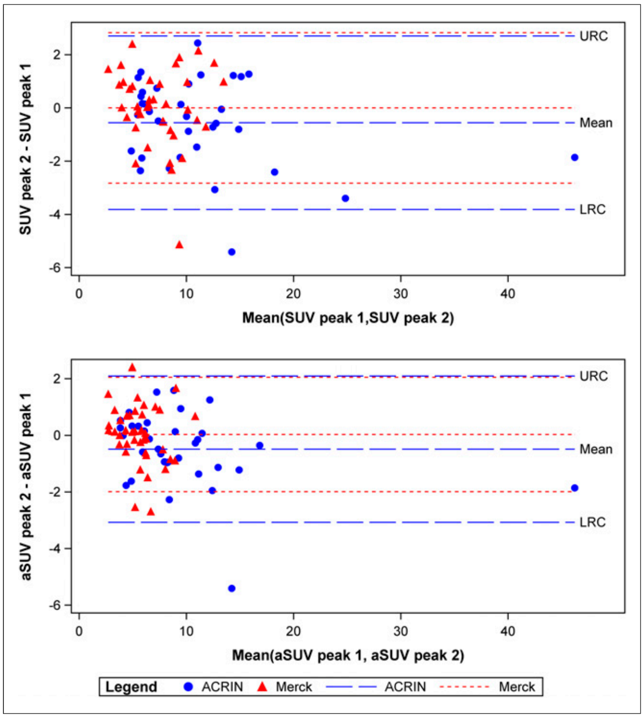

FIGURE 1. Bland–Altman plots showing repeatability of tumor ^{18}F-FDG uptake measured by SUV$_{peak}$ for most active target lesion (top) or SUV$_{peak}$ averaged for several lesions (aSUVpeak, bottom). SUV unit is g/mL. LRC = lower RC; URC = upper RC.

MATERIALS AND METHODS

The Merck and ACRIN trials both included patients aged 18 y or older with locally advanced or metastatic stage III or IV NCSLC (*15*). Other inclusion criteria included a performance status of 0–2 on the ECOG scale and the presence of measurable disease on CT according to the Response Evaluation Criteria in Solid Tumors 1.0 (*16*). For the ACRIN trial, this was either the primary tumor or a metastatic lesion in the chest. Exclusion criteria included previous chemotherapy within 2 wk of study entry, radiotherapy or surgery of the chest within 3 mo before entering the study, pregnancy, breastfeeding, and poorly controlled diabetes mellitus. The ACRIN trial additionally excluded patients with postobstructive pneumonia and patients with pure bronchioloalveolar carcinoma. The institutional review board of each participating site approved the study, and all subjects signed a written informed consent form.

PET/CT Imaging

To participate in the ACRIN trial, sites had to meet all of the criteria described in the ACRIN PET Qualifying Application (www.acrin.org/6678_protocol.aspx). Qualification included tests for correct PET/CT scanner calibration as well as submission of test images to ACRIN. The test images were reviewed by staff of the ACRIN imaging core laboratory for quality control and compliance. Merck used a similar process of site training and qualification that required phantom scans to be submitted to an imaging contract research organization before first-subject enrollment. PET/CT images were acquired in accordance with guidelines of the National Cancer Institute (*17*). The imaging procedure is described in the Supplemental Methods section (supplemental materials are available at http://jnm.snmjournals.org).

Image Analysis

Activity concentrations in the attenuation-corrected PET images were converted to standardized uptake values (SUVs) normalized to patient body weight. For analysis of test–retest repeatability, the intrathoracic lesion with a diameter of more than 2 cm on CT that showed the highest ^{18}F-FDG uptake on the baseline scan was studied (target lesion). A user-defined volume of interest (VOI) was placed around this lesion, and the maximum SUV of target lesion (SUV$_{max}$) within this volume was determined. If the SUV$_{max}$ was less than 4.0 g/mL, the patient was excluded from analysis. This threshold value was based on previous studies that indicated that the repeatability of SUVs (expressed as relative changes from baseline) deteriorates with decreasing tumor ^{18}F-FDG uptake (*12,18*).

A cylindric VOI 1.5 cm in diameter and 3 slices in height was centered on the voxel with maximum ^{18}F-FDG uptake using an automated program (written in MIMVista; MIM Software) (*19*). The SUV$_{max}$ (representing the single voxel with the highest activity concentration) and average SUV (SUV$_{peak}$) within this VOI were determined for further analysis. VOIs were placed in the same way in up to 6 additional lesions. In participants with more than 6 metastatic lesions, a maximum of 3 lesions were analyzed in the same organ. In each organ, the lesions with the highest ^{18}F-FDG uptake were selected for analysis. No minimum SUV or minimum size was required for those additional lesions.

For quality control purposes, a large circular region of interest (ROI) (diameter, ≥5 cm) was placed in normal liver tissue. The average SUV in liver (SUV$_{mean}$) in this ROI was recorded. When it was not feasible to place 1 large ROI in normal liver tissue because of multiple metastases, several small ROIs, comprising approximately the same number of pixels as one 5-cm ROI, were placed in normal liver tissue. ^{18}F-FDG uptake within these ROIs was then averaged for further analysis.

The ACRIN 6678 and the Merck images were analyzed in the ACRIN Imaging Core Laboratory by 1 of 3 nuclear medicine physicians with at least 5 y of experience in assessing PET/CT scans. Both PET/CT studies of individual patients were always analyzed by the same observer.

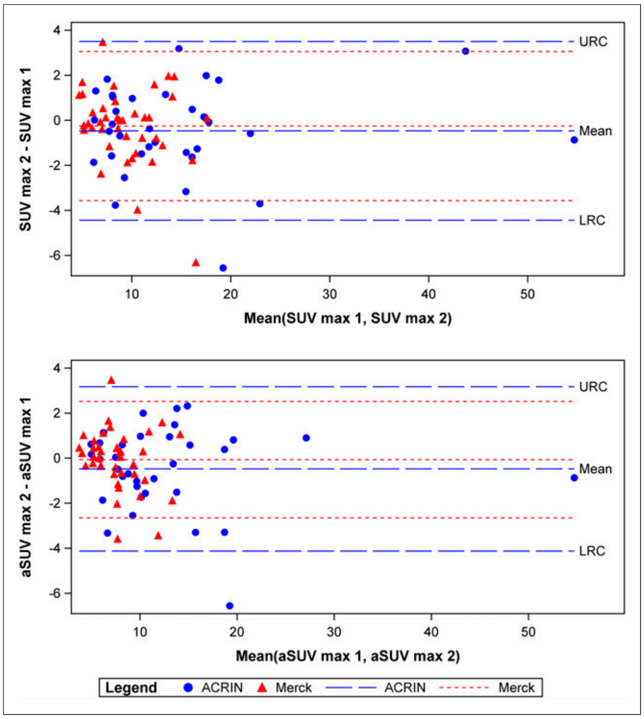

FIGURE 2. Bland–Altman plots showing repeatability of tumor ^{18}F-FDG uptake measured by SUV$_{max}$ (top) or SUV$_{max}$ averaged for several lesions (aSUVmax, bottom). SUV unit is g/mL. LRC = lower RC; URC = upper RC.

TABLE 3
RCs and Their 95% CIs

Parameter	Lower 95% RC	Upper 95% RC	95% CI for lower RC	95% CI for upper RC
SUV_{max}-A	−27%	37%	−34% to −23%	29% to 52%
SUV_{max}-M	−29%	41%	−36% to −25%	32% to 55%
SUV_{max}-P	−28%	39%	−33% to −25%	33% to 48%
SUV_{peak}-A	−28%	39%	−35% to −23%	31% to 54%
SUV_{peak}-M	−35%	53%	−42% to −29%	41% to 72%
SUV_{peak}-P	−32%	47%	−37% to −28%	39% to 59%
aSUVmax-A	−27%	36%	−34% to −22%	28% to 51%
aSUVmax-M	−28%	38%	−34% to −23%	30% to 52%
aSUVmax-P	−27%	38%	−32% to −24%	32% to 47%
aSUVpeak-A	−25%	33%	−31% to −21%	26% to 46%
aSUVpeak-M	−31%	45%	−38% to −26%	36% to 62%
aSUVpeak-P	−29%	41%	−33% to −25%	34% to 50%

Results of ACRIN trial are denoted with -A, those from Merck trial with -M, and those from pooled ACRIN and Merck data with -P.

Statistical Analysis

The data from the 2 studies were first analyzed separately. An analysis of the pooled data was also performed. For each analysis, variability was assessed by calculating the difference of paired SUV_{max} and SUV_{peak} measurements at the time of the 2 PET/CT studies:

$$D_i = u_{i2} - u_{i1}, \qquad \text{Eq. 1}$$

where u_{i1} and u_{i2} are the SUV measurements for a lesion at the time of the baseline and the follow-up scan, respectively. The parameter D was plotted against various parameters with potential influence on the repeatability of the SUV measurements. Then quantile–quantile plots were generated to determine whether the distribution of D deviated from a normal distribution. As this was found to be the case, further analyses were performed on the differences of log-transformed SUV measurements:

$$d_i = \ln(u_{i2}) - \ln(u_{i1}). \qquad \text{Eq. 2}$$

Because

$$\ln(u_{i2}) - \ln(u_{i1}) = \ln\left(\frac{u_{i2}}{u_{i1}}\right), \qquad \text{Eq. 3}$$

analysis of differences of log-transformed data provides information on the repeatability of relative changes in SUVs.

To quantify the test–retest repeatability of SUV measurements, repeatability coefficients (RCs) and their 95% confidence intervals (CIs) were calculated (*20*). This calculation was performed on the log-transformed data using the formula

$$RC_{ln} = 1.96 sd(d), \qquad \text{Eq. 4}$$

with sd being the SD of d. Assuming a normal distribution of d, the probability that measurements of d are larger than $+RC_{ln}$ or smaller than $-RC_{ln}$ is about 5%. To express the repeatability coefficient as a percentage change of SUVs, RC_{ln} was exponentiated using the following formula:

$$RC = (\exp(RC_{ln}) - 1) \times 100\%. \qquad \text{Eq. 5}$$

RC is the repeatability coefficient for the percentage change of SUVs.

The 95% CI of RC was calculated using the χ^2 distribution as previously described (*14*). The repeatability of SUV_{max} and SUV_{peak} was also displayed graphically by Bland–Altman plots of SUV differences on the original and the log scale.

These analyses were performed for SUV_{max} and SUV_{peak}. In addition, SUV_{max} and SUV_{peak} of all measured lesions in an individual patient were averaged, and the repeatability of these parameters ($aSUV_{peak}$, $aSUV_{max}$) was determined in the same way as for the target lesion.

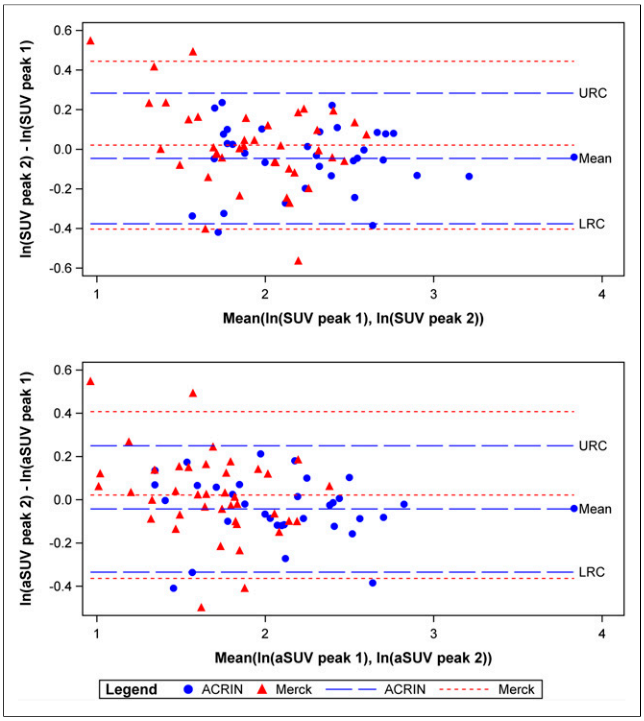

FIGURE 3. Bland–Altman plots showing repeatability of tumor ^{18}F-FDG uptake measured by log-transformed SUV_{peak} (top) or log-transformed SUV_{peak} averaged for several lesions (aSUVpeak, bottom). SUV unit is g/mL. LRC = lower RC; URC = upper RC.

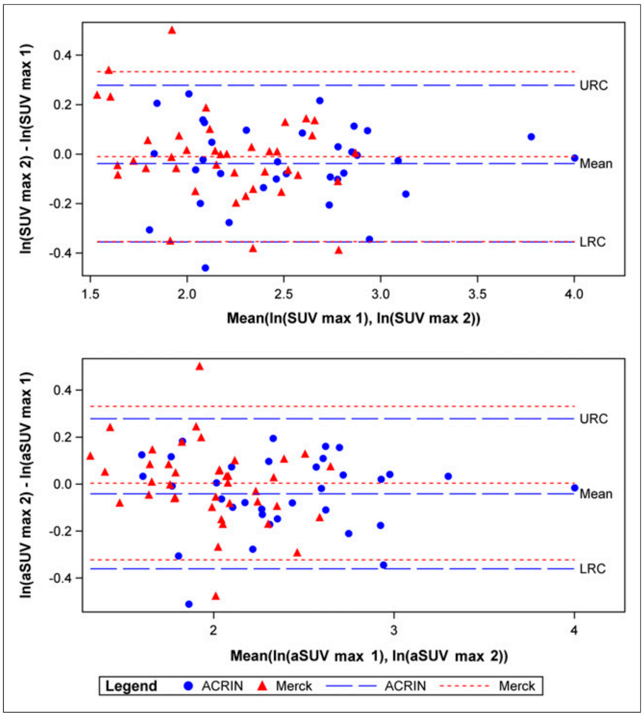

FIGURE 4. Bland–Altman plots showing repeatability of tumor ^{18}F-FDG uptake measured by log-transformed SUV_{max} (top) or log-transformed SUV_{max} averaged for several lesions (aSUVmax, bottom). SUV unit is g/mL. LRC = lower RC; URC = upper RC.

Quantitative parameters are presented as mean ± SD and 95% CIs of the mean, as indicated. The correlation between SUV_{peak} and SUV_{max} was evaluated by Spearman correlation coefficients. Systematic changes in quantitative parameters between the first and second PET scans were analyzed by the Wilcoxon signed-ranked test. Statistical analyses were generated using SAS/STAT software (version 9.3; SAS Institute Inc.).

RESULTS

Ninety-six patients were accrued at 17 sites (Supplemental Table 3) for the ACRIN 6678 trial. Of these, 45 (recruited at 10 sites) consented to participate in the evaluation of test–retest repeatability; evaluable data are available for 34 of these patients (Supplemental Fig. 1).

Merck provided data from 47 patients who were accrued at 14 centers in Europe and Asia from February 2009 to May 2010. Evaluable data are available for 40 of these patients (Supplemental Fig. 2). Table 1 summarizes the body weight, uptake time, blood glucose level, and injected activity for the ACRIN and Merck studies.

To evaluate changes in the whole-body distribution of ^{18}F-FDG between the 2 PET/CT studies, ^{18}F-FDG uptake in the liver was analyzed. As shown in Table 2, liver ^{18}F-FDG uptake remained stable at the time of the 2 PET/CT scans, with low interpatient variability in both the Merck and the ACRIN trials (Table 2).

SUV_{peak} and SUV_{max} parameters for the target lesion showed no systematic increase or decrease from the first to the second PET/CT scan. Overall, SUV_{peak} and SUV_{max} differences demonstrated similar variability as evident from the Bland–Altman plots (Figs. 1 and 2) and the correlation coefficients shown in Supplemental Table 1. Also, averaging SUV_{max} and SUV_{peak} for all lesions in an individual patient to calculate aSUVmax and aSUVpeak had no major relevant effect on the repeatability of the measurements (Figs. 1 and 2; Table 3; Supplemental Table 1). The distribution of the SUV differences in the original scale was similar for the ACRIN and Merck trials. Because lesion SUVs were, on average, more than 1.6 times lower for the Merck than for the ACRIN trial, the difference of log-transformed SUVs was larger for the Merck trial than for the ACRIN trial (Figs. 3 and 4). The higher tumor SUVs of the ACRIN patients may be related to differences in the biodistribution of ^{18}F-FDG because body weight of the ACRIN patient population was 1.3 times higher than that of the Merck patient population (*21*). However, additional factors are likely involved, because tumor SUVs normalized to lean body mass (*21*) were also markedly higher for the ACRIN patients (average SUV 10.6 for the ACRIN population, compared with 6.8 for the Merck population, at the time of the first PET/CT scan).

To identify factors that may explain the variability of SUV measurements, we correlated the differences in SUV_{peak} measurements with various parameters that potentially affect tumor ^{18}F-FDG uptake. Specifically, we analyzed whether body weight, age, clinical stage, blood glucose levels, location of the target lesion, and number of lesions are correlated with the variability of SUV differences. Supplemental Figure 3 indicates that none of these factors had a clear impact on the variability of SUV measurements, although there was a trend of higher variability for the pulmonary lesions. In addition to these patient-related factors, we also analyzed whether differences in uptake time had an impact on the variability of SUV measurements. As shown in Figure 5, no correlation between differences in uptake time and differences in SUV_{peak} measurements was observed. Thus, differences in radiotracer uptake time over the range encountered in our study population had no major influence on the observed variability of SUVs in this study.

The log-transformed SUVs were used to define the 95% RC for the various studied parameters (Table 3). Supplemental Table 2 shows the corresponding coefficients of variation for comparison with prior studies. Overall, all parameters demonstrated similar RCs, with widely overlapping CIs. Figure 6 shows that, in most of the patients, SUVs of the target lesion and additional lesions changed in the same direction, which explains why the analysis of multiple lesions only slightly reduced the variability of the measurements.

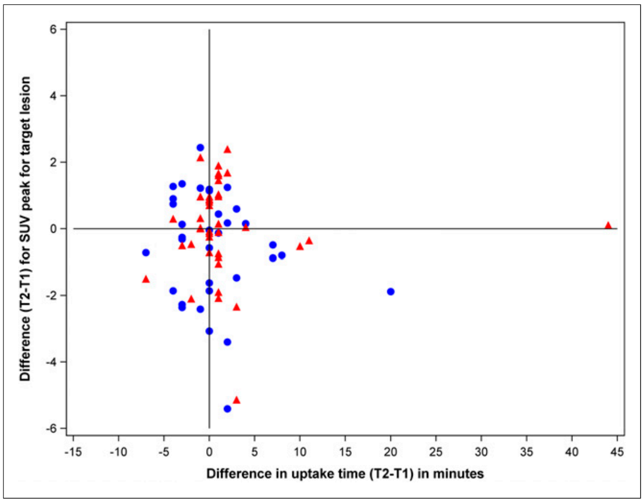

FIGURE 5. Correlation between differences in SUV_{peak} and differences in uptake time for target lesions. SUV unit is g/mL.

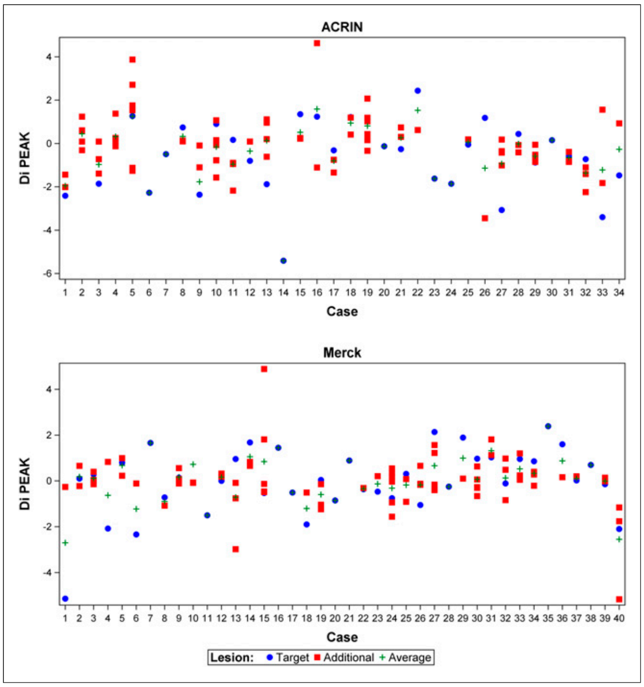

FIGURE 6. Scatterplot showing differences between SUV_{peak} for target lesions (red) and additional lesions (blue) by patient ID. In many cases, SUV differences are in same direction and of similar magnitude for target lesion and additional lesions.

DISCUSSION

The variability of SUVs observed in this study is slightly higher than in previous single-center studies of patients with lung cancer and other malignancies but similar to the results of a previous multicenter study in patients with gastrointestinal malignancies (*14*). There are several potential reasons why the repeatability in NSCLC may be better or worse than in other tumors. On the one hand, the high metabolic activity of most NSCLCs may facilitate quantitative measurements (*18*). On the other hand, respiratory movement may cause errors in quantitative parameters because the PET signal is averaged over several breathing cycles and misregistration of the liver on PET and CT scans can result in considerable underestimation of tumor ^{18}F-FDG uptake (*22*). Respiratory gating can minimize these measurement errors, but because it is not yet widely used clinically, it was not applied in this study (*22*).

To understand the variability of SUV measurements in our patient population, we investigated several clinical factors that may affect the measurements (*23*). We observed only a trend of higher SUV measurement variability in the lungs (Supplemental Fig. 3), which may be due to respiratory movement.

There is a longstanding discussion on how tumor ^{18}F-FDG uptake should be measured on PET/CT studies. Phantom studies have indicated that SUV_{max} measurements are more influenced by variations in acquisition and reconstruction protocols than measurements averaging the information of several voxels (*24,25*), although the differences are mitigated by the noise correlations introduced during image reconstruction (*26*). On the other hand, measurement of mean tumor ^{18}F-FDG uptake requires defining tumor borders, which may introduce interobserver variability. SUV_{peak} measurements represent a compromise between SUV_{max} and SUV_{mean} measurements and have been used in several studies to assess tumor response to therapy (*19*). In the present study, we did not observe differences in the test–retest variability of SUV_{max} and SUV_{peak} measurements. Similarly, a recent study in patients with ovarian cancer (*11*) has reported almost identical repeatability for SUV_{max} and SUV_{mean} measurements, suggesting that for the typical count statistics of whole-body ^{18}F-FDG PET studies, the repeatability of SUV_{max} and SUV_{mean} measurements are not fundamentally different.

A related question is whether to use measurements of tumor ^{18}F-FDG uptake for a single lesion or to average SUV_{max} or $\text{SUV}_{\text{mean}}/\text{SUV}_{\text{peak}}$ measurements for multiple lesions. Most previous studies on the test–retest variability of tumor ^{18}F-FDG uptake have used data for a single lesion (*7,9,12*). In the present study, averaging SUV_{peak} or SUV_{max} measurements for several lesions in an individual patient (aSUVmax and aSUVpeak) had no major impact on the test–retest repeatability of these measurements, because in many patients all lesions changed in a similar way between the first and second PET/CT scan (Fig. 6). Nevertheless, parameters that average information from several lesions may correlate more strongly with patient outcome (*27*). Therefore, further studies on treatment monitoring with ^{18}F-FDG PET should systematically compare measurements of single and multiple lesions.

The test–retest repeatability of SUVs can be analyzed on the basis of either SUV differences or SUV ratios. We focused on SUV ratios because tumor response has generally been defined as a percentage change in pretreatment SUVs and calculated 95% RCs for percentage changes in SUVs. RCs allow an objective definition of criteria for tumor response or progression: if tumor ^{18}F-FDG uptake after therapy decreases by more than the RC, there is a less than 2.5% probability that this decrease is due to variability of the measurement process. Such a change in ^{18}F-FDG uptake most likely indicates an effect of therapy. Therefore, our findings indicate that a decrease in ^{18}F-FDG uptake by 30% likely reflects a metabolic response to therapy. A larger relative increase in ^{18}F-FDG uptake is needed for confidence that the change represents metabolic progression (Table 3). The asymmetry of the RCs is a consequence of the log transformation of the original measurements and exponentiation in Equation 5 (*14*). At first glance, the asymmetry may seem counterintuitive, but it is the appropriate way to express the repeatability of relative SUV changes, as seen in the following example: a decrease in SUV from 5 to 4 represents a relative change of 20%, whereas an increase in SUV from 4 to 5 represents a 25% relative change, although both pairs of SUV measurement have the same variability. Therefore, symmetric RCs are not suitable for changes of a parameter relative to a baseline measurement.

CONCLUSION

Both trials of patients with advanced NSCLC suggest that for lesions greater than 2 cm in size and with SUV_{max} greater than 4.0, decreases in tumor ^{18}F-FDG uptake by more than 30% and increases by more than 40% are unlikely to reflect variability of the measurement process and could therefore be used to define metabolic response and metabolic progression, respectively. Thus, our data support the recently published PET Response Criteria in Solid Tumors for assessing tumor response on PET (*19*).

DISCLOSURE

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. This work was conducted by The American College of Radiology Imaging Network (now the ECOG-ACRIN Cancer Research Group) and supported by funding from the Foundation for the National Institutes of Health and by National Cancer Institute grants U01-CA-80098 and U01-CA-79778, under a cooperative agreement with the Cancer Imaging Program, Division of Cancer Treatment and Diagnosis. Merck & Co Inc. has no financial investments in medical imaging. No other potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We gratefully acknowledge the participating institutions and the principal investigators at each site (Supplemental Table 3). We also thank the many radiologists, radiation and medical oncologists, nuclear medicine technologists, and research coordinators at the participating institutions, the American College of Radiology Imaging Network (ACRIN) staff who supported ACRIN 6678, and the Biostatistical and Data Management Center at Brown University in Providence, Rhode Island (Supplemental Table 4). Without the diligent efforts of all of these individuals, this study would not have been possible.

REFERENCES

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. *CA Cancer J Clin.* 2014;64:9–29.
2. Schiller JH, Harrington D, Belani CP, et al. Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *N Engl J Med.* 2002;346:92–98.
3. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med.* 2009;361:947–957.
4. Skoura E, Datseris IE, Platis I, Oikonomopoulos G, Syrigos KN. Role of positron emission tomography in the early prediction of response to chemotherapy in patients with non–small-cell lung cancer. *Clin Lung Cancer.* 2012;13:181–187.
5. Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. *Nat Rev Drug Discov.* 2012;11:201–214.
6. Hatt M, Cheze-Le Rest C, Aboagye EO, et al. Reproducibility of ^{18}F-FDG and 3′-deoxy-3′-^{18}F-fluorothymidine PET tumor volume measurements. *J Nucl Med.* 2010;51:1368–1376.
7. 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.
8. Kumar V, Nath K, Berman CG, et al. Variance of SUVs for FDG-PET/CT is greater in clinical practice than under ideal study settings. *Clin Nucl Med.* 2013;38:175–182.
9. 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.
10. 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.
11. Rockall AG, Avril N, Lam R, et al. Repeatability of quantitative FDG-PET/CT and contrast-enhanced CT in recurrent ovarian carcinoma: test-retest measurements for tumor FDG uptake, diameter, and volume. *Clin Cancer Res.* 2014;20:2751–2760.
12. 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.
13. 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.
14. 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.
15. Greene F, Balch C, Haller D, et al., ed. American Joint Committee on Cancer. *AJCC Cancer Staging Manual.* 6th ed. Philadelphia, PA: LippincottRaven; 2002.
16. 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.
17. Shankar LK, Hoffman JM, Bacharach 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.
18. de Langen AJ, Vincent A, Velasquez LM, et al. Repeatability of ^{18}F-FDG uptake measurements in tumors: a metaanalysis. *J Nucl Med.* 2012;53:701–708.
19. 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.
20. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135–160.
21. Tahari AK, Chien D, Azadi JR, Wahl RL. Optimum lean body formulation for correction of standardized uptake value in PET imaging. *J Nucl Med.* 2014;55:1481–1484.
22. Nehmeh SA, Erdi YE, Ling CC, et al. Effect of respiratory gating on quantifying PET images of lung cancer. *J Nucl Med.* 2002;43:876–881.
23. Boellaard R, O'Doherty MJ, Weber WA, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging.* 2010;37:181–200.
24. 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.
25. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. *J Nucl Med.* 2012;53:1041–1047.
26. Doot RK, Scheuermann JS, Christian PE, Karp JS, Kinahan PE. Instrumentation factors affecting variance and bias of quantifying tracer uptake with PET/CT. *Med Phys.* 2010;37:6035–6046.
27. Reck M, Heigener DF, Mok T, Soria JC, Rabe KF. Management of non-small-cell lung cancer: recent developments. *Lancet.* 2013;382:709–719.