Repeatability of Quantitative ^{18}F-NaF PET: A Multicenter Study

Christie Lin[1], Tyler Bradshaw[2], Timothy Perk[1], Stephanie Harmon[1], Jens Eickhoff[3], Ngoneh Jallow[4], Peter L. Choyke[5], William L. Dahut[6], Steven Larson[7], John Laurence Humm[7], Scott Perlman[2,8], Andrea B. Apolo[6], Michael J. Morris[9], Glenn Liu[1,8], and Robert Jeraj[1,8]

[1]Department of Medical Physics, University of Wisconsin, Madison, Wisconsin; [2]Department of Radiology, University of Wisconsin, Madison, Wisconsin; [3]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin; [4]Department of Radiology and Imaging Sciences, Emory University, Atlanta, Georgia; [5]Molecular Imaging Program, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland; [6]Medical Oncology Branch, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland; [7]Molecular Imaging and Therapy Service, Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, New York; [8]University of Wisconsin Carbone Cancer Center, Madison, Wisconsin; and [9]Department of Medical Oncology, Memorial Sloan Kettering Cancer Center, New York, New York

^{18}F-NaF, a PET radiotracer of bone turnover, has shown potential as an imaging biomarker for assessing the response of bone metastases to therapy. This study aimed to evaluate the repeatability of ^{18}F-NaF PET–derived SUV imaging metrics in individual bone lesions from patients in a multicenter study. **Methods:** Thirty-five castration-resistant prostate cancer patients with multiple metastases underwent 2 whole-body (test–retest) ^{18}F-NaF PET/CT scans 3 ± 2 d apart from 1 of 3 imaging sites. A total of 411 bone lesions larger than 1.5 cm^3 were automatically segmented using an SUV threshold of 15 g/mL. Two levels of analysis were performed: lesion-level, in which measures were extracted from individual-lesion regions of interest (ROI), and patient-level, in which all lesions within a patient were grouped into a patient ROI for analysis. Uptake was quantified with SUV_{max}, SUV_{mean}, and SUV_{total}. Test–retest repeatability was assessed using Bland–Altman analysis, intraclass correlation coefficient (ICC), coefficient of variation, critical percentage difference, and repeatability coefficient. The 95% limit of agreement (LOA) of the ratio between test and retest measurements was calculated. **Results:** At the lesion level, the coefficient of variation for SUV_{max}, SUV_{mean}, and SUV_{total} was 14.1%, 6.6%, and 25.5%, respectively. At the patient level, it was slightly smaller: 12.0%, 5.3%, and 18.5%, respectively. ICC was excellent (>0.95) for all SUV metrics. Lesion-level 95% LOA for SUV_{max}, SUV_{mean}, and SUV_{total} was (0.76, 1.32), (0.88, 1.14), and (0.63, 1.71), respectively. Patient-level 95% LOA was slightly narrower, at (0.79, 1.26), (0.89, 1.10), and (0.70, 1.44), respectively. We observed significant differences in the variance and sample mean of lesion-level and patient-level measurements between imaging sites. **Conclusion:** The repeatability of SUV_{max}, SUV_{mean}, and SUV_{total} for ^{18}F-NaF PET/CT was similar between lesion- and patient-level ROIs. We found significant differences in lesion-level and patient-level distributions between sites. These results can be used to establish ^{18}F-NaF PET–based criteria for assessing treatment response at the lesion and patient levels. ^{18}F-NaF PET demonstrates repeatability levels useful for clinically quantifying the response of bone lesions to therapy.

Key Words: sodium fluoride; PET; repeatability; metastatic prostate cancer; multicenter clinical trial

J Nucl Med 2016; 57:1872–1879
DOI: 10.2967/jnumed.116.177295

Prostate cancer is distinct among solid tumors in that its advancement presents largely as clinically detectable osteoblastic bone metastases (*1*). Currently, there are no established tools to reliably and quantitatively measure functional changes in bone metastases in response to therapy (*2*). The development of imaging biomarkers to measure response by bone can improve clinical care, particularly in advanced prostate cancer.

Radiolabeled sodium fluoride, ^{18}F-NaF, was first introduced by Blau et al. in 1972 (*3*) for the detection of bone lesions with PET. However, ^{18}F-NaF was largely replaced by bone scintigraphy using ^{99m}Tc because of superior imaging characteristics with conventional γ-cameras and the readily available supply of ^{99m}Tc (*3–6*). With recent technologic advances in PET, ^{18}F-NaF PET has been increasingly used for detecting bone metastases because of its higher specificity and sensitivity as compared with planar bone scintigraphy and SPECT (*4,5,7–10*). ^{18}F-NaF PET shows potential for longitudinal disease assessment, as its SUV in both normal and pathologic bone is representative of changes in bone metabolism (*11–13*).

To accurately assess tumor response it is necessary to measure a biomarker's repeatability, defined as the variation in measurements when an experiment is repeated under the same conditions (*14*). The repeatability of ^{18}F-FDG PET based on double-baseline studies has been well studied, permitting the development of PERCIST (*15–17*). No such criteria exist for evaluating quantitative ^{18}F-NaF PET response.

A previous study on ^{18}F-NaF PET evaluated the repeatability of bone uptake within the whole body (*18*). However, the repeatability of uptake in individual bone-lesion regions of interest (ROIs) can also be evaluated, allowing assessment of how a

Received May 10, 2016; revision accepted Jun. 18, 2016.
For correspondence or reprints contact: Robert Jeraj, Department of Medical Physics, Wisconsin Institutes for Medical Research, 1111 Highland Ave., Room 7003, Madison, WI 53705.
E-mail: rjeraj@wisc.edu
Published online Jul. 21, 2016.

tumor's response may uniquely contribute to the disease burden on the patient as a whole. The ability to evaluate the repeatability of uptake in an individual lesion would allow for assessment of response heterogeneity within the patient.

Here, we report on the first (to our knowledge) multicenter study assessing the repeatability of ^{18}F-NaF PET uptake at the lesion level. In addition, we compared repeatability between 3 sites in a multicenter trial.

MATERIALS AND METHODS

Patient Population and Study Design

This was a prospective, nonrandomized, 2-arm, multicenter pharmacodynamic-imaging trial with the primary objective of determining the repeatability of ^{18}F-NaF PET/CT imaging for evaluating osseous metastases in patients with metastatic castration-resistant prostate cancer. Eligible patients aged 18 y or older with progressive metastatic castration-resistant histologically proven prostate adenocarcinoma and bone scan–confirmed osseous metastases were enrolled for either docetaxel-based chemotherapy or androgen receptor–directed therapy between February 2012 and September 2014 at the University of Wisconsin Carbone Cancer Center (UWCCC), Memorial Sloan Kettering Cancer Center (MSKCC), or the National Cancer Institute (NCI). The exclusion criteria included active systemic treatment for prostate cancer, palliative radiation within 4 wk of registration, or any prior radioisotope treatment for prostate cancer. The Institutional Review Board and Radiation Safety Committee of each participating institution approved this study, and all subjects signed a written informed consent form. A sample size of 20 patients per site was proposed to evaluate repeatability. This sample size provided sufficient power (≥80%) to detect the anticipated excellent level of repeatability at each of the 3 study sites at the 1-sided 0.0167 significance level.

Quantitative Image Acquisition

Test–retest ^{18}F-NaF PET/CT whole-body scans were to be performed 2–5 d apart and before the start of therapy. Patients were injected intravenously with a bolus of 111–185 MBq (3–5 mCi) of ^{18}F-NaF and imaged 60 min after injection for 3 min per bed position from feet to skull vertex. Scans at UWCCC and MSKCC were acquired on a Discovery VCT PET/CT scanner (GE Healthcare), and scans at NCI were acquired on a Gemini PET/CT scanner (Philips Healthcare). The PET images were corrected for attenuation and scatter.

Scanner Harmonization

The scanners were quantitatively harmonized to obtain equivalent image quality and quantitative accuracy across scanners. The Discovery VCTs were harmonized to the Gemini using a uniform phantom (the National Electrical Manufacturers Association International Electrotechnical Commission body phantom) to measure the signal-to-noise ratio. Absolute calibration was measured by the recovery coefficient, defined as the ratio of the mean measured activity concentration to the true activity concentration in the ROI. Differences in recovery coefficient and signal-to-noise ratio between scanners were minimized by systemically varying the reconstruction parameters, such as number of iterations, number of subsets, and postreconstruction filter.

ROI Definition

Lesions were automatically identified and segmented by applying a CT mask to exclude soft-tissue uptake, followed by application of an SUV threshold of 15 g/mL to exclude additional activity with a low statistical likelihood of being malignant (*18,19*). Lesion contours on PET/CT images were verified by an experienced nuclear medicine physician, and contours smaller than 1.5 cm^3 as measured by PET volume were excluded. Corresponding lesions were automatically matched between paired scans using articulated registration (*20*).

Two levels of SUV analysis were performed: lesion level, in which SUV metrics were extracted from each lesion ROI, and patient level, in which all lesions for a single patient were grouped into a patient ROI before SUV analysis. For both ROI levels, SUV$_{max}$ was defined as the maximum SUV of the ROI and SUV$_{total}$ was defined as the total summed SUV of the ROI normalized to voxel volume. SUV$_{mean}$ was defined as the mean SUV within the lesion ROI or the mean of the SUV$_{mean}$ of all lesions within the patient ROI. The 2 levels of analysis are differentiated here using the terms *lesion SUV* for lesion-level SUV metrics and *patient SUV* for patient-level SUV metrics.

Statistical Analysis

The primary outcome measures for evaluating the repeatability of SUV metrics were intraclass correlation coefficient (ICC) and repeatability coefficient. Repeatability coefficient was calculated at an α-level of 0.05. ICC was estimated using a 2-way mixed-effects model.

We also investigated additional statistical measures for the repeatability of quantitative imaging biomarkers as recommended by the Quantitative Imaging Biomarkers Alliance or as previously reported in the literature (*21*). Test–retest agreement for each ROI was evaluated using Bland–Altman analysis for repeated observations (*22,23*).

Because the distribution of SUV metrics was highly skewed, statistical analyses were performed on natural-log transformations of measurements (*21,22,24*). Statistical analysis was conducted using MATLAB (The MathWorks), version R2014B; R (R Development Core Team), version 3.0; and SPSS (IBM Corp.), version 22.

For lesion-level analysis, ANOVA with repeated measurements was used to account for correlations between multiple lesions within the same patient and to calculate σ, the SD of differences between test and retest measurements (*23*).

The coefficient of variation of within-subject measurements was calculated as the ratio of σ to the grand mean. The critical percentage difference is the minimum percentage change needed to designate a change as significant (*18*), defined as $[\exp(1.96\sqrt{2}\,\sigma) - 1] \times 100\%$.

The 95% limit of agreement (LOA) was calculated for the ratio between test (m_A) and retest (m_B) measurements. Within the 95% LOA lies the ratio of m_B/m_A with a probability of 95%:

$$95\%\ \mathrm{LOA} = \left(\mathrm{e}^{(B - RC)}, \mathrm{e}^{(B + RC)}\right), \qquad \text{Eq. 1}$$

where the bias B is the mean ratio between test and retest measurements. The 95% LOA is reported as the ratio of measurements in original units such that it can be applied to evaluate SUV data in original units (e.g., 95% LOA of (0.80, 1.20) would indicate that with 95% frequency, the ratio m_B/m_A will fall within this interval).

One-way ANOVA with pairwise comparisons and 2-sample t testing were used to assess whether the bias for each SUV metric significantly differed between sites. Two-sample F testing was used to evaluate variability across sites.

RESULTS

In total, we evaluated 411 ^{18}F-NaF–avid bone lesions from 35 patients with metastatic castration-resistant prostate cancer imaged at 1 of the 3 sites (Fig. 1). The patients were injected intravenously with 159.8 ± 9.7 MBq (mean ± SD) of ^{18}F-NaF, and test–retest ^{18}F-NaF PET/CT whole-body scans were performed 63 ± 7 min after injection (3 ± 2 d apart). Dose infiltration near

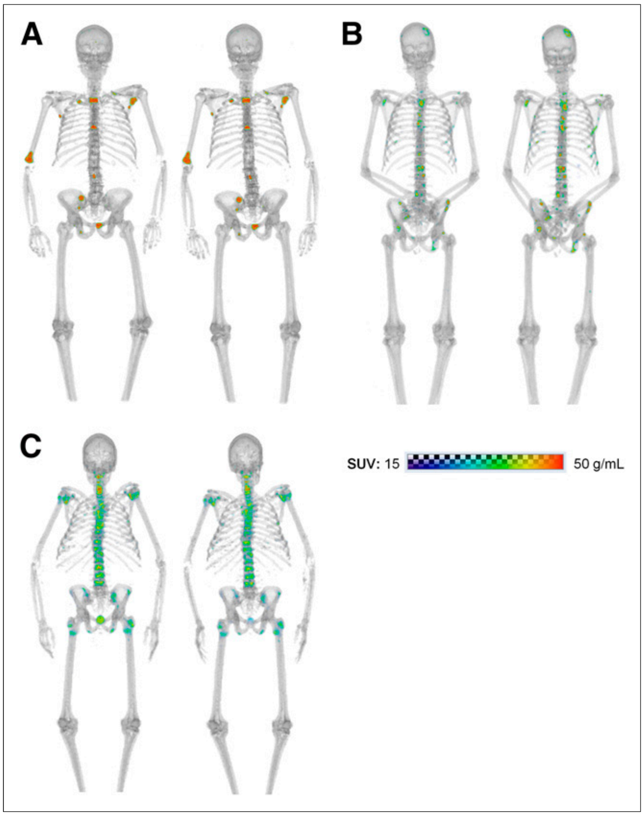

FIGURE 1. Whole-body paired baseline ^{18}F-NaF PET/CT scans of men with metastatic castration-resistant prostate cancer: a 74-y-old imaged 3 d apart at UWCCC (A), a 57-y-old imaged 2 d apart at MSKCC (B), and a 69-y-old imaged 1 d apart at NCI (C).

the injection site was minimal in all scans. Two of the 35 patients underwent partial whole-body scans because the patient was repositioned during the scan. The lesion and patient characteristics are summarized in Table 1. The harmonization reconstruction parameters, including reconstruction method, grid size, subset, iteration, and postreconstruction filter, for each of the scanners are summarized in Table 2.

The median number of lesions per patient at baseline was 8 (range, 1–69). The lesions were located across the skeleton, with the predominant site being the spine. For all lesions, median SUV_{max} was 44.8 (range, 19.6–225.5), SUV_{mean} 23.7 (16.7–75.8), and SUV_{total} 116.7 (26.4–5,628.0) g/mL. For all patients, median SUV_{max} was 86.4 (29.6–225.5), SUV_{mean} 25.4 (18.4–51.1), and SUV_{total} 2,429.3 (47.7–21,447) g/mL.

The relative difference between test and retest scans tended to be slightly greater at the lesion level than at the patient level. For all SUV metrics, relative difference had a narrower distribution for patient ROI than for lesion ROI (Fig. 2). SUV_{mean} had the smallest relative difference for both ROI levels. For lesion ROI, SUV_{mean} was the most repeatable (interquartile range, 2.5%) followed by SUV_{max} (4.4%) and SUV_{total} (5.1%). For patient ROI, SUV_{mean} was the most repeatable (2.0%), followed by SUV_{total} (2.6%) and SUV_{max} (3.3%).

Figure 3 shows Bland–Altman plots for each lesion SUV metric. SUV_{mean} had the smallest variability (repeatability coefficient, 0.13), followed by SUV_{max} (0.27) and SUV_{total} (0.49). Figure 4 shows Bland–Altman plots for each patient SUV metric; again, SUV_{mean} was the most repeatable (0.10), followed by SUV_{max} (0.24) and SUV_{total} (0.36). Both mean and difference values have been log-transformed from SUV (g/mL). Both lesion-level and patient-level distributions had approximately normal distributions and heteroscedasticity.

According to the repeatability coefficient, coefficient of variation, and critical percentage difference, SUV_{mean} was the most repeatable, followed by SUV_{max} and SUV_{total}, at both the lesion level and the patient level (Tables 3 and 4). The 95% LOA defines the interval containing the test-to-retest measurement ratio for each SUV metric. At each site, there was a wide overlap in 95% LOA for all 3 metrics. At the lesion level, the 95% LOA was the narrowest for SUV_{mean} (test-to-retest ratio, 1.00; 95% LOA, (0.88, 1.14)), followed by SUV_{max} (1.00; (0.76, 1.32)) and SUV_{total} (1.04; (0.63, 1.71)). At the patient level, the overall test-to-retest ratio was 0.99 for SUV_{mean} (95% LOA, (0.89, 1.10)), 1.00 for SUV_{max} (0.79, 1.26), and 1.00 for SUV_{total} (0.70, 1.44). Across

TABLE 1
Patient Demographics

Demographic	UWCCC	MSKCC	NCI
Patients (*n*)	18	11	6
Age (y)			
Median	72.5	75.0	68
Range	47–87	57–81	57–83
Height (cm)			
Median	178	177	171
Range	166–191	162–191	161–189
Weight (kg)			
Median	92.3	94.0	84.6
Range	70.7–145.0	73.0–119.0	75.4–91.6
PSA			
Median	71.2	8.1	85.9
Range	1.6–310.0	2.5–246.8	32.0–460.7
Gleason score (*n*)			
6	1 (6%)	2 (18%)	1 (17%)
7	7 (39%)	5 (45%)	2 (33%)
8	4 (22%)	1 (9%)	2 (33%)
9	3 (17%)	3 (27%)	1 (17%)
LDH (U/L)			
Median	200	219	264
Range	139–470	157–251	119–903
Hemoglobin (g/dL)			
Median	12.8	13.8	11.8
Range	7.7–14.9	11.3–15.3	9.0–13.9
Lesions (*n*)			
≤5	6 (33%)	5 (45%)	2 (33%)
6–10	0 (0%)	4 (36%)	1 (17%)
11–20	10 (56%)	2 (18%)	2 (33%)
20	2 (11%)	0 (0%)	1 (17%)

PSA = prostate-specific antigen; LDH = lactic acid dehydrogenase.

SUV metrics, the 95% LOA was consistently narrowest for $\text{SUV}_{\text{mean.}}$ Across sites, the 95% LOA was consistently narrowest, though not significantly different, for UWCCC.

A comparison of overall coefficient of variation and ICC is shown in Figure 5. At both the lesion level and the patient level, ICC was the highest for $\text{SUV}_{\text{total}}$, followed by SUV_{mean} and SUV_{max}. Consistently, patient-level SUV metrics presented a lower coefficient of variation than did lesion-level metrics.

Shown in Figure 6 are Bland–Altman plots of lesion-level SUV_{max} by site. Both mean and difference values have been log-transformed from SUV (g/mL). MSKCC had a sample mean that was statistically significantly different ($P = 0.004$) from the other sites, and UWCCC had a significantly smaller variance ($P < 0.001$). In addition, the variance in SUV_{mean} ($P < 0.001$) and $\text{SUV}_{\text{total}}$ ($P < 0.001$) was significantly smaller at UWCCC than at the other sites.

At the patient level, the sole difference between sites was a significantly smaller variance in $\text{SUV}_{\text{total}}$ at UWCCC ($P = 0.003$) than at the other sites.

TABLE 2
Scanner Harmonization Parameters

Parameter	UWCCC	MSKCC	NCI
Scanner	Discovery VCT	Discovery VCT	Gemini
Reconstruction	3D OSEM	3D OSEM	3D OSEM
Grid size	256 × 256	256 × 256	144 × 144
Subset	14	14	33
Iteration	2	2	2
Postprocessing filter	4 mm	4 mm	—

3D OSEM = 3-dimensional ordered-subsets expectation maximization.

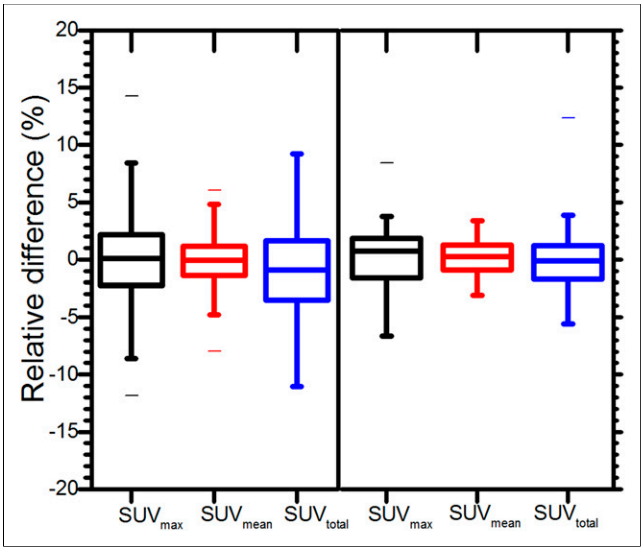

FIGURE 2. Box plots of relative differences in each SUV metric (log-transformed) for lesion-level ROIs (left; 411 lesions) and patient-level ROIs (right; 35 patients). Whiskers extend from minimum to maximum values.

DISCUSSION

To our knowledge, this was the first multicenter study with results demonstrating the repeatability of multiple ^{18}F-NaF PET SUV metrics—SUV_{max}, SUV_{mean}, and $\text{SUV}_{\text{total}}$—for both lesion-level and patient-level ROIs.

Although different guidelines exist for the interpretation of ICC, one of the most common guidelines defines an ICC range of 0.40–0.75 as moderate repeatability and an ICC higher than 0.75 as excellent repeatability (*25*). Although, at the lesion level, the 95% confidence intervals of the ICC for SUV_{max}, SUV_{mean}, and $\text{SUV}_{\text{total}}$ were excellent for all sites, those at the patient level for SUV_{mean} and SUV_{max} at MSKCC and NCI were not fully contained within the region of excellent repeatability. The patient accrual goal was not met because of an imbalance in accrual between the two arms of therapy, thus decreasing the statistical power for evaluating ICC.

In many cases in this study, there were multiple lesions per patient. As shown in the lesion-level Bland–Altman plots of SUV_{max} in Figure 6, multiple lesions within the same patient tended to show correlated repeatability. Thus, it was not possible to regard each lesion as independent. The intrapatient correlations were

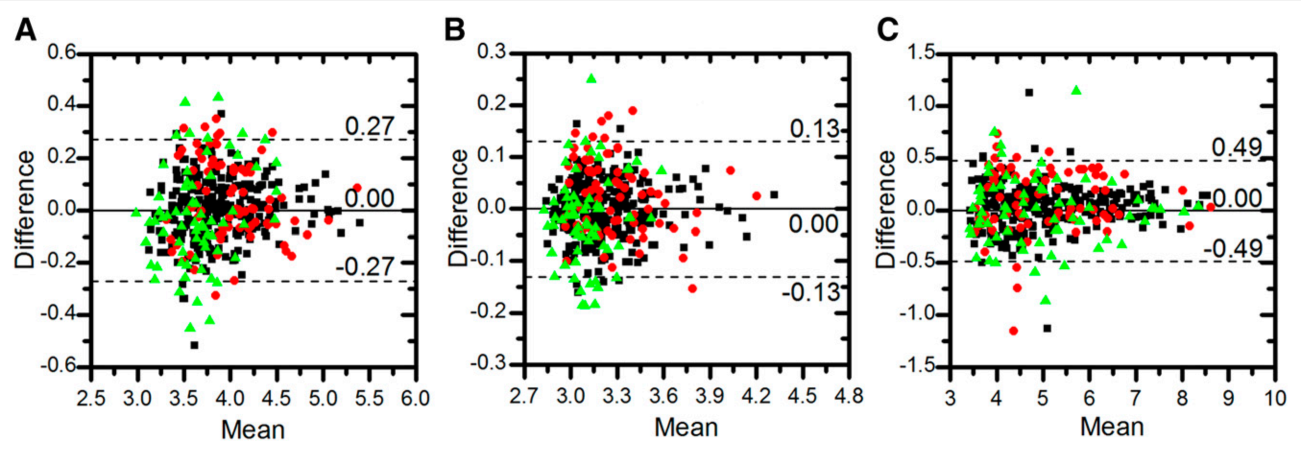

FIGURE 3. Bland–Altman plots of SUV metrics for all lesion-level ROIs (411 lesions): SUV_{max} (A), SUV_{mean} (B), and $\text{SUV}_{\text{total}}$ (C). Different sites are indicated by different symbols (■ = UWCCC, ● = MSKCC, and ▲ = NCI). Solid line denotes mean difference, and dotted lines denote upper and lower 95% LOA. Both mean and difference uptake values have been log-transformed.

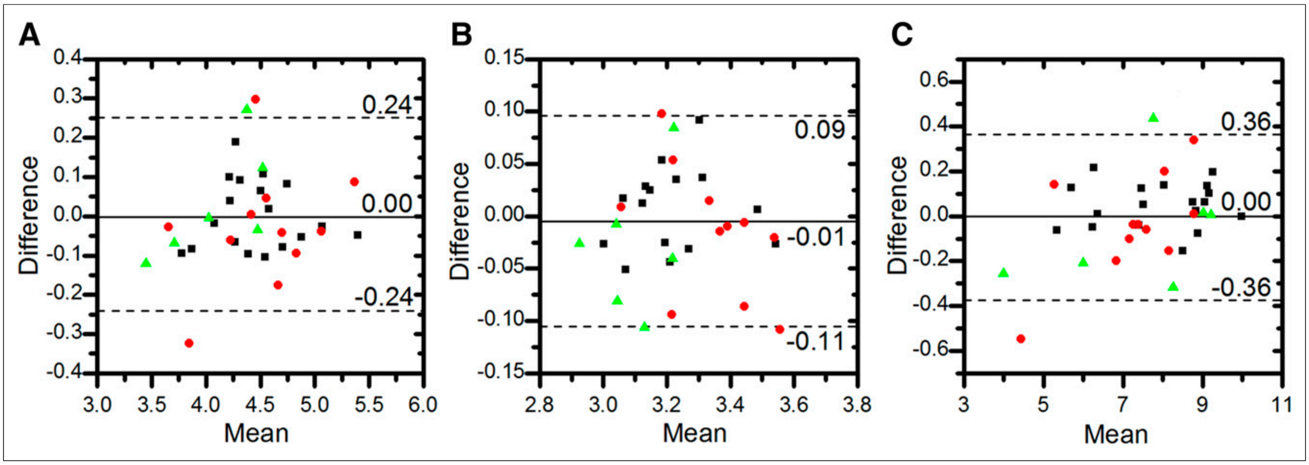

FIGURE 4. Bland–Altman plots of SUV metrics for all patient-level ROIs (35 patients): SUV_{max} (A), SUV_{mean} (B), and SUV_{total} (C). Different sites are indicated by different symbols (■ = UWCCC, ● = MSKCC, and ▲ = NCI). Solid line denotes mean difference, and dotted lines denote upper and lower 95% LOA. Both mean and difference values have been log-transformed.

considered by implementing the Bland–Altman analysis for repeated measures (*23*).

Our repeatability results at the patient level support those of a previous ^{18}F-NaF PET study on bone lesions by Kurdziel et al. (*18*). Despite differences in lesion segmentation methods, our ICC and critical percentage difference findings for SUV_{max}, SUV_{mean}, and SUV_{total} were similar to those of the previous study.

The application of both an uptake threshold and a volume threshold was used to minimize the probability of identifying benign disease. Although Kurdziel et al. used a segmentation SUV threshold of 10 (*18*), a later study by Rohren et al. showed that lesion ROIs identified using this threshold still included normal bone activity (*19*). One study showed that a lesion SUV_{max} of less than 12 g/mL always represented a site of benign disease (*26*). Another

TABLE 3
Repeatability of Lesion ^{18}F-NaF PET SUV Metrics

Metric	RC	ICC*	CV (%)	CPD (%)	B†
UWCCC (265 lesions)					
SUV_{max}	0.23	0.980 (0.974, 0.984)	11.7	37.5	1.00 (0.79, 1.25)
SUV_{mean}	0.10	0.983 (0.979, 0.987)	5.5	15.9	1.00 (0.90, 1.11)
SUV_{total}	0.40	0.990 (0.987, 0.992)	20.7	75.9	1.04 (0.69, 1.56)
MSKCC (78 lesions)					
SUV_{max}	0.31	0.958 (0.935, 0.973)	16.8	54.3	1.04 (0.75, 1.45)
SUV_{mean}	0.14	0.970 (0.953, 0.981)	7.8	22.2	1.03 (0.88, 1.19)
SUV_{total}	0.60	0.990 (0.985, 0.994)	32.7	133.6	1.08 (0.57, 2.06)
NCI (68 lesions)					
SUV_{max}	0.37	0.865 (0.791, 0.915)	20.6	69.2	0.97 (0.65, 1.46)
SUV_{mean}	0.16	0.876 (0.807, 0.922)	9.2	26.2	0.98 (0.82, 1.17)
SUV_{total}	0.65	0.993 (0.989, 0.996)	36.6	151.4	1.00 (0.49, 2.06)
All sites (411 lesions)					
SUV_{max}	0.27	0.969 (0.963, 0.975)	14.1	47.2	1.00 (0.76, 1.32)
SUV_{mean}	0.13	0.975 (0.970, 0.980)	6.6	19.6	1.00 (0.88, 1.14)
SUV_{total}	0.49	0.990 (0.988, 0.992)	25.5	100.4	1.04 (0.63, 1.71)

*Data in parentheses are 95% confidence intervals.

†Data in parentheses are 95% LOA.

RC = repeatability coefficient for α = 0.05 (log-transformed SUV); CV = log-transformed coefficient of variation; CPD = critical percentage difference; B = ratio of test-to-retest bias.

B and 95% LOA have been back-transformed to original units.

TABLE 4
Repeatability of Patient ^{18}F-NaF PET SUV Metrics

Metric	RC	ICC*	CV (%)	CPD (%)	B†
UWCCC (18 patients)					
SUV_{max}	0.17	0.984 (0.959, 0.994)	8.8	27.6	1.00 (0.84, 1.19)
SUV_{mean}	0.08	0.990 (0.974, 0.996)	4.2	12.3	1.01 (0.93, 1.09)
SUV_{total}	0.20	0.993 (0.981, 0.999)	10.1	32.2	1.05 (0.86, 1.28)
MSKCC (11 patients‡)					
SUV_{max}	0.30	0.965 (0.874, 0.990)	15.5	53.8	0.96 (0.71, 1.32)
SUV_{mean}	0.13	0.920 (0.731, 0.978)	6.3	19.0	0.99 (0.87, 1.11)
SUV_{total}	0.45	0.950 (0.825, 0.986)	23.1	89.9	0.96 (0.61, 1.51)
NCI (6 patients)					
SUV_{max}	0.28	0.921 (0.548, 0.989)	14.4	49.2	1.03 (0.77, 1.36)
SUV_{mean}	0.13	0.826 (0.190, 0.974)	6.7	20.2	0.97 (0.85, 1.11)
SUV_{total}	0.54	0.985 (0.895, 0.999)	27.6	115.0	0.95 (0.55, 1.63)
All sites (35 patients)					
SUV_{max}	0.24	0.974 (0.949, 0.987)	12.0	39.5	1.00 (0.79, 1.26)
SUV_{mean}	0.10	0.981 (0.962, 0.990)	5.3	16.0	0.99 (0.89, 1.10)
SUV_{total}	0.36	0.989 (0.978, 0.994)	18.5	67.1	1.00 (0.70, 1.44)

*Data in parentheses are 95% confidence intervals.

†Data in parentheses are 95% LOA.

‡Two patients underwent partial whole-body scans.

RC = repeatability coefficient for α = 0.05 (log-transformed SUV); CV = log-transformed coefficient of variation; CPD = critical percentage difference; B = ratio of test-to-retest bias.

B and 95% LOA have been back-transformed to original units.

study showed that the lesion SUV_{mean} for benign degenerative disease was 11.1 ± 3.8 g/mL (*27*). Therefore, in this study, we applied an SUV threshold of 15 to minimize the inclusion of benign disease.

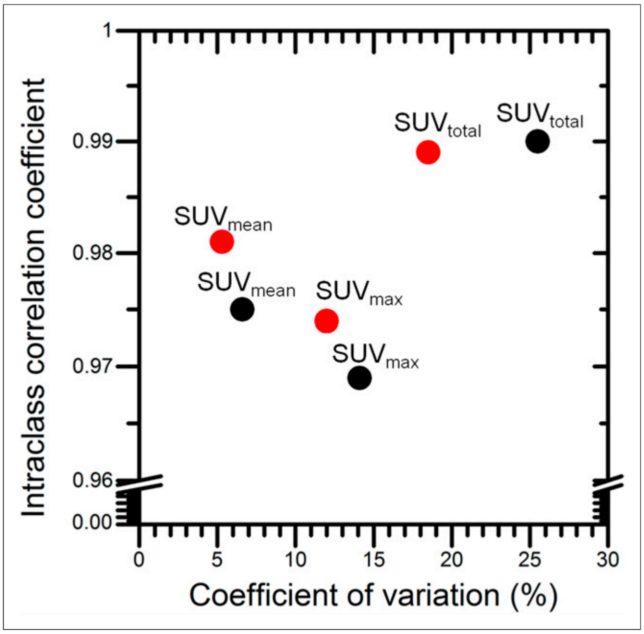

FIGURE 5. Overall ICC plotted against overall coefficient of variation of lesion-level (black) and patient-level (red) SUV metrics.

The ^{18}F-NaF PET findings were more repeatable than the findings of a multicenter ^{18}F-FDG PET study on patients with lung cancer and gastrointestinal malignancies (*17*). Such effects as respiratory motion may lead to increased random error in ^{18}F-FDG PET images of certain regions, more so in soft tissue than in bone (*17*). In comparing the repeatability of SUV metrics, one study also found SUV_{mean} to be more repeatable than the SUV_{max} of individual lesions (*28*).

One important aspect of this multicenter study was that although the PET scans were acquired on different scanners with different acquisition parameters, the scanners were harmonized. Despite image harmonization, we found that for all 3 SUV metrics, the variance in lesion-level test–retest measurements was significantly smaller at UWCCC than at the other sites. The repeatability differences between sites might have been due to physiologic factors such as circadian rhythm or different degrees of conformation to the imaging protocol (*29,30*). For example, the mean (±SD) postinjection time (61 ± 1 min at UWCCC vs. 69 ± 9 min at MSKCC) and injected dose (178 ± 9 MBq at UWCCC vs. 136 ± 32 MBq at NCI) varied by site (Supplemental Table 1; supplemental materials are available at http://jnm.snmjournals.org).

There is active discussion on whether it is lesion or patient measurements that should be used to assess treatment response. In ^{18}F-FDG PET, there are previous studies on the test–retest variability in uptake for individual lesions and for the whole patient (*31*). Weber et al. found that averaging the measurements of several lesions in a patient did not significantly affect the

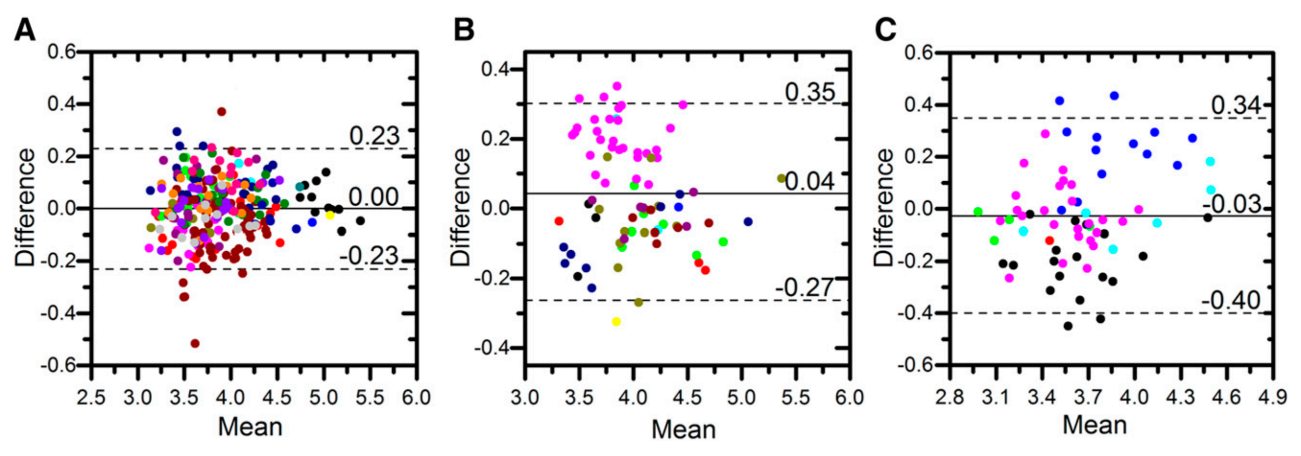

FIGURE 6. Bland–Altman plots of lesion SUV_{max} by site: UWCCC (265 lesions) (A), MSKCC (78 lesions) (B), and NCI (68 lesions) (C). Each point represents a lesion, and different subjects are indicated by different colors. Solid lines denote site-specific mean difference, and dotted lines denote site-specific upper and lower 95% LOA. Both mean and difference uptake values have been log-transformed.

repeatability of the SUV metrics (*17*). Our study confirmed similar repeatability between lesion and patient ROIs. Measuring the repeatability of lesion ROIs enables evaluation of the lesion-specific response to therapy and may more comprehensively represent patient response.

The statistical limits of agreement for ^{18}F-NaF PET SUV metrics were established at both the lesion level and the patient level such that 95% LOA ($\alpha = 0.05$) could be applied to reflect true changes in uptake. An SUV percentage decrease to less than the 95% LOA lower limit can be considered response, and an increase to more than the upper limit can be considered progression.

CONCLUSION

The repeatability of ^{18}F-NaF PET/CT–derived SUV_{max}, SUV_{mean}, and SUV_{total} was assessed for both lesion-level and patient-level ROIs in a multicenter prospective study on CRPC metastatic to bone. Low repeatability coefficients, high ICCs, and small coefficients of variation in test–retest scans were found. Patient-level repeatability was slightly superior to lesion-level repeatability, justifying the use of SUV both in individual lesions and across the whole body. These results can be used to establish quantitative criteria for ^{18}F-NaF PET assessment of treatment response in patients with CRPC metastatic to bone.

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 study was supported by the Prostate Cancer Foundation (PCF) through the PCF Creativity Award and the PCF Mazzone Challenge Award to Drs. Liu and Jeraj and was conducted within the Prostate Cancer Clinical Trials Consortium (PCCTC). No other potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We thank the patients who volunteered their time, and we thank the imaging technologists who acquired the data.

REFERENCES

1. Logothetis CJ, Lin SH. Osteoblasts in prostate cancer metastasis to bone. *Nat Rev Cancer.* 2005;5:21–28.
2. Costelloe CM, Chuang HH, Madewell JE, Ueno NT. Cancer response criteria and bone metastases: RECIST 1.1, MDA and PERCIST. *J Cancer.* 2010;1:80–92.
3. Blau M, Ganatra R, Bender MA. ^{18}F-fluoride for bone imaging. *Semin Nucl Med.* 1972;2:31–37.
4. Schirrmeister H, Glatting G, Hetzel J, et al. Prospective evaluation of the clinical value of planar bone scans, SPECT, and ^{18}F-labeled NaF PET in newly diagnosed lung cancer. *J Nucl Med.* 2001;42:1800–1804.
5. Even-Sapir E, Metser U, Mishani E, Lievshitz G, Lerman H, Leibovitch I. The detection of bone metastases in patients with high-risk prostate cancer: ^{99m}Tc-MDP planar bone scintigraphy, single- and multi-field-of-view SPECT, ^{18}F-fluoride PET, and ^{18}F-fluoride PET/CT. *J Nucl Med.* 2006;47:287–297.
6. Czernin J, Satyamurthy N, Schiepers C. Molecular mechanisms of bone ^{18}F-NaF deposition. *J Nucl Med.* 2010;51:1826–1829.
7. Iagaru A, Mittra E, Dick DW, Gambhir SS. Prospective evaluation of Tc-99m MDP scintigraphy, F-18 NaF PET/CT, and F-18 FDG PET/CT for detection of skeletal metastases. *Mol Imaging Biol.* 2012;14:252–259.
8. Mick CG, James T, Hill JD, Williams P, Perry M. Molecular imaging in oncology: ^{18}F-sodium fluoride PET imaging of osseous metastatic disease. *AJR.* 2014;203:263–271.
9. Morisson C, Jeraj R, Liu G. Imaging of castration-resistant prostate cancer: development of imaging response biomarkers. *Curr Opin Urol.* 2013;23:230–236.
10. Wondergem M, van der Zant FM, van der Ploeg T, Knol RJ. A literature review of ^{18}F-fluoride PET/CT and ^{18}F-choline or ^{11}C-choline PET/CT for detection of bone metastases in patients with prostate cancer. *Nucl Med Commun.* 2013;34:935–945.
11. Front D, Israel O, Jerushalmi J, et al. Quantitative bone-scintigraphy using SPECT. *J Nucl Med.* 1989;30:240–245.
12. Brenner W, Vernon C, Muzi M, et al. Comparison of different quantitative approaches to F-18-fluoride PET scans. *J Nucl Med.* 2004;45:1493–1500.
13. Hawkins RA, Choi Y, Huang SC, et al. Evaluation of the skeletal kinetics of fluorine-18-fluoride ion with PET. *J Nucl Med.* 1992;33:633–642.
14. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–310.
15. 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.
16. 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.

17. Weber WA, Gatsonis CA, Mozley PD, et al. Repeatability of ^{18}F-FDG PET/CT in advanced non-small cell lung cancer: prospective assessment in 2 multicenter trials. *J Nucl Med.* 2015;56:1137–1143.
18. Kurdziel KA, Shih JH, Apolo AB, et al. The kinetics and reproducibility of ^{18}F-sodium fluoride for oncology using current PET camera technology. *J Nucl Med.* 2012;53:1175–1184.
19. Rohren EM, Etchebehere EC, Araujo JC, et al. Determination of skeletal tumor burden on ^{18}F-fluoride PET/CT. *J Nucl Med.* 2015;56:1507–1512.
20. Yip S, Jeraj R. Use of articulated registration for response assessment of individual metastatic bone lesions. *Phys Med Biol.* 2014;59:1501–1514.
21. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res.* 2015;24:27–67.
22. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135–160.
23. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat.* 2007;17:571–582.
24. 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.
25. Portney L, Watkins MP. *Foundations of Clinical Research: Applications to Practice.* Philadelphia, PA: F.A. Davis Company; 2015:588–598.
26. Muzahir S, Jeraj R, Liu G, et al. Differentiation of metastatic vs degenerative joint disease using semi-quantitative analysis with F-18-NaF PET/CT in castrate resistant prostate cancer patients. *Am J Nucl Med Mol Imaging.* 2015;5:162–168.
27. Oldan JD, Hawkins AS, Chin BB. F-18 sodium fluoride PET/CT in patients with prostate cancer: quantification of normal tissues, benign degenerative lesions, and malignant lesions. *World J Nucl Med.* 2016;15:102–108.
28. 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.
29. Binns DS, Pirzkall A, Yu W, et al. Compliance with PET acquisition protocols for therapeutic monitoring of erlotinib therapy in an international trial for patients with non-small cell lung cancer. *Eur J Nucl Med Mol Imaging.* 2011;38:642–650.
30. Generali D, Berruti A, Tampellini M, et al. The circadian rhythm of biochemical markers of bone resorption is normally synchronized in breast cancer patients with bone lytic metastases independently of tumor load. *Bone.* 2007;40:182–188.
31. 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.