

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

Christie Lin¹, Tyler Bradshaw², Timothy Perk¹, Stephanie Harmon¹, Jens Eickhoff³, Ngoneh Jallow⁴, Peter L. Choyke⁵, William L. Dahut⁶, Steven Larson⁷, John Laurence Humm⁷, Scott Perlman^{2,9}, Andrea B. Apolo⁶, Michael J. Morris⁸, Glenn Liu^{1,9}, Robert Jeraj^{1,9}

¹Department of Medical Physics, University of Wisconsin – Madison, WI, USA

²Department of Radiology, University of Wisconsin – Madison, WI, USA

³Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison, WI, USA

⁴Department of Radiology and Imaging Sciences, Emory University – Atlanta, GA, USA

⁵Molecular Imaging Program, Center for Cancer Research, National Cancer Institute – Bethesda, MD, USA

⁶Medical Oncology Branch, Center for Cancer Research, National Cancer Institute – Bethesda, MD, USA

⁷Molecular Imaging and Therapy Service, Department of Radiology, Memorial Sloan Kettering – New York, NY, USA

⁸Department of Medical Oncology, Memorial Sloan Kettering – New York, NY, USA

⁹University of Wisconsin Carbone Cancer Center – Madison, WI, USA

*Correspondence should be addressed to:

Department of Medical Physics, Wisconsin Institutes for Medical Research, 1111 Highland Ave, Room 7003, Madison, WI 53705; Tel: 608-263-8619; Fax: 608-262-2413

Corresponding author: Robert Jeraj, PhD; Email: rjeraj@wisc.edu

First author: Christie Lin, MS; Email: clin232@wisc.edu

Running Title: Repeatability of ^{18}F -NaF PET

Article category: scientific article

Key words: sodium fluoride, PET, repeatability, metastatic prostate cancer, multicenter clinical trial

Word Count:

Abstract: 326

Text: 2658

Total: 4947

ABSTRACT

Purpose

¹⁸F-NaF, a positron emission tomography (PET) radiotracer of bone turnover, has shown potential as an imaging biomarker for assessing therapeutic response of bone metastases. This study aims to evaluate the repeatability of ¹⁸F-NaF PET-derived SUV metrics in individual bone lesions from patients in a multicenter study.

Methods

Thirty-five metastatic castrate-resistant prostate cancer patients with multiple metastases received two whole-body (test-retest) ¹⁸F-NaF PET/computed tomography (CT) scans 3±2 days apart from one of three imaging sites. A total of 411 bone lesions larger than 1.5 cm³ were automatically segmented using standardized uptake value (SUV) >15 g/mL threshold. Two levels of analysis were performed: lesion-level, where measures were extracted from individual lesion regions of interest (ROI), and patient-level, where all lesions within a patient were grouped into a single ROI for analysis. Uptake was quantified with the maximum (SUV_{max}), average (SUV_{mean}), and total (SUV_{total}) SUV. Test-retest repeatability was assessed using Bland-Altman analysis, intraclass correlation coefficient (ICC), coefficient of variation (CV), critical percent difference, and repeatability coefficient. The 95% limits of agreement (LOA_{95%}) of the ratio between test-retest measurements were calculated.

Results

At the lesion-level, CV for SUV_{max}, SUV_{mean}, and SUV_{total} were 14.1%, 6.6%, and 25.5%, respectively. At the patient-level, CV was slightly smaller: 12.0%, 5.3%, and 16.5%, respectively. ICC was excellent (ICC>0.95) for all imaging metrics. Lesion-level LOA_{95%} for SUV_{max}, SUV_{mean}, and SUV_{total} were (0.76, 1.32), (0.88, 1.14), and (0.63, 1.71), respectively. Patient-level LOA_{95%} were 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

Repeatability of ^{18}F -NaF PET/CT of SUV_{max} , SUV_{mean} , and $\text{SUV}_{\text{total}}$ were similar between lesion- and patient-level regions of interest. We found significant differences in lesion-level and patient-level difference distributions between sites. These results can be used to establish NaF PET-based treatment response assessment criteria at the lesion- and patient-levels. NaF PET demonstrates repeatability levels useful for clinical quantification of bone lesion response to therapy.

INTRODUCTION

Prostate cancer is distinct among solid tumors in that its advancement largely presents 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 in bone can improve clinical care, particularly in advanced prostate cancer.

Radiolabeled sodium fluoride, ^{18}F -NaF, was first introduced by Blau *et al* in 1962 for the detection of bone lesions with positron emission tomography (PET). However, NaF was largely replaced by bone scintigraphy utilizing $^{99\text{m}}\text{Tc}$ because of superior imaging characteristics with conventional gamma cameras and the readily available supply of $^{99\text{m}}\text{Tc}$ (3-6). With recent technological advances in PET, NaF PET has been increasingly used for detecting bone metastases because of its higher specificity and sensitivity as compared to planar bone scintigraphy and SPECT (3,4,7-10). NaF PET shows potential for longitudinal disease assessment, as its standardized uptake values (SUV) in both normal and pathologic bone are representative of changes in bone metabolism (11-13).

Repeatability of a biomarker, defined as the variation of repeated measurements in an experiment performed under the same conditions, is necessary to measure in order to accurately assess tumor response (14). Repeatability of ^{18}F -FDG based on double baseline studies has been well studied, permitting the development of PET Response Criteria In Solid Tumors (15-17). No such criterion exists for evaluating quantitative ^{18}F -NaF PET response.

A previous study evaluated the repeatability of NaF PET activity in bone uptake within the whole body (18). However, repeatability can also be evaluated of individual bone lesion regions of interest (ROIs), allowing the assessment of how a tumor's response may uniquely contribute to the disease burden on the patient as a whole. The evaluation of repeatability of uptake in an individual lesion would allow for assessment of response heterogeneity across within the patient.

Here we report on the first multi-center study assessing the repeatability of NaF PET uptake at the lesion-level. In addition, we compared the repeatability values between 3 different imaging sites in a multicenter trial.

METHODS

Patient population and study design

This was a prospective, non-randomized two-arm, multi-institutional pharmacodynamic imaging clinical trial with the primary objective to determine the repeatability of NaF PET/computed tomography (CT) imaging for evaluating osseous metastases in metastatic castrate-resistant prostate cancer (mCRPC) patients. Eligible patients aged 18 or older with progressive mCRPC with bone scan confirmed osseous metastases were enrolled in either docetaxel-based chemotherapy or androgen receptor directed therapy between 2/2012 and 9/2014 at University of Wisconsin (UWCCC), Memorial Sloan-Kettering Cancer Center (MSKCC), or the National Cancer Institute (NCI). Patients demonstrated histologically proven adenocarcinoma. Exclusion criteria included active systemic treatment for prostate cancer, palliative radiation within 4 weeks of registration, or any prior radioisotope treatment for prostate cancer. The Institutional Review Board and Radiation Safety Committee of each participating institution approved this retrospective study and all subjects signed a written informed consent. A sample size of n=20 patients per site was proposed to evaluate repeatability. This sample size provided sufficient power ($\geq 80\%$) to detect the anticipated excellent level of repeatability at each of the three study sites at the one-sided 0.0167 significance level.

Quantitative image acquisition

Test-retest NaF PET/CT whole-body scans were to be performed 2-5 days apart and prior to start of therapy. Patients were injected intravenously with a bolus of 111-185 MBq (3-5 mCi) of NaF and imaged 60 minutes post-injection for 3 minutes per bed position from feet to skull vertex. Scans at UWCCC and MSKCC were acquired on the Discovery VCT (GE Healthcare, Waukesha, WI) PET/CT scanner, and scans at NCI were acquired on the Gemini

(Philips Healthcare, Amsterdam, Netherlands) PET/CT scanner. PET images were attenuation and scatter-corrected.

Scanner harmonization

Quantitative harmonization of scanners was achieved to obtain equivalent image quality and quantitative accuracy across scanners. The GE scanners were harmonized to the Philips scanner. Harmonization was performed using a uniform phantom to measure the signal-to-noise ratio, and the National Electrical Manufacturers Association International Electrotechnical Commission body phantom. Absolute calibration was measured by the recovery coefficient, defined as the ratio of the mean measured activity concentration in to the true activity concentration in the ROI. Difference between scanners in recovery coefficient and signal-to-noise ratio was minimized by systemically varying reconstruction parameters such as number of iterations, number of subsets, and post-reconstruction filter.

ROI definition

Automatic identification and segmentation of lesions was achieved with a CT mask applied to exclude soft tissue uptake followed by a $SUV > 15$ g/mL threshold to exclude additional activity with low statistical likelihood of being malignant (18,19). Lesion contours were verified by an experienced nuclear medicine physician on PET/CT images and contours smaller than 1.5 cm^3 as measured by PET volume were excluded. Matching of corresponding lesions between paired scans was performed automatically using articulated registration (20).

Two levels of SUV analysis were performed: lesion-level analysis, where SUV metrics were extracted from each individual lesion ROI (iROI), and patient-level analysis, where all lesions for a patient were grouped into a single patient ROI (pROI) before SUV analysis. For both ROI levels, SUV_{max} is defined as the maximum SUV of the ROI. SUV_{total} is defined as the total summed SUV of the ROI normalized to voxel volume. SUV_{mean} of iROI is the mean SUV within the ROI and SUV_{mean} of pROI is the mean of SUV_{mean} of all lesions in the patient. The two

different levels of analysis will be differentiated using the notation iSUV for individual lesion-level SUV metrics, and pSUV for patient-level SUV metrics.

Statistical analysis

Primary outcome measures for evaluating repeatability of SUV metrics were intra-class correlation coefficient (ICC) and repeatability coefficient (RC). RC was calculated at $\alpha = 0.05$. ICC was estimated using a two-way mixed effects model.

Additional statistical measures for evaluating repeatability of quantitative imaging biomarkers as recommended by the Quantitative Imaging Biomarkers Alliance or previously reported in literature were investigated (21). Test-retest agreement for each ROI was evaluated with the Bland-Altman analysis method for repeated observations (22,23).

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

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

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

Limits of agreement ($LOA_{95\%}$) were calculated for the ratio of the test (m_A) to retest (m_B) measurements. Within $LOA_{95\%}$ lies the ratio of m_B/m_A with a probability of 95%:

$$LOA_{95\%} = (e^{(B-RC)}, e^{(B+RC)}) \quad (1)$$

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

One-way analysis of variance with pairwise comparisons and two-sample t-test were used to assess whether the bias for each metric significantly differed between sites. Two-sample F-tests were used to evaluate if variability across sites.

RESULTS

A total of 411 NaF-avid bone lesions from 35 mCRPC patients imaged at one of three sites were evaluated (Fig. 1). Patients were injected intravenously with 159.8 ± 9.7 MBq of NaF and test-retest NaF PET/CT whole-body scans were performed 63 ± 7 minutes post-injection (3 ± 2 days apart). Dose infiltration near the injection site was minimal in all scans. Two of 35 patients received partial whole-body scans due to patient repositioning during the scan. Lesion and patient characteristics are summarized in Table 1. Harmonization reconstruction parameters including reconstruction method, grid size, subset, iteration, and post reconstruction filter, for each of the scanners are summarized in **Error! Reference source not found.**

The median number of lesions per patient at baseline was 8 (range: 1-69 lesions). Lesions were located across the skeleton, with the predominant lesion site being the spine. For all lesions, median $iSUV_{max}$ was 44.8 (range: 19.6–225.5), $iSUV_{mean}$ was 23.7 (range: 16.7–75.8), and $iSUV_{total}$ was 116.7 (range: 26.4–5628.0) g/mL. For all patients, median $pSUV_{max}$ was 86.4 (range: 29.6–225.5), $pSUV_{mean}$ was 25.4 (range: 18.4–51.1), and $pSUV_{total}$ was 2429.3 (range: 47.7–21,447) g/mL.

Relative difference between test-retest scans tend to be slightly greater at the lesion-level than at the patient-level. For all metrics, distributions of relative difference were narrower for pROI than iROI (Fig. 2). SUV_{mean} had the smallest relative difference for both ROIs. For iROI, $iSUV_{mean}$ was most repeatable (inner-quartile range, IQR=2.5%) followed by SUV_{max} (IQR=4.4%)

and $iSUV_{total}$ (IQR=5.1%). For pROI, $pSUV_{mean}$ was most repeatable (IQR=2.0%), followed by $pSUV_{total}$ (IQR=2.6%), and $pSUV_{max}$ (IQR=3.3%).

Figs. 3 and 4 contain Bland-Altman plots, which demarcate RC. For lesion-level SUV metrics, $iSUV_{mean}$ had the smallest variability (RC=0.13), followed by $iSUV_{max}$ (RC=0.27), and $iSUV_{total}$ (RC=0.49). Fig. 4 contains Bland-Altman plots for each patient-level SUV metric. Similarly, $pSUV_{mean}$ was most repeatable (RC=0.10), followed by $pSUV_{max}$ (RC=0.24) and $pSUV_{total}$ (RC=0.36). Both lesion-level and patient-level distributions have approximately normal distributions and heteroscedasticity.

According to RC, CV, and CPD, SUV_{mean} was the most repeatable followed by SUV_{max} , and SUV_{total} at both the lesion- and patient-levels (Tables 3 and 4). The $LOA_{95\%}$ of the ratio of test-retest measurements define the interval containing the ratio of test-retest measurements for each imaging metric. $LOA_{95\%}$ from each site were widely overlapping for all three metrics. At the lesion-level, $LOA_{95\%}$ was narrowest for $iSUV_{mean}$ at 1.00 ($LOA_{95\%}$: 0.88, 1.14), followed by $iSUV_{max}$ at 1.00 ($LOA_{95\%}$: 0.76, 1.32), and $iSUV_{total}$ at 1.04 ($LOA_{95\%}$: 0.63, 1.71). At the patient-level, the overall ratio between test-retest of $pSUV_{mean}$ was 0.99 ($LOA_{95\%}$: 0.89, 1.10), ratio of $pSUV_{max}$ was 1.00 ($LOA_{95\%}$: 0.79, 1.26), and ratio of $pSUV_{total}$ was 1.00 ($LOA_{95\%}$: 0.70, 1.44), respectively. Across imaging metrics, $LOA_{95\%}$ was consistently narrowest for SUV_{mean} . Across sites, $LOA_{95\%}$ was consistently narrowest, though not significantly different, for UWCCC.

A comparison of overall CV and ICC are shown in Fig. 5. At both the lesion- and patient-levels, ICC was highest for $iSUV_{total}$ followed by $iSUV_{mean}$ and $iSUV_{max}$. Consistently, patient-level SUV metrics present lower CV than do lesion-level SUV metrics.

Shown in Fig. 6 are Bland-Altman plots of lesion-level $iSUV_{max}$ by site. MSKCC had statistically significantly different sample mean ($p = .004$) and UWCCC had significantly smaller variance ($p < .001$) as compared to the other two sites. In addition, the variance of $iSUV_{mean}$ ($p < .001$) and $iSUV_{total}$ ($p < .001$) at UWCCC were significantly smaller as compared to Sites 2 and 3. The sample mean of $iSUV_{mean}$ at MSKCC was differed significantly from the rest ($p < .001$).

For pROI, the sole difference between sites was that the variance of pSUV_{total} was significantly smaller for UWCCC ($p = .003$).

DISCUSSION

This is the first multicenter study with results demonstrating the repeatability of multiple NaF PET SUV metrics, SUV_{max}, SUV_{mean}, and SUV_{total}, for both lesion-level and patient-level ROIs.

While different guidelines exist for the interpretation of ICC, one of the most common guidelines defines the range $0.40 < ICC < 0.75$ as moderate repeatability, and > 0.75 as excellent repeatability (25). While 95% confidence intervals of ICC of SUV_{max}, SUV_{mean}, and SUV_{total} at the lesion-level were excellent for all sites, the 95% confidence intervals of the ICCs of pSUV_{mean} and pSUV_{max} at MSKCC and NCI are not fully contained within the region of excellent repeatability. The target patient accrual goal was not met due to an imbalance of accrual between the two arms of therapy, thus decreasing 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 iSUV_{max} in Fig. 6, multiple lesions within the same patient have a tendency to show correlated repeatability. Thus, it is important to note that it was not possible to regard each lesion as independent. The intra-patient correlations were taken into account by implementing the Bland-Altman analysis for repeated measures (23).

Our repeatability results at the patient-level support those of the previous NaF PET study in bone lesions. At the patient-level, our findings show similar levels of repeatability for pSUV_{max}, pSUV_{mean}, and pSUV_{total} as compared to a study conducted by Kurdziel *et al* (18). Despite the differences in lesion segmentation methods, our study shows similar ICC and CPD results for the three SUV metrics investigated in the Kurdziel study.

The application of both an uptake threshold and volume threshold were applied to minimize the probability of identifying benign disease. While Kurdziel *et al* used a segmentation threshold of SUV > 10 , a later study by Rohren *et al* showed that ROIs identified with iSUV_{max} $>$

10 still included normal bone activity (19). A study showed that $iSUV_{max} < 12$ g/mL always represented a site of benign disease (26). Another study showed that $iSUV_{mean}$ of benign degenerative disease was 11.1 ± 3.8 g/mL (27). Therefore in this study, we applied the threshold of $SUV > 15$ to minimize the inclusion of benign disease.

NaF PET findings demonstrated higher repeatability as compared to a multicenter study of ^{18}F -FDG PET imaging in patients with lung cancer and gastrointestinal malignancies (17). Patient effects such as respiratory motion may lead to increased random error in FDG PET of regions, to a greater effect in tissue than in bone structures (17). In comparing the repeatability of SUV metrics, SUV_{mean} was also found to be more repeatable than SUV_{max} of individual lesions (28).

One important aspect of this multicenter study was that although PET scans were acquired on different scanners with different acquisition parameters, the scanners were harmonized. Despite image harmonization, we found that UWCCC had significantly smaller variance in lesion-level test-retest measurements as compared to the other sites for all three imaging metrics. Rather, the repeatability differences between sites may be due to physiological factors such as circadian rhythm or different degrees of conformation to the imaging protocol (29,30). The mean (standard deviation) of post-injection time (UWCCC: 61(1) min vs. MSKCC: 69(9) min) and injected dose (UWCCC: 178(9) MBq vs. NCI: 135(32) MBq) varied by site.

There is active discussion on whether to use single lesion or patient-level measurements for treatment response assessment. In ^{18}F -FDG PET, there are previous studies on the test-retest variability of ^{18}F -FDG PET uptake for individual lesions and for the whole patient (31). Weber *et al* found that averaging measurements of several lesions in a patient did not have significant impact on the repeatability of the SUV metrics (17). Our study confirms similar repeatability between lesion-level and patient-level ROIs. Measuring the repeatability of lesion-level ROIs enables the evaluation of lesion-specific response to therapy, which may provide a more comprehensive representation of patient response.

Statistical limits of agreement for NaF PET metrics were established at both lesion- and patient-level such that $LOA_{95\%}$ ($\alpha=0.05$) can be applied to reflect true changes in uptake; percent decline in SUV less than the lower limit of $LOA_{95\%}$ can be considered response, and increase in SUV greater than the upper limit can be considered progression.

CONCLUSION

Repeatability of NaF PET/CT-derived SUV_{max} , SUV_{mean} , and SUV_{total} were assessed for both lesion- and patient-level ROI in a multicenter prospective study in bone-metastatic CRPC. Low repeatability coefficients, high intraclass correlation coefficients, and small coefficient of variations in test-retest scans were found. Patient-level repeatability was slightly superior to that of lesion-level, justifying the use of SUV both in individual lesions and across the whole body. Results can be used to establish quantitative criteria for treatment response assessment using NaF PET in patients with bone-metastatic CRPC.

Acknowledgements

We would like to thank the patients who volunteered their time and imaging technologists for data acquisition. This study was supported by the Prostate Cancer Foundation (PCF) through the PCF Creativity Award (Liu, Jeraj) and PCF Mazzone Challenge Award (Jeraj, Liu), and conducted within the Prostate Cancer Clinical Trials Consortium (PCCTC).

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. Schirrmeyer 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.
4. 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: 99mTc-MDP Planar bone scintigraphy, single- and multi-field-of-view SPECT, 18F-fluoride PET, and 18F-fluoride PET/CT. *J Nucl Med*. 2006;47:287-297.
5. Blau M, Ganatra R, Bender MA. 18 F-fluoride for bone imaging. *Semin Nucl Med*. 1972;2:31-37.
6. Czernin J, Satyamurthy N, Schiepers C. Molecular mechanisms of bone 18F-NaF deposition. *J Nucl Med*. 2010;51:1826-1829.
7. Iagaru A, Mitra 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 Am J Roentgenol*. 2014;203:263-271.
9. Morisson C, Jeraj R, Liu G. Imaging of castration-resistant prostate cancer: development of imaging response biomarkers. *Curr Opin in Urol*. 2013;23:230-236.
10. Wondergem M, van der Zant FM, van der Ploeg T, Knol RJ. A literature review of 18F-fluoride PET/CT and 18F-choline or 11C-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;50:122s-150s.
16. Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of 18F-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 18F-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 18F-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 F-18-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, Hall, LT, Rio, AM, Perk, T, Jaskowiak, C, Perlman, SB. 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 J, 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 F-18-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.

FIGURES

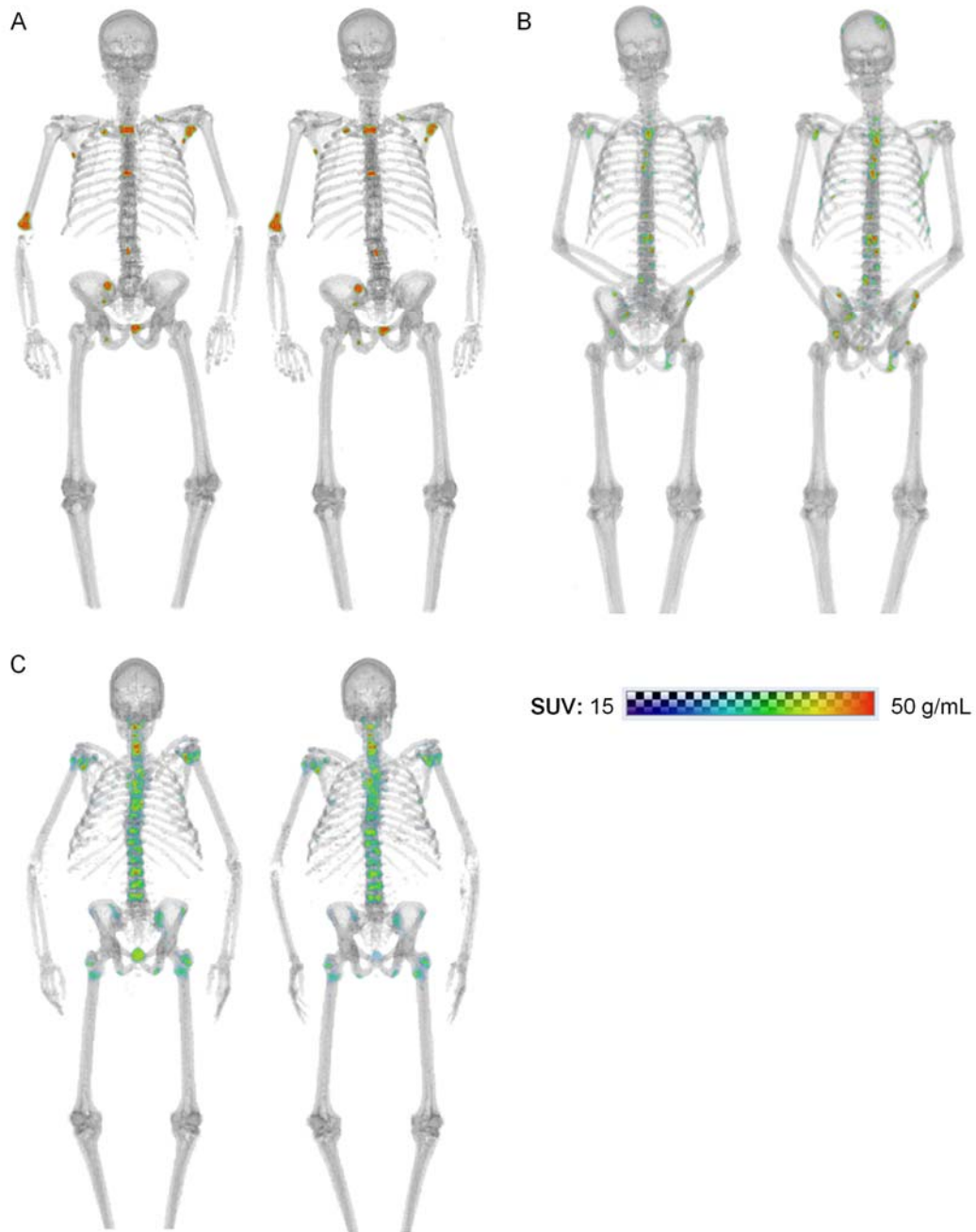


Figure 1. Whole body paired baseline scans on NaF PET/CT of males with metastatic castrate-resistant prostate cancer: (A) 74 year-old images 3 days apart at UWCCC. (B) 57 year-old imaged 2 days apart at MSKCC. (C) 69 year-old imaged 1 day apart at NCI

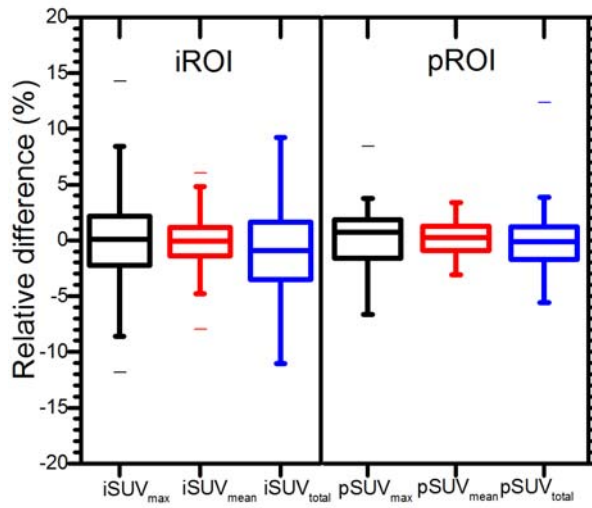


Figure 2. Box plots of relative differences (%) of each NaF PET SUV metrics (log-transformed) for all lesion-level ROIs (*left*, $n = 411$ lesions) and for patient-level ROIs (*right*, $n = 35$ patients). The whiskers extend from minimum to maximum values.

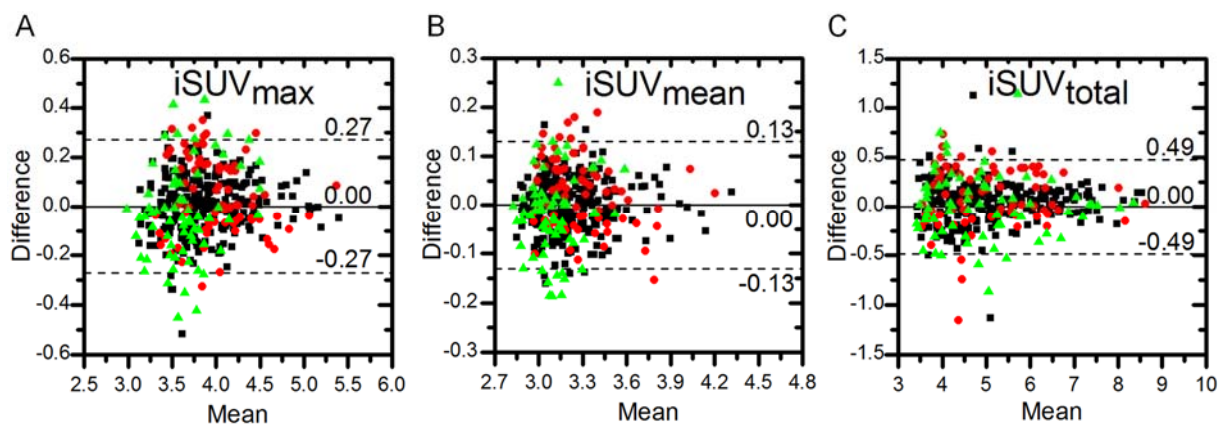


Figure 3. Bland-Altman plots of imaging metrics for all lesion-level regions of interest (n=411 lesions), including (A) $iSUV_{max}$, (B) $iSUV_{mean}$, and (C) $iSUV_{total}$. Sites are indicated by the symbol (■ UWCCC, ● MSKCC, ▲ NCI). Solid horizontal line denotes the mean difference, and the 95% LOAs are indicated by the dotted lines. Both the mean and difference uptake values have been log-transformed.

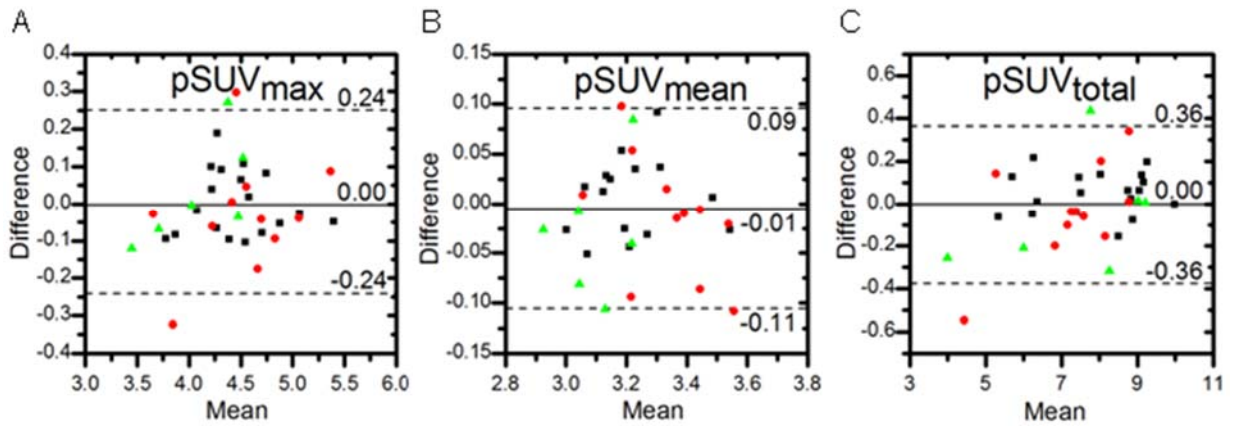


Figure 4. Bland-Altman plots of imaging metrics for all patient-level regions of interest (n=35), including (A) pSUV_{max}, (B) pSUV_{mean}, and (C) pSUV_{total}. Sites are indicated by the symbol (■ UWCCC, ● MSKCC, ▲ NCI). Solid horizontal line denotes the mean difference, and the 95% limits of agreement are indicated by the dotted lines. Both the mean and difference values have been log-transformed.

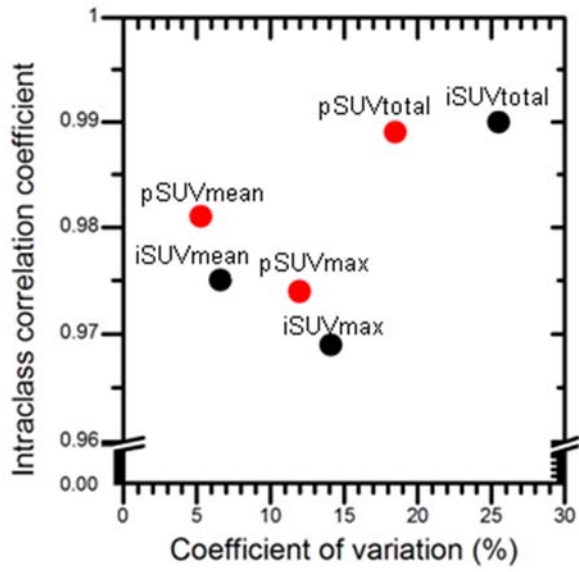


Figure 4. Overall intraclass correlation coefficient plotted against the overall coefficient of variation of lesion-level (black) and patient-level (red) NaF PET SUV metrics.

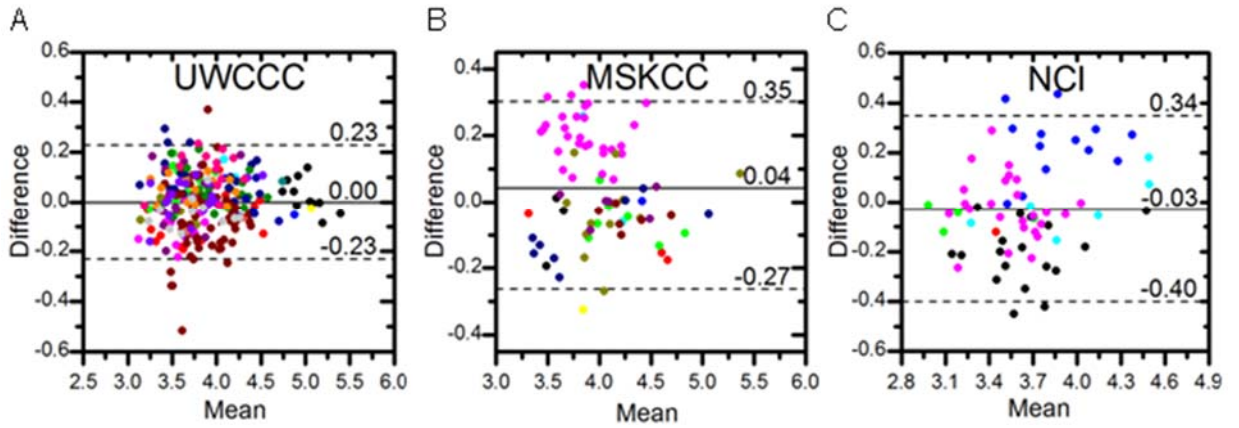


Figure 6. Bland-Altman plots of $iSUV_{max}$ by site: (A) UWCCC ($n = 265$), (B) MSKCC ($n = 78$), and (C) NCI ($n = 68$). Each point represents a lesion and each color represents a subject. Solid horizontal line denotes the site-specific mean difference, and the dotted line denotes the site-specific upper and lower 95% limits of agreement. Both the mean and difference uptake values have been log-transformed.

TABLES

Table 1

Patient demographics. All patients had metastatic castrate-resistant prostate adenocarcinoma with metastatic bone lesions identified by NaF PET/CT.

Characteristics	Patients by imaging site		
	UWCCC (n = 18)	MSKCC (n = 11)	NCI (n = 6)
Age (years)			
median (range)	72.5 (47 - 87)	75.0 (57 - 81)	68 (57 - 83)
Height (cm)			
median (range)	178 (166 - 191)	177 (162 - 191)	171 (161 - 189)
Weight (kg)			
median (range)	92.3 (70.7 - 145.0)	94.0 (73.0 - 119.0)	84.6 (75.4 - 91.6)
PSA			
median (range)	71.2 (1.6 - 310.0)	8.1 (2.5 - 246.8)	85.9 (32.0 - 460.7)
Gleason Score			
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 (range)	200 (139 - 470)	219 (157 - 251)	264 (119 - 903)
Hgb (g/dL)			
median (range)	12.8 (7.7 - 14.9)	13.8 (11.3 - 15.3)	11.8 (9.0 - 13.9)
Number of lesions			
≤ 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%)

Table 2
Scanner harmonization parameters by imaging site.

	UWCCC	MSKCC	NCI
Scanner	GE Discovery VCT	GE Discovery VCT	Philips Gemini
Reconstruction	3D OSEM	3D OSEM	3D OSEM
Grid Size	256x256	256x256	144x144
Subset	14	14	33
Iteration	2	2	2
Post filter	4 mm	4 mm	--

Table 3
Repeatability of lesion-level NaF PET SUV metrics.

UWCCC (n = 265 lesions)	RC	ICC (CI_{95%})	CV (%)	CPD (%)	B (LOA_{95%})
iSUV _{max}	0.23	0.980 (0.974, 0.984)	11.7	37.5	1.00 (0.79, 1.25)
iSUV _{mean}	0.10	0.983 (0.979, 0.987)	5.5	15.9	1.00 (0.90, 1.11)
iSUV _{total}	0.40	0.990 (0.987, 0.992)	20.7	75.9	1.04 (0.69, 1.56)
MSKCC (n = 78 lesions)					
iSUV _{max}	0.31	0.958 (0.935, 0.973)	16.8	54.3	1.04 (0.75, 1.45)
iSUV _{mean}	0.14	0.970 (0.953, 0.981)	7.8	22.2	1.03 (0.88, 1.19)
iSUV _{total}	0.60	0.990 (0.985, 0.994)	32.7	133.6	1.08 (0.57, 2.06)
NCI (n = 68 lesions)					
iSUV _{max}	0.37	0.865 (0.791, 0.915)	20.6	69.2	0.97 (0.65, 1.46)
iSUV _{mean}	0.16	0.876 (0.807, 0.922)	9.2	26.2	0.98 (0.82, 1.17)
iSUV _{total}	0.65	0.993 (0.989, 0.996)	36.6	151.4	1.00 (0.49, 2.06)
All Sites (n = 411 lesions)					
iSUV _{max}	0.27	0.969 (0.963, 0.975)	14.1	47.2	1.00 (0.76, 1.32)
iSUV _{mean}	0.13	0.975 (0.970, 0.980)	6.6	19.6	1.00 (0.88, 1.14)
iSUV _{total}	0.49	0.990 (0.988, 0.992)	25.5	100.4	1.04 (0.63, 1.71)

RC = repeatability coefficient for $\alpha=0.05$ (log-transformed SUV); ICC = intraclass correlation coefficient; CI_{95%} = 95% confidence intervals; CV = coefficient of variation (log-transformed); CPD = critical percent difference; B = the ratio of the test-retest bias; LOA_{95%} = 95% limits of agreement of the ratio of test-retest measurements. B and LOA_{95%} have been back-transformed to original units.

Table 4
Repeatability of patient-level NaF PET SUV metrics.

UWCCC (n = 18 patients)	RC	ICC (CI_{95%})	CV (%)	CPD (%)	B (LOA_{95%})
pSUV _{max}	0.17	0.984 (0.959, 0.994)	8.8	27.6	1.00 (0.84, 1.19)
pSUV _{mean}	0.08	0.990 (0.974, 0.996)	4.2	12.3	1.01 (0.93, 1.09)
pSUV _{total}	0.20	0.993 (0.981, 0.999)	10.1	32.2	1.05 (0.86, 1.28)
MSKCC (n = 11* patients)					
pSUV _{max}	0.30	0.965 (0.874, 0.990)	15.5	53.8	0.96 (0.71, 1.32)
pSUV _{mean}	0.13	0.920 (0.731, 0.978)	6.3	19.0	0.99 (0.87, 1.11)
pSUV _{total}	0.45	0.950 (0.825, 0.986)	23.1	89.9	0.96 (0.61, 1.51)
NCI (n = 6 patients)					
pSUV _{max}	0.28	0.921 (0.548, 0.989)	14.4	49.2	1.03 (0.77, 1.36)
pSUV _{mean}	0.13	0.826 (0.190, 0.974)	6.7	20.2	0.97 (0.85, 1.11)
pSUV _{total}	0.54	0.985 (0.895, 0.999)	27.6	115.0	0.95 (0.55, 1.63)
All Sites (n = 35 patients)					
pSUV _{max}	0.24	0.974 (0.949, 0.987)	12.0	39.5	1.00 (0.79, 1.26)
pSUV _{mean}	0.10	0.981 (0.962, 0.990)	5.3	16.0	0.99 (0.89, 1.10)
pSUV _{total}	0.36	0.989 (0.978, 0.994)	18.5	67.1	1.00 (0.70, 1.44)

*two patients received partial whole-body scans

RC = repeatability coefficient for $\alpha=0.05$ (log-transformed SUV); ICC = intraclass correlation coefficient; CI_{95%} = 95% confidence intervals; CV = coefficient of variation (log-transformed); CPD = critical percent difference; B = the ratio of the test-retest bias; LOA_{95%} = 95% limits of agreement of the ratio of test-retest measurements. B and LOA_{95%} have been back-transformed to original units.