Practical Approach for Comparative Analysis of Multilesion Molecular Imaging Using a Semiautomated Program for PET/CT

Josef J. Fox*[1], Estelle Autran-Blanc*[2], Michael J. Morris[3], Somali Gavane[1], Sadek Nehmeh[4], André Van Nuffel[5], Mithat Gönen[6], Heiko Schöder[1], John L. Humm[5], Howard I. Scher[3], and Steven M. Larson[1]

[1]Department of Radiology, Memorial Sloan-Kettering Cancer Center, New York, New York; [2]Department of Nuclear Medicine, Cochin Hospital, Paris, France; [3]Genitourinary Oncology Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York; [4]Department of Medical Physics, Memorial Sloan-Kettering Cancer Center, New York, New York; [5]General Electric Healthcare, Brussels, Belgium; and [6]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York

We propose a standardized approach to quantitative molecular imaging (MI) in cancer patients with multiple lesions. **Methods:** Twenty patients with castration-resistant prostate cancer underwent ^{18}F-FDG and ^{18}F-16β-fluoro-5-dihydrotestosterone (^{18}F-FDHT) PET/CT scans. Using a 5-point confidence scale, 2 readers interpreted coregistered scan sets on a workstation. Two hundred three sites per scan (specified in a lexicon) were reviewed. ^{18}F-FDG–positive lesion bookmarks were propagated onto ^{18}F-FDHT studies and then manually accepted or rejected. Discordance-positive ^{18}F-FDHT lesions were similarly bookmarked. Lesional SUV_{max} was recorded. Tracer- and tissue-specific background correction factors were calculated via receiver-operating-characteristic analysis of 65 scan sets. **Results:** Readers agreed on more than 99% of ^{18}F-FDG– and ^{18}F-FDHT–negative sites. Positive-site agreement was 83% and 85%, respectively. Consensus-lesion maximum standardized uptake value (SUV_{max}) was highly reproducible (concordance correlation coefficient $>$ 0.98). Receiver-operating-characteristic curves yielded 4 correction factors (SUV_{max} 1.8–2.6). A novel scatterplot (Larson-Fox-Gonen plot) depicted tumor burden and change in SUV_{max} for response assessments. **Conclusion:** Multilesion molecular imaging is optimized with a 5-step approach incorporating a confidence scale, site lexicon, semiautomated PET software, background correction, and Larson-Fox-Gonen graphing.

Key Words: molecular imaging; PET/CT; ^{18}F-FDG; ^{18}F-FDHT; semi-automated

J Nucl Med 2011; 52:1727–1732
DOI: 10.2967/jnumed.111.089326

Molecular imaging (MI) with ^{18}F-FDG PET is widely used for assessing the effect of treatment on tumors (*1,2*). Numerous additional agents for imaging the hallmarks of cancer, such as rapid proliferation, apoptosis, amino acid synthesis, hypoxia, and more specific molecules expressed on tumors, are under development (*3,4*). Evaluation of these potential imaging biomarkers requires a reproducible and expeditious system to identify disease, quantify metabolic activity, and follow the course of the lesion over time, particularly in patients with a multitude of lesions. With these requirements in mind, our group developed a PET image segmentation technique based on adaptive thresholding (*5*). This method produced precise volume measurements and eliminated the subjectivity of manual contouring. Furthermore, a coordinate system was devised to facilitate longitudinal tumor tracking on serial ^{18}F-FDG scans and characterization of tumor heterogeneity with diverse tracers (*6*). These tools served as a foundation for semiautomated image-based PET/CT analysis programs now produced by various manufacturers. PET volume computer-assisted reading (VCAR), an application of the Advantage Workstation (GE Healthcare), is one such program that incorporates precise examination-to-examination coregistration, using the companion CT scan as a fiduciary marker, and threshold-based image segmentation. These features permit unambiguous lesion tracking and efficient analysis of large datasets, essential for streamlining pharmacodynamic and response assessments in clinical trials. In this brief communication, we report a 5-step approach intended to standardize implementation of semiautomated image analysis programs such as PET VCAR, thereby facilitating the successful codevelopment of novel MI biomarkers and therapies.

MATERIALS AND METHODS

To further develop and validate our approach, we chose a clinical situation in which 2 radiotracers were used to image a group of patients with multiple metastatic bone or soft-tissue lesions. In the context of an institutional review board–approved protocol, 65 consecutive patients with progressive castration-resistant prostate cancer underwent paired ^{18}F-FDG and ^{18}F-16β-fluoro-5-dihydrotestosterone (^{18}F-FDHT) PET/CT scans within a 24-h period. ^{18}F-FDG scans were acquired approximately 60 min after about 370 MBq of ^{18}F-FDG had been injected. ^{18}F-FDHT scans were acquired approx-

Received Mar. 10, 2011; revision accepted May 18, 2011.
For correspondence or reprints contact: Josef J. Fox, Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021.
E-mail: foxj@mskcc.org
*Contributed equally to this work.
Published online Oct. 7, 2011.

imately 40 min after about 333 MBq of ^{18}F-FDHT had been injected (*7*). Patients were imaged from skull base to upper thighs on a Discovery STE PET/CT scanner (GE Healthcare). Reconstructed images were loaded onto a PET VCAR workstation. Two experienced readers interpreted a randomized subset of 20 scan sets in a masked fashion. The reader reviewed 203 sites per scan, prespecified in an anatomic lexicon (Supplemental Table 1). Scans were first interpreted qualitatively on a 5-point confidence scale for the absence or presence of malignancy (0 = definitely negative, 1 = probably negative, 2 = equivocal, 3 = probably positive, and 4 = definitely positive). Foci of activity visually higher than local background and not explained by physiologic or benign processes were considered positive. Sites rated 0–2 were recorded as negative; sites rated 3–4 were recorded as positive. All discrete lesions within positive sites were segmented with the threshold-based isocontour tool set at a default of 42% from maximum standardized uptake value (SUV_{max}) (*5*). Coalescing lesions that could not be clearly separated were segmented as 1 lesion. Lesions occupying 2 contiguous sites were considered distinct lesions. Paired ^{18}F-FDG and ^{18}F-FDHT scans were automatically coregistered by PET VCAR using a system of coarse and fine adjustments based on the characteristics of the bone and soft tissue on the companion CT scan. Bookmarked regions of interest for ^{18}F-FDG lesions were automatically duplicated and propagated onto the coregistered ^{18}F-FDHT images. Propagated bookmarks were accepted or rejected using the confidence scale, and regions of interest were manually adjusted by the reader, as needed. Discordance-positive ^{18}F-FDHT lesions were segmented in a similar fashion. SUV_{max} (body weight) was obtained for every lesion and cataloged site by site. The results of the 2 readers were compared on a per-site and per-lesion basis to determine interobserver variability. Reproducibility of SUV_{max} measurements for consensus lesions was assessed with Bland–Altman plots and calculation of concordance correlation coefficient (CCC) (*8*). SUV_{max} reproducibility was further analyzed after correction for background activity.

With the rationale that lesional metabolic activity is composed of tracer bound in tumor and unbound tracer in stroma, we aimed to subtract the contribution of stromal signal from the measured SUV_{max}. We hypothesized that establishing a population-based background would provide an approximate measure of stromal signal and could also serve as a threshold for better discrimination between benign and malignant uptake. To determine this value, all 65 scan sets were interpreted by consensus. In addition to lesional uptake, SUV_{max} of background activity was recorded. For bone background, a region of interest was placed in the posterior iliac crest or other uninvolved bone if the iliac crest harbored tumor. For soft tissue, a region of interest was placed in gluteal muscle (chosen for ease of measurement). Tracer- and tissue-specific receiver-operator-characteristic curves were constructed by plotting background SUVs against lesion SUVs. The point on the curve closest to perfect classification (0,1) was chosen as the background/threshold SUV_{max}. The 4 resulting values were applied as a correction factor for all segmented lesions within the respective tracer and tissue categories: (lesion SUV_{max}) – (background SUV_{max}). Background-corrected lesions with an SUV_{max} of 0 or less were reassigned as PET-negative.

RESULTS

For the interobserver analysis of ^{18}F-FDG scans, 3,852 (94.9%) of 4,060 sites were classified as negative by both readers, 173 (4.2%) as positive by both, and 35 (0.9%) as positive by only one. For the 4,060 ^{18}F-FDHT sites, the respective classifications were 3,838 (94.5%), 189 (4.7%), and 33 (0.8%). This translates to 83.2% (173/208) agreement for positive ^{18}F-FDG sites and 85.1% (189/222) agreement for positive ^{18}F-FDHT sites. As several positive sites contained more than 1 discrete lesion, the number of recorded lesions was greater than the number of positive sites. The 2 readers agreed on 80.8% (194/240) of all recorded ^{18}F-FDG lesions and 78.7% (211/268) of all ^{18}F-FDHT lesions. SUV_{max} measurements for these consensus lesions were highly concordant: for ^{18}F-FDG, CCC was 0.994 (95% confidence interval, 0.992–0.996); for ^{18}F-FDHT, CCC was 0.981 (95% confidence interval, 0.976–0.986). Consensus lesion SUV_{max} reproducibility is depicted graphically with Bland–Altman plots in Figure 1.

The background analysis yielded 4 separate values with an SUV_{max} of 1.8–2.6 (Tables 1 and 2). Interobserver reproducibility for background-corrected consensus lesion SUVs was nearly identical to the precorrection scenario: for background-corrected ^{18}F-FDG, CCC was 0.994

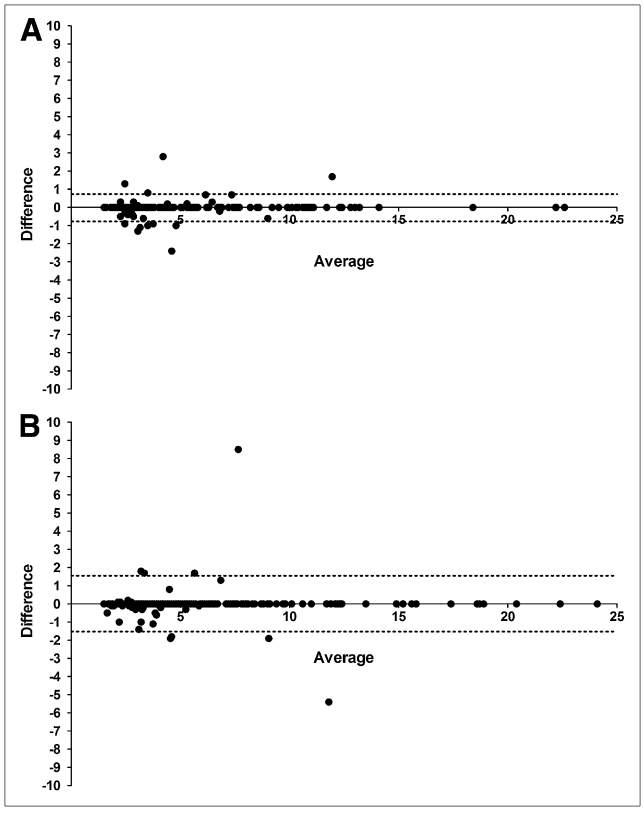

FIGURE 1. Bland–Altman plots for ^{18}F-FDG (A) and ^{18}F-FDHT (B) demonstrating high reproducibility of interobserver consensus lesion SUV_{max} measurements. For ^{18}F-FDG, bias is 0.016 and 95% limits of agreement are −0.77 to 0.74. For ^{18}F-FDHT, bias is −0.015 and 95% limits of agreement are −1.56 to 1.53. Slightly wider confidence limits for ^{18}F-FDHT indicate higher variability in measurements, largely due to 2 outlying lesions, both of which can be seen on plot.

TABLE 1
Lesion and Background Data from 65 ^{18}F-FDG and ^{18}F-FDHT Scan Sets Used in Receiver-Operator-Characteristic Curve Background Analysis

Tracer	Site	*n*	Mean SUV_{max}	SD	Minimum SUV_{max}	Maximum SUV_{max}
Bone ^{18}F-FDG	Lesion	1,079	5.6	5.4	0.6	47.2
	Background	65	1.4	0.3	0.8	2.3
Bone ^{18}F-FDHT	Lesion	1,014	6.3	3.9	1.0	28.5
	Background	65	1.9	0.5	0.8	2.8
Soft ^{18}F-FDG	Lesion	225	5.6	3.6	0.8	22.6
	Background	50	1.2	0.4	0.5	2.0
Soft ^{18}F-FDHT	Lesion	196	8.1	4.5	1.5	20.5
	Background	50	1.4	0.5	0.5	3.0

(95% confidence interval, 0.993–0.996); for background-corrected ^{18}F-FDHT, CCC was 0.979 (95% CI, 0.973–0.985).

Representative response data for 2 patients were graphed on a novel scatterplot designed to facilitate multilesion response assessments. We refer to this graph here as the Larson-Fox-Gonen (LFG) plot (Figures 2 and 3).

DISCUSSION

MI offers the potential for improved detection of disease and quantitation of alterations in molecular targets. In the context of clinical trials, MI can assist in determining the proof of mechanism for an experimental drug and, separately, treatment efficacy. A variety of PET-based methods has been proposed for quantitating treatment response, including the recently proposed PERCIST (PET Response Criteria in Solid Tumors) criteria (*9*). These methods generally recommend assessment of only a selected number of target lesions, modeled after structure-based criteria such as RECIST 1.1 (Response Evaluation Criteria in Solid Tumors) (*10*). However, RECIST-type criteria are largely based on pragmatism, with limited supporting evidence (*11–14*). In patients with many metastatic lesions, this reductive approach risks the overlooking of key lesions that are outliers in terms of behavior and are potentially responsible for a poor patient outcome. The introduction of semiautomated data analysis programs such as PET VCAR can account for all lesions in outcome assessments, thus helping elucidate optimal parameters of response. In addition, this platform can be used to compare the uptake of multiple tracers in various lesions and to monitor similarities and differences in response to treatment.

Our standardized approach to comparative analysis of total-lesion MI builds on the capabilities of these semiautomated systems (Fig. 4).

Step 1: 5-Point Confidence Scale Is Used for Initial Qualitative Assessment

Overall, there was high interobserver agreement (>99%) with respect to qualitatively classifying the 4,060 anatomic sites as negative or positive for both ^{18}F-FDG and ^{18}F-FDHT scans. The agreement rate fell to roughly 84% when only positive sites were the focus and to 80% when all recorded lesions were considered (some sites contained multiple lesions). An ordinal confidence scale mitigates, but cannot completely resolve, the inherent and unavoidable subjectivity of diagnostic imaging interpretation, irrespective of the workstation used. MI with PET is arguably more prone to interobserver variability than conventional structural imaging. Nevertheless, a recent paper looking at CT interpretation reported major interobserver disagreements in 26%–32% of cases (*15*), supporting the notion that disagreement in qualitative interpretation is unavoidable. As a solution to this problem, we recommend that preliminary training sessions or consensus readouts should be integrated into imaging protocols.

TABLE 2
Receiver-Operator-Characteristic Curve Analyses of Lesion and Background SUV_{max} Data in Table 1

Tracer	SUV_{max} Threshold	Specificity	Sensitivity	Distance from perfect marker
Bone ^{18}F-FDG	2.0	99.53%	84.80%	0.153
Bone ^{18}F-FDHT	2.6	99.13%	90.04%	0.100
Soft ^{18}F-FDG	1.8	94.03%	94.12%	0.084
Soft ^{18}F-FDHT	2.3	99.47%	96.94%	0.031

Four distinct tracer- and tissue-dependent threshold values were obtained for optimal discrimination between benign and malignant uptake. For any given threshold, tradeoff exists between sensitivity and specificity. When distance from perfect marker was similar for more than one value, we opted for greater specificity at expense of lower sensitivity.

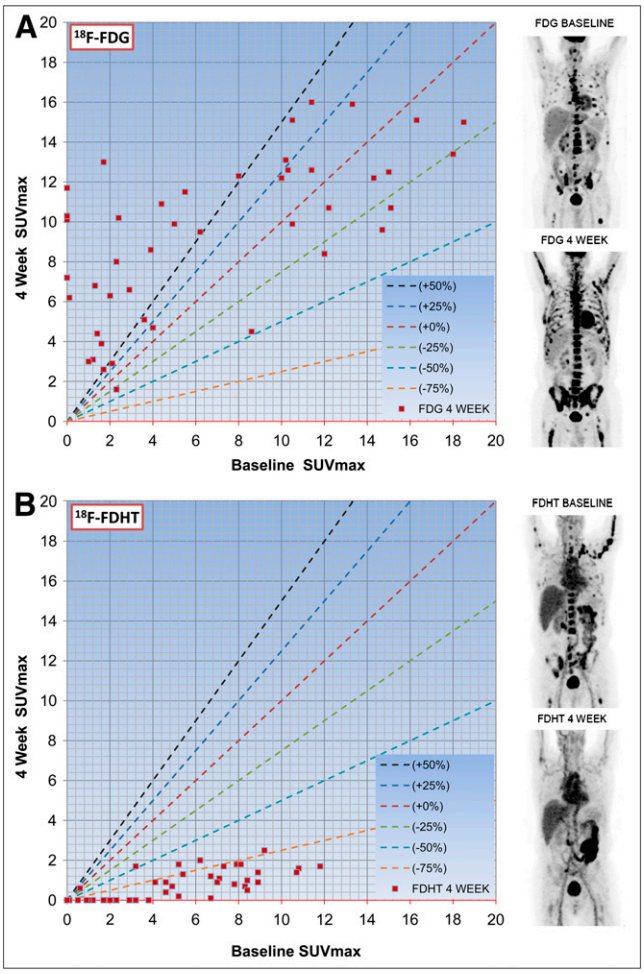

FIGURE 2. Representative ^{18}F-FDG (A) and ^{18}F-FDHT (B) LFG plots in nonresponding castration-resistant prostate cancer patient receiving androgen receptor–targeted therapy. Identity line indicates no change in SUV between baseline and follow-up (change in SUV_{max}, 0%). Rays around identity line indicate various levels of percentage change. New lesions fall on *y*-axis when value of zero for baseline SUV_{max} is imputed. In this example, total-lesion (n = 51 at baseline) ^{18}F-FDG and ^{18}F-FDHT background-corrected SUV_{max} data are plotted, demonstrating marked hypermetabolism at baseline and metabolic progression at 4 wk (increase in ^{18}F-FDG uptake > 50% for several lesions, as well as several new lesions). ^{18}F-FDHT plot shows concomitant suppression of ^{18}F-FDHT uptake (>75% reduction in most lesions), despite apparent ^{18}F-FDG progression. Corresponding maximum-intensity-projection PET images, at baseline and after 4 wk of therapy, are found to right of plots.

Step 2: Standardized Lexicon for Lesion Nomenclature Is Adopted

A lexicon minimizes ambiguities in lesion assignment, particularly in the context of a total-lesion cataloging effort. A lexicon also facilitates correlation with more conventional imaging modalities such as bone scanning, CT, and MRI.

Step 3: Scans Are Analyzed Semiautomatically

First, positive lesions are bookmarked with a threshold-based segmentation algorithm. An isocontour tool clearly defines the 3-dimensional borders of the lesion, ensuring that the voxel containing the SUV_{max} is within the confines of the lesion. Second, PET/CT studies are automatically coregistered. Coregistration enables automatic propagation of lesion bookmarks and facilitates unambiguous lesion tracking. In contrast to the qualitative assessment (Step 1), quantitative agreement was excellent, reflected by high SUV_{max} reproducibility for consensus lesions (CCC > 0.98 for both ^{18}F-FDG and ^{18}F-FDHT). These results are at least similar to the interobserver reproducibility of SUV_{max} measurements obtained on a standard workstation (intraclass correlation coefficient, 0.93) (*16*).

Step 4: Positive Lesions Are Corrected for Background Activity

Background correction in this context serves 2 purposes: to eliminate the contribution of signal from unbound tracer

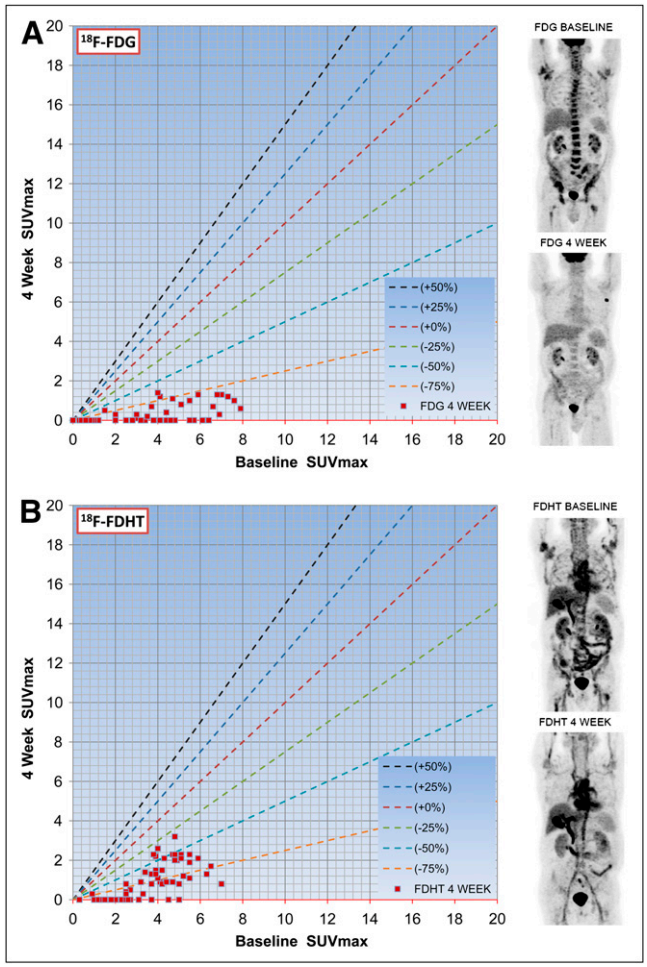

FIGURE 3. Representative ^{18}F-FDG (A) and ^{18}F-FDHT (B) LFG plots in responding castration-resistant prostate cancer patient receiving androgen receptor–targeted therapy. Total-lesion (n = 61) ^{18}F-FDG and ^{18}F-FDHT background-corrected SUV_{max} data are graphed, depicting favorable metabolic response (>75% reduction in ^{18}F-FDG uptake for most lesions) and concomitant suppression of ^{18}F-FDHT uptake (>50% in most lesions). Corresponding maximum-intensity-projection images, at baseline and after 4 wk of therapy, are found to right of plots. Focal activity in left axilla on 4-wk ^{18}F-FDG scan represents artifact (i.e., benign nodal uptake related to radiotracer injection).

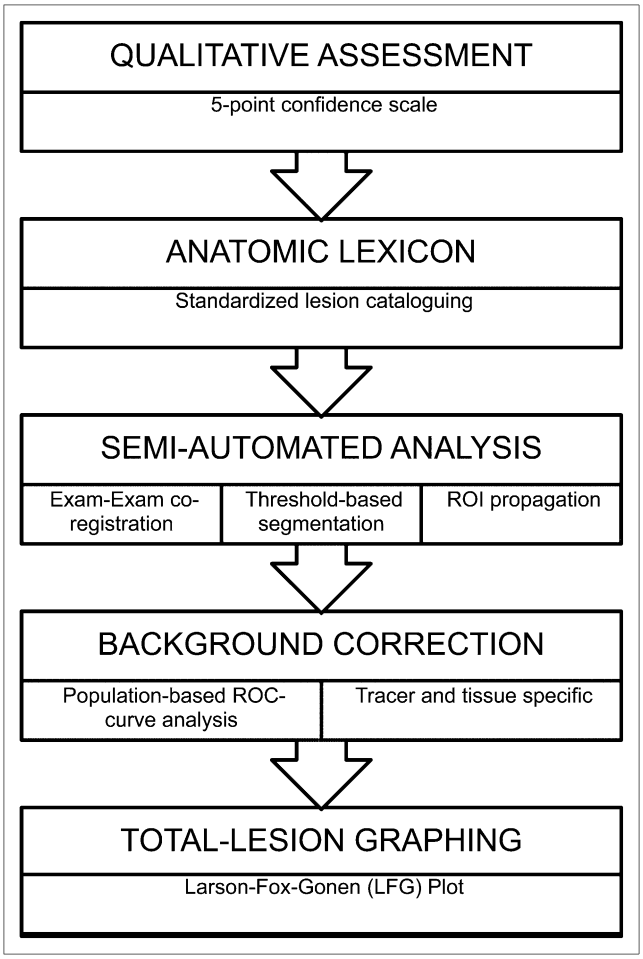

FIGURE 4. Summary diagram of 5-step approach to comparative analysis of total-lesion MI.

in stroma and to optimally discriminate between benignity and malignancy. We used a population-based receiver-operator-characteristic curve analysis to establish a standard background level, which was then applied as a correction factor. Four separate thresholds were calculated to account for the distinct properties of each tracer in bone and soft tissue. When applicable, we opted for greater specificity over sensitivity, given the plethora of lesions.

Step 5: SUV Data Are Graphed on LFG Plot

An LFG plot allows for representation of large amounts of comparison data while clearly depicting absolute and percentage change in SUV_{max} for individual lesions, new lesions, and trends for the total-lesion burden. Individual lesions with aberrant behavior are easily detected.

A limitation of the study is the lack of a gold standard comparator to confirm the accuracy of the segmented lesions. Nevertheless, the purpose of this brief communication is not to present specific outcome data for ^{18}F-FDG or ^{18}F-FDHT in castration-resistant prostate cancer. Rather, our goal is to describe a standardized and practical approach for multilesion assessments, as an aid for future work with MI. We intend to further validate the receiver-operator-characteristic–based background analysis in the context of pending pharmacodynamic and response assessments, as well as with tissue correlation, when available.

CONCLUSION

We have described our approach to the challenging problem of MI-based quantitative analysis of multiple lesions in individual patients or patient populations. We propose that this type of analysis benefits from semi-automated software such as PET VCAR, which allows for unambiguous lesion tracking and reproducible quantitative assessment. A novel summary plot, the LFG plot, was developed to visualize data in a manner that is intuitive and permits easy assessment of treatment response. In future work, we plan to compare the largest group of lesions with smaller subsets of target lesions to determine the optimal number needed for prediction of clinical endpoints such as overall survival. Ultimately, we propose that this biologically sound approach will lead to the qualification of robust imaging biomarkers.

DISCLOSURE STATEMENT

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.

ACKNOWLEDGMENTS

Support for this research came from the MSKCC Center for Molecular Imaging in Cancer from the National Cancer Institute (grant P50-CA086438) and from the Memorial Sloan-Kettering Cancer Center Specialized Program of Research Excellence (SPORE) (grant in prostate cancer P50-CA92629). No other potential conflict of interest relevant to this article was reported.

REFERENCES

1. Kelloff GJ, Hoffman JM, Johnson B, et al. Progress and promise of FDG-PET imaging for cancer patient management and oncologic drug development. *Clin Cancer Res.* 2005;11:2785–2808.
2. 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.
3. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* 2000;100:57–70.
4. Larson SM, Schoder H. New PET tracers for evaluation of solid tumor response to therapy. *Q J Nucl Med Mol Imaging.* 2009;53:158–166.
5. Erdi YE, Mawlawi O, Larson SM, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. *Cancer.* 1997;80 (suppl)2505–2509.
6. Erdi YE, Srivastava NC, Humm JL, Larson SM. A coordinate system for tumor identification in positron emission tomography (PET) imaging. *Clin Positron Imaging.* 2000;3:131–136.
7. Larson SM, Morris M, Gunther I, et al. Tumor localization of 16β-^{18}F-fluoro-5α-dihydrotestosterone versus ^{18}F-FDG in patients with progressive, metastatic prostate cancer. *J Nucl Med.* 2004;45:366–373.
8. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45:255–268.

9. 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.
10. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;45:228–247.
11. Schwartz LH, Mazumdar M, Brown W, Smith A, Panicek DM. Variability in response assessment in solid tumors: effect of number of lesions chosen for measurement. *Clin Cancer Res.* 2003;9:4318–4323.
12. Hillman SL, An MW, O'Connell MJ, et al. Evaluation of the optimal number of lesions needed for tumor evaluation using the response evaluation criteria in solid tumors: a North Central Cancer Treatment Group investigation. *J Clin Oncol.* 2009;27:3205–3210.
13. Darkeh MH, Suzuki C, Torkzad MR. The minimum number of target lesions that need to be measured to be representative of the total number of target lesions (according to RECIST). *Br J Radiol.* 2009;82:681–686.
14. Moskowitz CS, Jia X, Schwartz LH, Gonen M. A simulation study to evaluate the impact of the number of lesions measured on response assessment. *Eur J Cancer.* 2009;45:300–310.
15. Abujudeh HH, Boland GW, Kaewlai R, et al. Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists. *Eur Radiol.* 2010;20:1952–1957.
16. Jacene HA, Leboulleux S, Baba S, et al. Assessment of interobserver reproducibility in quantitative ^{18}F-FDG PET and CT measurements of tumor response to therapy. *J Nucl Med.* 2009;50:1760–1769.