Need for Standardization of ^{18}F-FDG PET/CT for Treatment Response Assessments

Ronald Boellaard

VU University Medical Center, Amsterdam, The Netherlands

Many factors affect standardized uptake values (SUVs) in ^{18}F-FDG PET/CT. The use of the SUV from a single PET scan in multicenter studies requires the standardization of ^{18}F-FDG PET/CT procedures. In the context of treatment response assessments (repeated PET scans), many factors may seem to have minor effects on percentage changes in SUVs, provided that imaging procedures are executed in a consistent manner for each subject. However, the use of ^{18}F-FDG PET/CT in a nonstandardized manner will result in unknown biases and reproducibilities of SUVs and SUV-based response measures. This article provides an overview of the need for standardization in relation to the specific use of SUVs and SUV changes in studies of treatment response assessments.

Key Words: ^{18}F-FDG; PET/CT; standardization; response; harmonization

J Nucl Med 2011; 52:93S–100S
DOI: 10.2967/jnumed.110.085662

Visual inspection of ^{18}F-FDG PET whole-body studies is important for diagnosis and staging (*1*). However, quantitative PET is increasingly being recognized as an important tool for prognosis and response monitoring (*2–5*). For the quantification of ^{18}F-FDG PET studies, various methods, which differ with regard to the complexity of data collection and (mathematic) analysis, have been described; these include tumor-to-background ratios, standardized uptake values (SUVs), and full kinetic analysis (*6,7*).

Semiquantitative analysis by means of SUVs is clinically feasible because an SUV is available in every clinically obtained whole-body scan. It is a simple index for glucose metabolism and can be obtained with good reliability, provided that ^{18}F-FDG PET/CT studies (including proper calibration procedures) are acquired in a standardized manner. A recent review of existing guidelines and factors affecting SUV results (*8*) demonstrated how to obtain accurate and reproducible SUV results in single-center and multicenter settings.

Measurements of glucose metabolism are used to differentiate between benign and malignant lesions, to try to provide relevant prognostic information at presentation—beyond TNM staging (*9*), and to predict or evaluate therapy outcomes. In some situations, a single SUV measurement suffices (i.e., an absolute SUV); in other situations, serial measurements are used (i.e., relative or percentage changes in SUVs). Guidelines have emphasized the procedures used to obtain absolute SUV data and less often have addressed the level of standardization needed for serial measurements. However, in a treatment response setting, tumor tracer uptake is measured with serial scans, that is, before, during, and after treatment (*5*).

Changes in uptake or SUVs can be used as a quantitative index for treatment responses. This application was recognized by the European Organization for Research and Treatment of Cancer (EORTC) in 1999, when it published recommendations for the use of ^{18}F-FDG PET/CT for treatment response assessments (*4,10*). Typically, relative or percentage changes in SUVs are considered to be an index for drug efficacy or a clinical response (*4,10*). Alternatively, the residual SUV, that is, the SUV from a single scan during or after therapy, may have predictive value (*11,12*). Many factors, such as the region of interest or the applied tracer uptake period, may seem to have relatively small effects on observed percentage changes in SUVs, provided that the same methodology is used during all scans of an individual patient (*13*). However, this is not the case when an SUV is used as a prognostic factor (stratification) or predictive factor (residual SUV), mainly because of the spatial resolution dependence of the SUV.

The large variability in currently applied PET procedures and methodologies across institutions emphasizes the need for the standardization of PET/CT examinations, as was recently shown in 2 important surveys (*14,15*). Moreover, Velasquez et al. (*16*) investigated test-retest repeatability in a multicenter setting, demonstrating deficiencies in the quality of PET/CT examinations. About 25% of all test-retest studies could not be analyzed properly because of technical issues or the use of "off-protocol" imaging procedures (e.g., uptake periods outside the designated interval). However, even after censoring of studies that did not meet

Received Feb. 14, 2011; revision accepted Sep. 15, 2011.
For correspondence or reprints contact: Ronald Boellaard, Department of Nuclear Medicine and PET Research, VU University Medical Center, Amsterdam, The Netherlands.
E-mail: r.boellaard@vumc.nl
COPYRIGHT © 2011 by the Society of Nuclear Medicine, Inc.

performance criteria, considerable differences in test-retest repeatability performance remained at various sites.

It is therefore beyond any doubt that standardization is needed for the use of quantitative ^{18}F-FDG PET/CT as an imaging biomarker. However, the extent to which standardization is needed and whether standardization is feasible for certain uses of quantitative ^{18}F-FDG PET/CT are not fully known. The purpose of this article is to review the need for the standardization of ^{18}F-FDG PET/CT in studies of treatment response assessments in multicenter settings. Moreover, the degree of standardization needed in relation to specific aspects of response criteria is addressed.

DEFINITIONS

A few definitions are provided here to facilitate discussion of the need for standardization. In this article, "execution" of an ^{18}F-FDG PET/CT study refers to all steps involved in obtaining quantitative uptake measures (including preparation of a patient, PET/CT acquisition, image reconstruction, data analysis, and PET/CT system calibration procedures).

As stated earlier, apart from standardization itself, the level of standardization needed is yet to be determined. In this article, a distinction is made between "minimal" and "harmonizing" standards.

Minimal performance standards require that sites, studies, or users meet a minimal threshold, but any performance above or beyond this threshold is considered to be sufficient. These standards typically do not primarily aim to reduce variability among users, studies, scanners, and imaging sites but set lower limits for performance or quality. An example of a minimal performance standard is a recommendation that the ^{18}F-FDG uptake period should be at least 60 min to ensure sufficient uptake in the tumor and contrast with the background, but the use of longer uptake periods (e.g., 90 min and 120 min) is allowed.

Harmonizing standards aim to minimize variability among subjects, studies, scanners, and sites. A harmonizing standard may imply that performance needs to be within a certain bandwidth (lower and upper thresholds). Performance within this bandwidth aims to reduce intersubject and interinstitution variability but may not necessarily provide the best possible performance or result for individual imaging systems or institutions. An example, taken from the ongoing work of the ^{18}F-FDG PET/CT UPICT Protocol Writing Committee (*17*), is the recommendation that the ^{18}F-FDG uptake period should be at least 60 min (lower limit) but should not exceed 75 min (upper limit).

Many guidelines contain a mixture of these standards. For some items, only minimal thresholds can be given (e.g., if calibration should be better than 10%, then it is clear that there is no lower limit), but for others, recommendations for a specific use of an SUV should be made (e.g., if the SUV itself rather than relative changes in SUVs is used in a multicenter setting, then the uptake period should be defined with harmonizing thresholds).

Additionally, in this article, a distinction is made between intrasubject (or within-subject) and intersubject (or between-subjects) standardization.

Intrasubject (or within-subject) standardization means the consistent execution of all (steps of) ^{18}F-FDG PET/CT examinations of a single subject at a single site. This guideline automatically implies that different ^{18}F-FDG PET/CT procedures and methods may be applied for other subjects within the same institution or trial. The only requirement is that the same procedures, scanners, and data analysis methods are applied for all PET/CT studies of a single subject (*13*).

Intersubject (or between-subjects) standardization means that all scans performed with all scanners at all sites are executed in the same manner. Therefore, this methodology is a harmonizing standard. This form of standardization includes matching of image quality, quantification, and reconstructed image resolution of all PET/CT systems at all sites and for all examinations (*18*).

USE OF ^{18}F-FDG PET/CT IN RESPONSE ASSESSMENT STUDIES

^{18}F-FDG PET/CT studies are used for various applications in both clinical practice and clinical trials. Such studies include both visual interpretation and quantitative reading on the basis of SUVs or relative changes in SUVs. Various uses of ^{18}F-FDG PET/CT in response assessment (clinical) studies or trials are summarized here.

Visual Image Interpretation

This article focuses on the use of quantitative ^{18}F-FDG PET/CT studies. However, visual interpretation is of utmost importance both for diagnostic purposes and in trials. For example, ^{18}F-FDG PET/CT studies are used clinically for TNM staging. In clinical trials, visual reading can be important for assessing the eligibility of patients to participate in a trial or for assigning subjects in a trial arm (stratification). Finally, an assessment of disease progression may be based on the visual assessment of new lesions, and responses may be determined by visual scoring of (changes in) tracer uptake. When studies are performed in a (quantitative) standardized manner, fixed color scales can be applied to all longitudinal scans, allowing for a more accurate visual assessment of changes in tumor tracer uptake.

Use of SUV from Single PET/CT Study

The baseline or residual SUV in response assessment (clinical) studies or trials can be used for target lesion selection, as a predictive factor, and as a prognostic factor.

Target Lesion Selection. Recently, it was suggested (*4*) that the eligibility of a lesion for response measurement should be determined on the basis of minimal uptake or metabolic volume. The idea behind setting a minimal uptake criterion for lesion selection is that lesions showing low ^{18}F-FDG uptake may not be able to show a decrease in uptake or that SUVs (or changes in SUVs) in small lesions

or lesions with a low avidity for ^{18}F-FDG cannot be determined with sufficient accuracy and precision (e.g., the SUVs do not exceed background and noise levels). It was suggested (*4*) that uptake in such lesions should be considered relative to normal liver uptake (e.g., twice the normal liver or blood pool uptake, or 1.5 times normal liver uptake plus 2 times the SD for "liver noise," assuming that hepatic uptake is constant among subjects and during treatment).

Predictive Factor. A few reports have indicated that the residual SUV after or early during therapy may have a predictive value (*12*); that is, it may be used to assess or predict treatment responses. The reasoning is that (relative) change alone does not account for the possible impact of the absolute baseline value (*19*). In this scenario, the SUV is used in an absolute (or single PET/CT study) manner.

Prognostic Factor. If validated, the SUV may be used to select or assign subjects to trial arms and may help to achieve a balanced (or even intended unbalanced) study design; that is, the SUV may be considered to be a prognostic factor (*2,20–22*). In this scenario, the SUV itself (rather than relative changes in SUVs) is used.

Use of Percentage Change in Uptake for Response Assessments

Treatment responses are often measured by use of percentage changes in SUVs. In 1999, Young et al. (*10*) described the EORTC guideline for the measurement of a treatment response. This guideline provided criteria for classifying patients as having metabolically progressive or stable disease as well as partial and complete responses. This classification relies on percentage or relative changes in SUVs (or the change in the rate of metabolism of glucose) in the same lesion(s) in a patient across all longitudinal studies of that patient.

Recently, Wahl et al. described the PERCIST criteria (*4*) for classifying responses by use of percentage changes in SUVs in the "hottest" lesion(s) per scan; that is, in each scan of the same patient in a longitudinal study, the lesion(s) with the highest level of uptake is identified, and the change in SUVs for the 2 (sets of) lesions is measured. The hottest lesions per scan are not necessarily the same over time. In addition to measurement of the percentage changes in SUVs, the PERCIST criteria recommend consideration of a minimal change of 0.8 unit of SUV normalized to lean body mass.

This summary highlights only a few key differences between the PERCIST and EORTC response criteria that are relevant in the context of the present article.

STANDARDIZATION FOR ^{18}F-FDG PET/CT–BASED RESPONSE ASSESSMENTS

SUVs are affected by technical, physical, and biologic factors (*5,8*). This article focuses on the need to standardize these factors in the setting of response assessments in multicenter studies rather than on replicating an extensive discussion of these factors (*8*).

On the basis of the earlier overview of the uses of PET/CT in treatment response studies, there is an obvious need for minimal performance standards for all technical factors, such as scanner calibration, clock synchronization, residual activities in syringes and lines after administration, and paravenous administration. Any systematic error, drift, or random variability in the technical factors above or beyond a minimal standard will result in (unnecessary) systematic errors and increased variability in both SUVs and observed changes in SUVs over time. Suboptimal performance of PET/CT systems will also affect the quality of visual interpretation.

The biologic (or physiologic) factors affecting SUVs are plasma glucose levels, uptake period, patient motion or breathing, patient comfort during the uptake period, and uptake due to an inflammatory reaction. For most of these factors, clear recommendations have been provided, as they directly affect SUVs or image interpretation. Obviously, there are no upper limits for interpretation, preparation of a patient during the period before administration, and patient comfort during the uptake period. Therefore, recommendations provided in various guidelines should be considered to be minimal standards and should be followed by all sites and for all subjects participating in a multicenter study. However, when relative changes in SUVs are considered to be an index for metabolic responses, it may be argued that some factors only need to be consistent or equal across longitudinal scans of the same subject (intrasubject standardization).

The third category of factors affecting SUVs and therefore ^{18}F-FDG PET–based treatment response assessments includes applied imaging and data analysis methods and settings. Imaging parameters, such as scan duration per bed position, acquisition mode, ^{18}F-FDG dose, and reconstruction methods and settings, directly affect image quality and quantification (*5*). Poor image quality results in an upward bias of the SUV that increases with lesion size, especially when the maximum uptake is based on a single voxel (*23*). Moreover, as indicated by Wahl et al. (*4*), any new ^{18}F-FDG–avid lesion seen on a PET image during or after treatment should be interpreted as metabolic progression. Image quality with respect to lesion detection is therefore important.

Two recent surveys on the multicenter variability of ^{18}F-FDG PET/CT methodology identified several factors for which substantial variability between imaging centers was shown (*14,15*). The following factors were found to be highly variable in both surveys:

- Uptake period (*24*), ranging from 45 to 90 min in the United States and from 20 to 90 min worldwide (*14,15*)
- Procedures for patients with diabetes and measurement of and dealing with plasma glucose levels (*25–28*)

- ^{18}F-FDG dose and imaging time per bed, both of which affect image quality and quantification
- Image reconstruction, processing, and data analysis

Biologic Factors

Uptake Period. Figure 1 shows data taken from 14 dynamic ^{18}F-FDG PET studies (*29*); the data show percentage changes in SUVs for 2 tracer uptake periods. Figure 1A shows the results obtained when the uptake times for both baseline and response scans matched exactly, and Figure 1B shows the results obtained when there was a 10-min difference in the baseline and response scan uptake times. When the baseline and response scan uptake times matched, there was no significant difference in observed percentage changes in SUVs for either 37.5- or 57.5-min uptake times. When the uptake times differed by 10 min, there were significant differences in observed responses (paired t test, $P < 0.0001$). These findings illustrate that, in case of deviation from the prescribed uptake time interval, the same uptake period (rather than the prescribed uptake period) should be used for all scans of the same patient (intrapatient or within-patient consistency). In theory, imaging sites therefore could decide to apply different uptake periods for different patients. However, this approach would require careful registration of the actual uptake period at baseline and the use of exactly the same uptake period during subsequent response studies. It is debatable whether, in a busy clinic, applying the same uptake period for all patients and for all scans would be more difficult than trying to schedule or reschedule subsequent response scans using a "baseline-driven" patient-specific uptake period.

Figure 2A shows SUV changes (i.e., response SUV minus baseline SUV) for 2 uptake periods, and Figure 2B shows SUV changes obtained when there was a 10-min difference in the uptake period between baseline and response studies. No significant difference in SUV changes was found when the uptake periods were exactly the same for baseline and response studies. However, when there was a 10-min difference in the uptake period between baseline and response studies, significantly different results for absolute SUV changes were obtained (paired t test, $P < 0.0001$). Besides the effect of the uptake period on SUV changes in longitudinal scans, the SUV itself depends on the uptake period (*23*). Therefore, the use of different uptake periods for patients and at imaging sites would prohibit any use of the SUV for target lesion selection or as a prognostic or predictive (response) factor.

Plasma Glucose Levels and Patients with Diabetes. Elevated plasma glucose levels result in decreased ^{18}F-FDG uptake and hence in lower SUVs (*28,30*). Consequently, variable plasma glucose levels in longitudinal studies of the same patient will likely cause artificial SUV changes, which may impair a proper assessment of the effect of therapy on glucose metabolism. Several guidelines (*8*) have suggested that fasting for 4–6 h before the ^{18}F-FDG injection results in fairly uniform plasma glucose levels of 4–7 mmol/L (72–126 mg/dL) (*18,31*). However, these plasma glucose levels may be difficult to obtain in patients with diabetes (known or unknown). A more feasible approach could be to keep plasma glucose levels as constant as possible for all scans of the same patient. The effect of elevated but constant glucose levels on ^{18}F-FDG PET/CT response assessments is not fully known. As a consequence, trying to achieve constant plasma glucose levels across all longitudinal studies of the same patient seems to be the minimal feasible requirement; achieving plasma glucose levels of 4–7 mmol/L for all patients does not seem to be feasible from a clinical point of view, particularly for patients with diabetes. Moreover, data on the need for intersubject standardization are lacking. However, several guidelines (*8*) have recommended attempts to achieve plasma glucose levels in the reference range (4–7 mmol/L), which is feasible for patients without diabetes, and to measure (and report) plasma glucose values using well-calibrated and validated methods.

There are a few strategies for dealing with plasma glucose levels in SUV calculations, but further research is needed. An SUV could be "corrected" for differences in plasma glucose levels within and between subjects by including the measured plasma glucose level as a multiplication factor in the SUV equation. There are several concerns about whether this correction should be applied or not. This correction assumes that ^{18}F-FDG uptake is inversely proportional to the plasma glucose level. This correction may be valid for plasma glucose levels close to the normal (fasting)

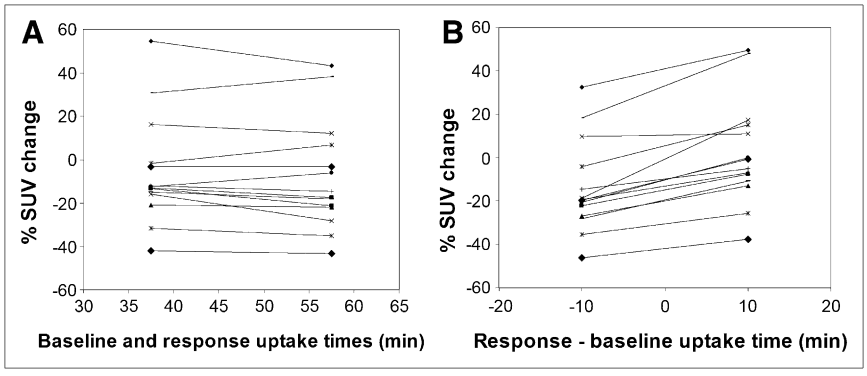

FIGURE 1. (A) Percentage changes in SUVs for 2 uptake periods. Tracer uptake times were equal in baseline and response studies. (B) Percentage changes in SUVs for 10-min mismatch in uptake time between response and baseline studies. In B, baseline uptake period was set at 48 min, and response uptake time was set at 38 min (−10 min) or 58 min (+10 min). Data were taken from dynamic ^{18}F-FDG PET studies recently described by Cheebsumon et al. (*29*). Each symbol and line represent data from single subject.

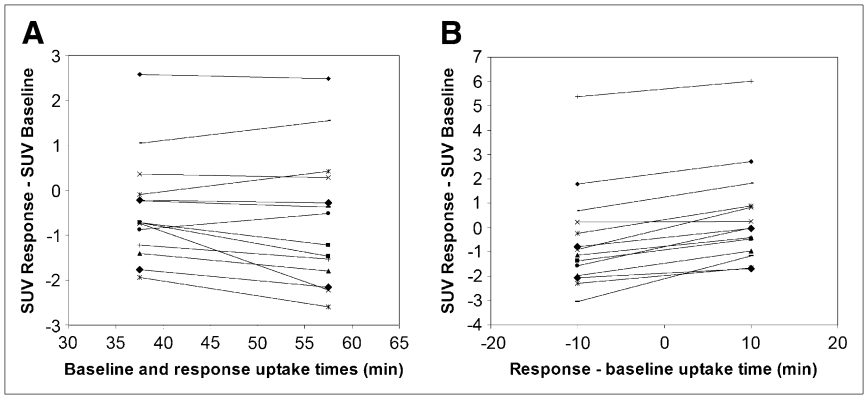

FIGURE 2. (A) Absolute SUV changes (response SUV minus baseline SUV) for 2 uptake periods. Tracer uptake times were equal in baseline and response studies. (B) Absolute SUV changes for 10-min mismatch in uptake time between response and baseline studies. In B, baseline uptake period was set at 48 min, and response uptake time was set at 38 min (−10 min) or 58 min (+10 min). Data were taken from dynamic ^{18}F-FDG PET studies recently described by Cheebsumon et al. (*29*). Each symbol and line represent data from single subject.

level, but it is unclear whether this correction is still accurate at very high levels (>11 mmol/L or >200 mg/dL). Recently, however, Wong et al. (*32*) demonstrated that correction for blood glucose levels will make the SUV a more robust outcome measure in tissues that show decreasing ^{18}F-FDG uptake with increasing blood glucose level. The introduction of an additional correction in the SUV equation will also result in additional noise in the SUV results; that is, any bias and uncertainty in the determination of plasma glucose levels will be propagated into the SUV results. Although the use of plasma glucose level corrections seems to improve test-retest variability in a single-center setting (*6,7*), such is not the case in a multicenter setting (*33*). Standards for the accuracy and precision of plasma glucose measurements therefore should be high (*34*).

In conclusion, trying to achieve constant plasma glucose levels across all longitudinal studies of the same patient, aiming to achieve a normal fasting range (4–7 mmol/L) for these levels in all patients, and properly reporting (*35*) measured plasma glucose values seem to be clinically feasible and justified. It is hoped that future studies with these goals will provide evidence or more detailed information about whether intrasubject or intersubject standardization is required.

Physical Factors: Image Quality, Resolution, and Data Analysis

A minimal standard aimed at providing PET images with a minimally acceptable spatial resolution but a sufficiently low noise level is required to ensure minimal performance of all PET/CT examinations for the detection of new lesions. However, imaging parameters (e.g., reconstruction settings) that are optimal for lesion detection may not be optimal for proper quantification. Figures 3C and 3D show SUV recovery coefficients, measured as indicated by Boellaard et al. (*18*), for scans with normal or clinically expected results (Fig. 3A) and for scans with high statistical quality (Fig. 3B). It is clear that a combination of high statistical quality and high spatial resolution is optimal for lesion detection. However, large biases may occur when reconstruction methods and settings that are more optimal for tumor detection than for quantification are applied. In this specific example,

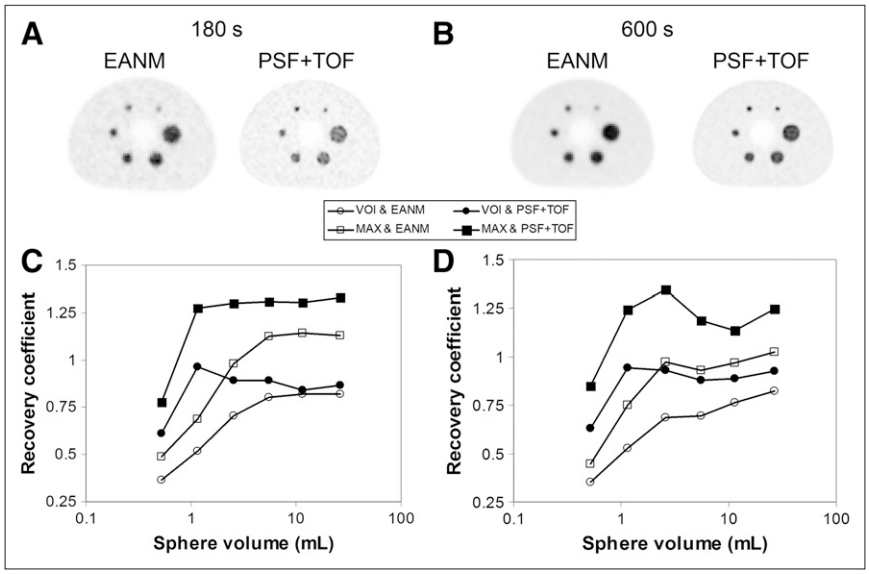

FIGURE 3. SUV recovery coefficients. Experiments were performed as described in EANM guideline for quantitative ^{18}F-FDG PET/CT studies (*18*). Data were collected by applying 180 s per bed (acquisitions with normal statistical quality) (A and C) and 600 s per bed (acquisitions with high statistical quality) (B and D). (A and B) Axial slices through modified National Electrical Manufacturers Association NU 2 image quality phantom. (C and D) SUV recovery coefficients obtained with reconstruction settings recommended by EANM (□ and ○) and point spread function plus time-of-flight (PSF+TOF) reconstruction settings (■ and ●). SUV recovery coefficients were derived from SUV_{max} (MAX) (■ and □) and mean SUV (VOI) (● and ○). Mean SUV was obtained by applying source to background adaptive 50% of maximum voxel value isocontour VOI per sphere (*18*). (Courtesy of S. Stroobants and M. Lambrechts, University of Antwerp.)

upward bias was seen mainly when the maximum SUV (SUV_{max}) was taken from images generated with a reconstruction method that incorporated the point spread function of the scanner for resolution recovery. This bias was caused by the increased upward bias of the SUV_{max} due to higher variance at the (smaller) voxel level and by the so-called "ringing" artifacts (Gibbs oscillations) introduced by the point spread function reconstruction. The latter artifacts appeared as an artificial enhancement of edges (*36*). Consequently, other image reconstruction settings (e.g., the European Association of Nuclear Medicine [EANM] guideline in Fig. 3) may be preferable for quantification.

SUV results are affected by the data analysis or volume-of-interest (VOI) method applied. The SUV_{max} is still the parameter that is most frequently used to quantify metabolism, mainly because of the lack of widely available standardized automated 3-dimensional VOI methods. The SUV_{max} results in large positive biases (overestimation) when settings that are optimized for lesion detection are used, even for images with high statistical quality; however, the SUV_{max} is more accurate and reliable when settings recommended by the EANM are used (*18*). In both scenarios, the SUV still depends on the size of the spheres, independent of the reconstruction method, settings, and VOI method used. Although the causes of this dependence on metabolic volume may be different (bias due to noise or Gibbs artifacts vs. partial-volume effects), it needs to be considered in response assessment studies. Consequently, changes in metabolic volume over time will have an additional effect (bias) on observed changes in SUVs. Therefore, image quality (noise and resolution) and data analysis strategies (VOI methods) should be closely matched when quantitative response measures are used, even for relative changes in SUVs.

A difficult decision with respect to image resolution is whether to specify minimal or harmonizing standards. Specifying a minimal threshold might ensure minimal performance for lesion detection, provided that image noise levels remain sufficiently low and sites are allowed to use the most optimal settings for their scans. However, quantification is resolution-dependent (*8,23,37*); therefore, harmonizing thresholds (i.e., indicating lower and upper limits for image resolution) are needed to match SUV results in a multicenter setting. This requirement applies to both absolute SUVs and relative changes in SUVs and is even more important with PERCIST criteria because of the selection of the hottest lesion(s) per scan, that is, potentially different lesions having different metabolic volumes. In the latter case, differences in image resolution between systems and sites might add to variability in observed responses.

Optimizing image quality for lesion detection therefore seems to be in conflict with optimizing image quality for quantification. One strategy for overcoming this problem is to use limits for image resolution and quality that are lower but that still provide acceptable diagnostic quality. Applying and accepting upper limits would also be necessary, even though certain systems and sites would be able to generate higher-resolution images. It is clear that the threshold for lower limits should not be determined by the system with the worst performance in a multicenter study, such that image quality would be determined by outdated technology. This argument is often made (for good reasons) by sites that wish to use their PET/CT systems to achieve the best possible lesion detection. The lower limit for image resolution therefore should be set to achieve minimally acceptable performance but not so high that it cannot be translated into clinical practice (in a multicenter setting). Another solution is to generate 2 sets of images for each PET/CT examination: one to provide optimal diagnostic quality and another to meet quantitative harmonizing standards for image characteristics (resolution and noise) (*38*). The second image dataset could be generated either by an additional image reconstruction process or, if the first image dataset was generated with higher standards, by an additional image processing or filtering step. Recently, Kelly and Declerck (*38*) used this strategy and demonstrated that it might be useful for harmonizing and optimizing quantitative results for different reconstruction methods and settings. However, this strategy requires that PET/CT systems be provided with acquisition/reconstruction protocols or data processing methods that allow the (additional) generation of images with characteristics that meet quantitative harmonizing standards.

Generating 2 image datasets also has the advantage of allowing the use of new technologies that enhance image quality for diagnostic purposes while facilitating the use of quantitative ^{18}F-FDG PET/CT studies in a multicenter setting. Moreover, both datasets may be analyzed quantitatively to help set standards. Still under consideration is whether to use 2 image datasets at the same time for response assessments. Imaging sites that can generate higher-quality images may detect new lesions earlier and therefore identify more subjects as being metabolically progressive (in accordance with the PERCIST criteria) than other sites. Consequently, how to deal with such potential site-dependent bias in larger multicenter studies needs to be considered.

Overview of Need for Standards

For each of the known factors, Table 1 shows the type of standardization (minimal or harmonizing and intrasubject or intersubject) that is recommended for ^{18}F-FDG PET/CT–based treatment assessments. In this article I have attempted to indicate the possible effects of nonstandardization and standardization on SUVs and relative changes in SUVs by using examples, data, or literature. The message is that minimal or harmonizing and intrasubject or intersubject standardization is needed, but for some factors, the quantitative impact on the results of multicenter studies is still unclear. On the other hand, when ^{18}F-FDG PET/CT studies are used clinically, the type and degree of standardization will have direct effects on response assessments in individual patients, even when relative changes in SUVs for the same lesion(s) in an individual patient are considered (Fig. 3).

TABLE 1
Recommendations for ^{18}F-FDG PET/CT–Based Treatment Response Assessments in Multicenter Settings

Factor	Standard needed for: Use of SUV (baseline or residual)* — Target lesion eligibility	Use of SUV (baseline or residual)* — Prognostic factor	Use of SUV (baseline or residual)* — Predictive factor	Use of percentage changes in SUVs (longitudinal studies) — Hottest lesion(s)/scan (PERCIST criteria)	Use of percentage changes in SUVs (longitudinal studies) — Same lesion(s) for all scans in subject's longitudinal study (EORTC criteria)
Biologic					
Uptake period	H	H	H	H	M-INTRA (>60 min)
Patient motion or breathing (instructions)	H	H	H	H	H
Patient comfort†	M	M	M	M	M
Inflammation‡	M	M	M	M	M
Physical					
Scan acquisition parameters	H	H	H	H	M-INTRA
Image reconstruction methods, image quality, and quantification	H	H	H	H	M-INTRA
ROI and VOI	H	H	H	H	H
SUV normalization	H	H	H	H	H
Blood glucose level correction	H	H	H	H	H
Contrast agents used during CT-AC	H	H	H	M	M

*Use of SUV from single scan requires harmonizing standards in all cases.
†Patient comfort before and during uptake period and during scanning should meet at least currently accepted recommendations for all patients.
‡Guidelines for correct interpretation and warning about false-positive results due to inflammation should be provided.
H = harmonizing performance standard (including intersubject [or between-subjects] consistency); M-INTRA = minimal performance standard but within-subject consistency of applied methodology; M = minimal performance standard; ROI = region of interest; CT-AC = CT for attenuation correction purposes.

FINAL CONSIDERATIONS

A frequently overlooked issue is that patients often are enrolled or agree to participate in a trial after an ^{18}F-FDG PET/CT examination has been performed. In many cases, the PET/CT studies have not been performed in accordance with current recommendations or guidelines; consequently, a second (baseline) PET/CT study may need to be performed. This situation could be avoided if all studies were performed in a consistent and standardized manner. Therefore, a standardized, globally accepted ^{18}F-FDG PET/CT procedure that is applicable not only to (all) quantitative (research) studies but also to clinical diagnostic PET/CT examinations is urgently needed—exactly why EANM guidelines pertain to trials as well as practice (*18*).

In addition, reducing variability in image quality and quantification between imaging centers could result in more consistent readings between imaging sites. Performing ^{18}F-FDG PET/CT studies in a standardized quantitative manner also allows the use of "fixed" color scales, which could be helpful for the visual interpretation of longitudinal studies; that is, all longitudinal scans of the same subject could be visually compared by use of the same color table and scale to allow a more accurate visual assessment of differences in uptake between the studies.

Finally, evidence will become available only when quantitative PET/CT studies (such as those for response assessments) are performed in a standardized manner, thereby minimizing any methodologic "noise." The latter is presently making a proper correlation (multicenter meta-analysis) between quantitative imaging results and clinical outcomes impossible.

I believe that the clinical validation of quantitative ^{18}F-FDG PET/CT (for response assessments) is even more important than striving to achieve the best possible image quality in individual cases. Therefore, standardization is required to move toward validated quantitative imaging for response assessments.

CONCLUSION

There is wide variability in applied ^{18}F-FDG PET/CT procedures. Complete standardization will reduce the number of quantitative ^{18}F-FDG PET/CT scans that are improperly performed (and therefore not usable) and may improve interinstitutional repeatability. To move toward validated quantitative PET-based response assessments, the radiology community cannot avoid rigorous standardization of ^{18}F-FDG PET/CT.

ACKNOWLEDGMENTS

I would like to thank Otto Hoekstra for his review of the article and all colleagues from the EANM Physics Committee, the EORTC Imaging Workgroup, the SNM Image Reconstruction Harmonization Workgroup, the FDG PET/CT Technical Committee of the QIBA/RSNA, the ^{18}F-FDG PET/CT UPICT Protocol Writing Committee, and many others. The discussions that we had were fruitful and helpful in the writing of this article.

REFERENCES

1. Fletcher JW, Djulbegovic B, Soares HP, et al. Recommendations on the use of ^{18}F-FDG PET in oncology. *J Nucl Med.* 2008;49:480–508.
2. de Geus-Oei LF, van der Heijden HF, Corstens FH, Oyen WJ. Predictive and prognostic value of FDG-PET in nonsmall-cell lung cancer: a systematic review. *Cancer.* 2007;110:1654–1664.
3. Hoekstra CJ, Stroobants SG, Hoekstra OS, et al. The value of [^{18}F]fluoro-2-deoxy-D-glucose positron emission tomography in the selection of patients with stage IIIA-N2 non-small cell lung cancer for combined modality treatment. *Lung Cancer.* 2003;39:151–157.
4. 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.
5. Weber WA. Use of PET for monitoring cancer therapy and for predicting outcome. *J Nucl Med.* 2005;46:983–995.
6. Hoekstra CJ, Paglianiti I, Hoekstra OS, et al. Monitoring response to therapy in cancer using [F-18]-2-fluoro-2-deoxy-D-glucose and positron emission tomography: an overview of different analytical methods. *Eur J Nucl Med.* 2000;27:731–743.
7. Hoekstra CJ, Hoekstra OS, Stroobants SG, et al. Methods to monitor response to chemotherapy in non–small cell lung cancer with ^{18}F-FDG PET. *J Nucl Med.* 2002;43:1304–1309.
8. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med.* 2009;50(suppl 1):11S–20S.
9. Paesmans M, Berghmans T, Dusart M, et al. Primary tumor standardized uptake value measured on fluorodeoxyglucose positron emission tomography is of prognostic value for survival in non-small cell lung cancer: update of a systematic review and meta-analysis by the European Lung Cancer Working Party for the International Association for the Study of Lung Cancer Staging Project. *J Thorac Oncol.* 2010;5:612–619.
10. Young H, Baum R, Cremerius U, et al. Measurement of clinical and subclinical tumour response using [^{18}F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. *Eur J Cancer.* 1999;35:1773–1782.
11. Akhurst T, Downey RJ, Ginsberg MS, et al. An initial experience with FDG-PET in the imaging of residual disease after induction therapy for lung cancer. *Ann Thorac Surg.* 2002;73:259–264.
12. Hoekstra CJ, Stroobants SG, Smit EF, et al. Prognostic relevance of response evaluation using [F-18]-2-fluoro-2-deoxy-D-glucose positron emission tomography in patients with locally advanced non-small-cell lung cancer. *J Clin Oncol.* 2005;23:8362–8370.
13. 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.
14. Beyer T, Czernin J, Freudenberg LS. Variations in clinical PET/CT operations: results of an international survey of active PET/CT users. *J Nucl Med.* 2011; 52:303–310.
15. Graham MM, Badawi RD, Wahl RL. Variations in PET/CT methodology for oncologic imaging at U.S. academic medical centers: an imaging response assessment team survey. *J Nucl Med.* 2011;52:311–317.
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. ^{18}F-FDG PET/CT UPICT Protocol Writing Committee. UPICT FDG Consolidated Statement (draft). Available at: http://upictwiki.ctsa-imaging.org/index.php?title = Main_Page#UPICT_Process_and_Template. Accessed October 10, 2011.
18. Boellaard R, O'Doherty MJ, Weber WA, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging.* 2010;37:181–200.
19. Quarles van Ufford HM, van Tinteren H, Stroobants SG, et al. Added value of baseline ^{18}F-FDG uptake in serial ^{18}F-FDG PET for evaluation of response of solid extracerebral tumors to systemic cytotoxic neoadjuvant treatment: a meta-analysis. *J Nucl Med.* 2010;51:1507–1516.
20. Borst GR, Belderbos JSA, Boellaard R, et al. Standardised FDG uptake: a prognostic factor for inoperable non-small cell lung cancer. *Eur J Cancer.* 2005;41:1533–1541.
21. Downey RJ, Akhurst T, Gonen M, et al. Fluorine-18 fluorodeoxyglucose positron emission tomographic maximal standardized uptake value predicts survival independent of clinical but not pathologic TNM staging of resected non-small cell lung cancer. *J Thorac Cardiovasc Surg.* 2007;133:1419–1427.
22. de Geus-Oei LF, Wiering B, Krabbe PF, et al. FDG-PET for prediction of survival of patients with metastatic colorectal carcinoma. *Ann Oncol.* 2006;17: 1650–1655.
23. Boellaard R, Krak NC, Hoekstra OS, Lammertsma AA. Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study. *J Nucl Med.* 2004;45:1519–1527.
24. Lowe VJ, DeLong DM, Hoffman JM, Coleman RE. Optimum scanning protocol for FDG-PET evaluation of pulmonary malignancy. *J Nucl Med.* 1995;36: 883–887.
25. Akhurst T. Lessons from the old masters: pragmatism or purity, FDG PET SUV, serum glucose and prediction of nodal status in non-small cell lung cancer. *J Surg Oncol.* 2006;94:547–548.
26. Lammertsma AA, Hoekstra CJ, Giaccone G, Hoekstra OS. How should we analyse FDG PET studies for monitoring tumour response? *Eur J Nucl Med Mol Imaging.* 2006;33(suppl 1):16–21.
27. Lee WW, Chung JH, Jang SJ, et al. Consideration of serum glucose levels during malignant mediastinal lymph node detection in non-small-cell lung cancer by FDG-PET. *J Surg Oncol.* 2006;94:607–613.
28. Lindholm P, Minn H, Leskinen-Kallio S, et al. Influence of the blood glucose concentration on FDG uptake in cancer: a PET study. *J Nucl Med.* 1993;34:1–6.
29. Cheebsumon P, Velasquez LM, Hoekstra CJ, et al. Measuring response to therapy using FDG PET: semi-quantitative and full kinetic analysis. *Eur J Nucl Med Mol Imaging.* 2011;38:832–842.
30. Diederichs CG, Staib L, Glatting G, et al. FDG PET: elevated plasma glucose reduces both uptake and detection rate of pancreatic malignancies. *J Nucl Med.* 1998;39:1030–1033.
31. Delbeke D, Coleman RE, Guiberteau MJ, et al. Procedure guideline for tumor imaging with ^{18}F-FDG PET/CT 1.0. *J Nucl Med.* 2006;47:885–895.
32. Wong KP, Sha W, Zhang X, Huang SC. Effects of administration route, dietary condition, and blood glucose level on kinetics and uptake of ^{18}F-FDG in mice. *J Nucl Med.* 2011;52:800–807.
33. Boellaard R, Hayes W, Hoetjes N, et al. Impact of quality assurance, centralized analysis, ROI method and glucose correction on SUV test-retest variability in a multi-center setting [abstract]. *J Nucl Med.* 2009;50(suppl 2):627.
34. Dai KS, Tai DY, Ho P, et al. Accuracy of the EasyTouch blood glucose self-monitoring system: a study of 516 cases. *Clin Chim Acta.* 2004;349:135–141.
35. van der Putten L, Hoekstra OS, de Bree R, et al. 2-Deoxy-2[F-18]FDG-PET for detection of recurrent laryngeal carcinoma after radiotherapy: interobserver variability in reporting. *Mol Imaging Biol.* 2008;10:294–303.
36. Rapisarda E, Bettinardi V, Thielemans K, Gilardi MC. Image-based point spread function implementation in a fully 3D OSEM reconstruction algorithm for PET. *Phys Med Biol.* 2010;55:4131–4151.
37. Krak NC, Boellaard R, Hoekstra OS, et al. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Imaging.* 2005;32:294–301.
38. Kelly M, Declerck J. SUVref: reducing reconstruction-dependent variation in PET SUV [abstract]. *J Nucl Med.* 2010;51(suppl 2):355.