Repeatability of SUV in Oncologic ^{18}F-FDG PET

Martin A. Lodge

Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, Maryland

Learning Objectives: On successful completion of this activity, participants should be able to (1) describe the way test–retest studies have been used to measure SUV repeatability, (2) summarize the different methodologic approaches and complexities when analyzing SUV test–retest data, and (3) understand the implications of SUV repeatability for the quantitative assessment of tumor response to treatment.

Financial Disclosure: This work was partially funded by grants from the National Institutes of Health (HHSN268201300071C and U01CA140204). The author of this article has indicated no other relevant relationships that could be perceived as a real or apparent conflict of interest.

CME Credit: SNMMI is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to sponsor continuing education for physicians. SNMMI designates each *JNM* continuing education article for a maximum of 2.0 AMA PRA category 1 credits. Physicians should claim only credit commensurate with the extent of their participation in the activity. For CE credit, SAM, and other credit types, participants can access this activity through the SNMMI website (http://www.snmmilearningcenter.org) through April 2020.

Quantitative analysis can potentially improve the accuracy and consistency of ^{18}F-FDG PET, particularly for the assessment of tumor response to treatment. Although not without limitations, SUV has emerged as the predominant metric for tumor quantification with ^{18}F-FDG PET. Growing literature suggests that the difference between SUVs measured before and after treatment can be used to predict tumor response at an early stage. SUV is, however, associated with multiple sources of variability, and to best use SUV for response assessment, an understanding of the repeatability of the technique is required. Test–retest studies involve repeated scanning of the same patient on the same scanner using the same protocol no more than a few days apart and provide basic information on the repeatability of the technique. Multiple test–retest studies have been performed to assess SUV repeatability, although a comparison of reports is complicated by the use of different methodologies and statistical metrics. This article reviews the available data, addressing issues such as different repeatability metrics, relative units, log transformation, and asymmetric limits of repeatability. When acquired with careful attention to protocol, tumor SUV has a within-subject coefficient of variation of approximately 10%. In a response assessment setting, SUV reductions of more than 25% and increases of more than 33% are unlikely to be due to measurement variability. Broader margins may be required for sites with less rigorous protocol compliance, but in general, SUV is a highly repeatable imaging biomarker that is ideally suited to monitoring tumor response to treatment in individual patients.

Key Words: repeatability; SUV; standardized uptake value; FDG; PET; reproducibility

J Nucl Med 2017; 58:523–532
DOI: 10.2967/jnumed.116.186353

Quantitative analysis can potentially improve the accuracy and consistency of oncologic ^{18}F-FDG PET, particularly for the assessment of tumor response to treatment (1,2). When tumor response is only partial or when small changes occur early after treatment, before the full treatment effect is complete, visual assessment can be problematic (3). Subjective interpretation can lead to inconsistency between readers, potentially undermining the value of the study. These concerns apply not only to clinical practice but also to clinical trials, in which there is a greater expectation for robust quantitative data. Growing evidence suggests that, for applications such as these, visual assessment can be enhanced by supplementary quantitative analysis (4), an approach to which PET is particularly well suited.

SUV was initially regarded with mixed enthusiasm (5), but as the methodology improved, it emerged as the predominant metric for tumor quantification with ^{18}F-FDG PET. Although it may lack the scientific rigor and conceptual attractiveness of more sophisticated kinetic modeling approaches (6), it has substantial advantages in terms of practicality and compatibility with clinical protocols. It also has a large base of evidence supporting its use for the noninvasive assessment of tumor response to treatment (7–12). Changes in SUV between baseline and follow-up studies can help determine whether tumors are responding to treatment. The follow-up PET evaluation can potentially be performed early after the end of treatment, well before a change in tumor size can be seen on anatomic imaging. The ability to assess tumor response early after treatment may, for example, allow nonresponders to be redirected to more appropriate treatment. Or in the case of clinical trials, early tumor assessment can aid drug development by identifying ineffective therapies before they are deployed in large, expensive multicenter trials.

Although simplicity and ease of use are among the strengths of SUV, the measurement is nevertheless vulnerable to many sources of unwanted variability (13). These include issues associated with biologic variability, patient preparation, scanner stability, image quantitative accuracy, and image analysis, including tumor volume-of-interest (VOI) techniques. Improved standardization of methodology has gone some way toward mitigating these problems, but many sources of variability remain. Knowledge of the repeatability

Received Nov. 29, 2016; revision accepted Feb. 21, 2017.

For correspondence or reprints contact: Martin A. Lodge, Division of Nuclear Medicine and Molecular Imaging, Johns Hopkins University School of Medicine, 601 N. Caroline St., Baltimore, MD 21287.

E-mail: mlodge1@jhmi.edu

Published online Feb. 23, 2017.

of SUV measurements is particularly relevant for response assessment studies because it provides a basis for interpreting the tumor SUVs obtained at baseline and follow-up. What change in SUV should be interpreted as a real change in a particular tumor? And what change in SUV should be attributed simply to measurement variability? Changes in SUV beyond the expected range of variability are not consistent with stable disease, and the extent of the difference can help guide or substantiate the reader's impression. In the clinical trial context, repeatability can determine the number of patient volunteers who need to be enrolled to confirm a particular effect (*14*). As such, repeatability can directly influence the cost of a trial and, in turn, the cost of developing new therapies. An understanding of the repeatability of SUV measurements is thus important for both clinical and research applications.

The literature on the repeatability of oncologic ^{18}F-FDG SUV has developed slowly, most likely because of the difficulty in acquiring the relevant data. Phantom studies (*15*) and simulation studies (*16*) are capable of capturing important components of variability, but more directly representative data require patient measurements acquired under test–retest conditions. Repeated scanning of the same patient on the same scanner using the same protocol no more than a few days apart provides basic information on the repeatability of the technique. Under the assumption that the tumor has not progressed over this short period, the SUVs would ideally be identical. In practice, measurement variability means that the two SUVs are not identical, and when data are acquired over a large group of patients, the expected range of repeatability can be estimated. The term *reproducibility* is sometimes used in this context, but this term is more correctly used to refer to studies performed in different settings (*17*), such as on different scanner systems. Although reproducibility is of interest, this review focuses on the data that are currently available, which are mostly repeatability data.

Several reports have been published describing the repeatability of tumor SUV with ^{18}F-FDG PET or PET/CT. However, a comparison of these papers is not straightforward because of differences in methodology, such as the use of different acquisition protocols or image analysis methods. In particular, because the literature includes different approaches to statistical analysis, repeatability is often expressed using metrics or nomenclature that are not the same even when the experimental methods are substantially similar. Consequently, the literature includes results that often are not directly comparable and may be somewhat confusing. This article attempts to review the available literature, reconcile differences between the publications, and clarify expectations for the repeatability of tumor SUV.

SUV REPEATABILITY LITERATURE

The scientific literature was reviewed with the aim of identifying publications related to ^{18}F-FDG PET and the repeatability of tumor SUV. The online databases PubMed (U.S. National Library of Medicine, National Institutes of Health) and Google Scholar (Google Inc.) were searched using terms such as *FDG, PET, SUV, repeatability,* and *reproducibility.* The main inclusion criterion was that each paper contained all of the following components: measurement of SUV repeatability in a test–retest study design, human as opposed to animal studies, quantification in tumors as opposed to normal organs or other disease states, and ^{18}F-FDG as opposed to other radiopharmaceuticals. For this purpose, we considered a test–retest study design to involve two imaging studies performed on the same patient on the same scanner system using the same acquisition and analysis protocol. To be clear, each of the two imaging studies had to involve separate ^{18}F-FDG administrations so as to capture the variability associated with biologic effects, patient preparation, and tracer administration. The interval between successive imaging studies was not rigidly specified in our search but was typically between 1 and 7 d. Importantly, we specified that no treatment or other significant interventions could take place between the two studies. Specifically excluded from further analysis were animal studies, phantom studies, and computer simulation studies. Although relevant, these studies are not expected to be directly comparable to human studies, which were the main interest. Also excluded were studies that involved repeated imaging after a single ^{18}F-FDG administration (*18*), studies that measured the repeatability of different readers analyzing the same images (*19*), and repeatability studies that did not include SUV quantification.

Table 1 shows the articles that were identified and included in this review. Sixteen papers (*20–35*), published between 1995 and 2016, met the inclusion criteria. All were reports on original research, although there was some overlap in the source data: Nakamoto et al. (*22*) performed a retrospective analysis of data previously published by Minn et al. (*20*); Krak et al. (*23*) analyzed SUV measurements derived from dynamic data originally presented by Hoekstra et al. (*36*); van Velden et al. (*31*) analyzed a subset of the data published by Velasquez et al. (*25*); and de Langen et al. (*28*) performed a metaanalysis pooling data from 5 previously published cohorts. Several closely related papers did not strictly meet the requirements of our review but are nevertheless relevant. Examples include the previously mentioned work of Hoekstra et al. (*36*), which included test–retest data on patients with non–small cell lung cancer but assessed the repeatability of tracer kinetic analysis as opposed to SUV. Kamibayashi et al. (*37*) assessed the reproducibility of tumor SUVs acquired using different scanner systems: one a PET-only scanner and the other a PET/CT system. Bengtsson et al. (*38*) reported on a study that involved repeated imaging, but in this case the interval between the imaging studies was extended (median, 21 d) and the patients received treatment in the intervening period, albeit treatment that proved to be ineffective. Although not included in the following analysis, some of these papers will be discussed subsequently.

DATA ACQUISITION

The range of tumors that have been included in test–retest studies is shown in Table 1. Lung cancer has been a particular focus, but a wide range of other cancer types has also been studied, including gastrointestinal malignancies, esophageal cancer, colorectal cancer, head and neck cancer, and ovarian cancer. Each of these studies involved a careful test–retest protocol with two repeated imaging sessions using the same protocol and scanner system for each patient. Four of the publications (*25,31–33*) included data acquired at multiple centers, although, to be clear, individual patients were always scanned on the same system. The remaining reports were on single-center studies. A limitation of many of these studies is the small number of patients that were included (median, 18). However, when all publications are considered as a whole, test–retest data have been obtained for over 300 patients.

Because the literature spans more than 20 y, different generations of PET instrumentation have been used, including both PET and PET/CT scanners from various manufacturers. Data acquisition

TABLE 1
Literature on Tumor SUV Repeatability

Publication	Year	Tumor type	Comments
Minn (*20*)	1995	Lung cancer, n = 10	PET; localized view (dynamic)
Weber (*21*)	1999	Various malignancies, n = 16	PET; localized view (dynamic)
Nakamoto (*22*)	2002	Lung cancer, n = 10	Retrospective analysis of Minn (*20*) data
Krak (*23*)	2005	Non–small cell lung cancer, n = 11	PET; localized view (dynamic)
Nahmias (*24*)	2008	Various malignancies, n = 26	PET/CT; whole-body
Velasquez (*25*)	2009	Advanced gastrointestinal malignancies, n = 61	PET and PET/CT; multicenter study
Hatt (*26*)	2010	Esophageal cancer, n = 14	PET/CT; whole-body
Heijmen (*27*)	2012	Liver metastases in colorectal cancer, n = 18	PET/CT; whole-body
de Langen (*28*)	2012	Various malignancies, n = 102 in largest subgroup	Metaanalysis
Hoang (*29*)	2013	Head and neck squamous cell carcinoma, n = 17	PET/CT; localized view
Kumar (*30*)	2013	Various malignancies, mostly colon cancer, n = 21	PET/CT; whole-body
van Velden (*31*)	2014	Colorectal cancer, n = 29	PET/CT; whole-body
Rockall (*32*)	2014	Ovarian cancer, n = 21	PET/CT; whole-body; two centers
Weber (*33*)	2015	Non–small cell lung cancer, n = 74	PET/CT; whole-body; multicenter study
Rasmussen (*34*)	2015	Head and neck squamous cell carcinoma, n = 24	PET/CT and PET/MR; localized view
Kramer (*35*)	2016	Non–small cell lung cancer, n = 9	PET/CT; whole-body

n = number of patients in each study.

methods reflected the evolving state of the technology over this period and have included bismuth germanate and lutetium oxyorthosilicate detectors, 2-dimensional and 3-dimensional acquisition geometries, and scanner systems with and without time-of-flight capability. Various reconstruction algorithms have been used, and although they were used consistently within a given study, we should not assume consistency between different studies. For example, Minn et al. (*20*) used filtered backprojection, producing an estimated spatial resolution of 12 mm in full width at half maximum, whereas Krak et al. (*23*) used an ordered-subset expectation-maximization iterative algorithm and estimated a spatial resolution of 7 mm in full width at half maximum.

Depending on the study, PET data were acquired as dynamic scans at a single bed position, localized head-and-neck studies (1 or 2 bed positions), or whole-body studies typically covering the base of the skull to mid-thigh (2–5 min per bed position). When dynamic data were acquired (*20,21,23*), a frame of 10–15 min starting approximately 60 min after injection was used for SUV calculation. For the static studies, the interval between ^{18}F-FDG administration and the start of the PET acquisition was typically 60 min, although Nahmias and Wahl (*24*) favored 90 min. Kramer et al. (*35*) assessed repeatability at both 60 min and 90 min. Careful adherence to maintaining consistent uptake periods was a feature of most studies. For example, Rockall et al. (*32*) reported that, for a given patient, the difference in the uptake periods between scan 1 and scan 2 averaged 1.9 min. Such careful control of uptake periods was important for optimizing repeatability but may not be typical of clinical conditions. The study by Kumar et al. (*30*) showed an average difference of 33 ± 20 min between corresponding uptake periods and may better reflect the repeatability that can be expected in a more typical setting (*39*).

The literature is complicated by the different tumor-sampling schemes that have been used. In general, there have been 3 different VOI approaches, with their corresponding SUVs being SUV_{max} (*22,23,25–27,29–35*), SUV_{mean} (*21,23–27,29–31,34,35*), and SUV_{peak} (*20,22,23,25,27,32–35*). As is usual, SUV_{max} was derived from the single tumor voxel with the highest uptake. Given its unambiguous definition, SUV_{max} would be expected to be most comparable between reports, although it should be noted that the voxel dimensions were not the same across studies (e.g., 2.3 × 2.3 × 3.3 mm for the head and neck (*29*) and 5.5 × 5.5 × 3.3 mm for the whole body (*30*)). SUV_{mean} was derived from the average value of all voxels within an extended VOI. These VOIs were usually defined by isocontour thresholding, typically based on a fixed percentage of SUV_{max} (e.g., 50%), occasionally including background correction. Other tumor segmentation approaches were also used, including fuzzy locally adaptive Bayesian methodology (*26,27*), manual delineation (*23,29*), and circular regions manually adjusted to the dimensions of the tumor (*24*). SUV_{peak} has been defined as the average of all voxels within a 1-mL spheric region positioned within the tumor so as to maximize its mean value (*1*). Some of the repeatability papers were published before the term SUV_{peak} was adopted and instead use other designations. In various cases, the peak region was defined slightly differently from the above criterion, frequently involving small (e.g., 12-mm) circular or square regions of interest centered over the maximum tumor voxel. For the purposes of this review, when a small fixed-size VOI with a volume of approximately 1 mL was used, we refer to this as SUV_{peak} even though the original article may not have used this term.

The number of tumors analyzed for each patient varied among studies, and some reports included multiple analyses. The most common approach was to analyze a single tumor per patient (*20–22,24,25,29,31,33,34*). Another approach allowed for the inclusion of a variable number of tumors per patient, analyzing all tumors collectively (*21,23,26,27,30,35*) or averaging the tumor SUVs for an individual patient and assessing the repeatability of the average SUV (*25,32,33,35*). Inclusion criteria in terms of minimum tumor size or SUV were not always specified. When these criteria were reported, a minimum diameter of 2 cm in all 3 orthogonal dimensions (*20*) or at least 3 cm in the largest direction (*35*) was typical. Rockall et al. (*32*) and Weber et al. (*33*) specified a minimum $\mathrm{SUV}_{\max}$ of 2.5 and 4.0, respectively. SUV was normalized using the patient's body mass or lean body mass (*20,22,23,25,35*), estimated using predictive equations. Lean body mass has the advantage of making SUVs more comparable between patients with different body compositions. Intersubject variability is reduced (e.g., normal-organ SUV), but lean body mass normalization would not be expected to alter within-subject variability, at least not in this test–retest setting.

REPEATABILITY ANALYSIS

With regard to statistical analysis, several slightly different approaches can be found in the literature. The relationships between the various statistical metrics (Table 2) are not immediately obvious and have caused some confusion. Older publications tend to characterize repeatability in terms of the mean absolute percentage difference (MAPD), whereas more recent papers tend to use the repeatability coefficient (RC) derived from Bland–Altman analysis (*40*). Both approaches reflect repeatability, but RC provides useful limits beyond which an SUV change is likely to reflect a true change in an individual tumor.

SUV_1 and SUV_2 denote corresponding SUV measurements of the same tumor under test–retest conditions. The difference *d* is given simply as

$$d = \mathrm{SUV}_2 - \mathrm{SUV}_1. \qquad \text{Eq. 1}$$

The parameter *d* has the units of the original SUV measurements (e.g., g/mL), but the difference can also be expressed in relative units (*D*):

$$D = \frac{\mathrm{SUV}_2 - \mathrm{SUV}_1}{\overline{\mathrm{SUV}}} \times 100\%, \qquad \text{Eq. 2}$$

where

$$\overline{\mathrm{SUV}} = \frac{1}{2}(\mathrm{SUV}_1 + \mathrm{SUV}_2). \qquad \text{Eq. 3}$$

Note that *D* is the difference expressed as a percentage of the average of the two measurements. The absolute value, $|D|$, can be averaged over multiple patient studies to determine the MAPD as follows:

$$\mathrm{MAPD} = \frac{1}{n}\sum_{i=1}^{n}|D_i|, \qquad \text{Eq. 4}$$

where D_i indicates the relative difference for multiple patients ($i = 1 \ldots n$).

An alternative statistical approach involves taking the SD of the test–retest differences. The data can be conveniently presented as a Bland–Altman plot (Fig. 1) in which the differences between two repeated measurements, in either original units (*d*) or relative units (*D*), are plotted as a function of their average ($\overline{\mathrm{SUV}}$). Subsequent analysis is based on meeting the following two conditions: that there be no proportionality between the magnitude of the difference data ($|d|$ or $|D|$) and the average ($\overline{\mathrm{SUV}}$), and that the difference data (*d* or *D*) be normally distributed. Confirmation of the first condition indicates that the variability of the measurement is independent of the magnitude of the SUV and that the resulting repeatability estimate is valid for tumors with very different SUVs. If this were not the case and, for example, |d| were proportional to $\overline{\mathrm{SUV}}$, estimates of repeatability would likely be too high for low-SUV tumors and too low for high-SUV tumors. Confirmation of the second condition allows 95% limits of repeatability to be estimated, because for normally distributed data we would expect 95% of the differences to be within approximately 2 SDs.

Having established that the data satisfy these conditions, we can determine the SD of the difference data. In most cases, relative data were used and the SD of *D* (DSD) can be considered a coefficient of variation. Note that DSD is not the variability in a single measurement, because *D* is subject to noise in both SUV_1 and SUV_2. The within-subject coefficient of variation (wCV) of a single measurement is given by DSD/√2 and is often reported as the primary metric of repeatability. RC is directly related to wCV

TABLE 2
Various Repeatability Metrics Found in the Literature

Parameter	Symbol	Definition	Comment
Difference	*d*	$\mathrm{SUV}_2 - \mathrm{SUV}_1$	Test–retest difference in units of original data
Relative difference	*D*	$100 \times d/[0.5 \times (\mathrm{SUV}_1 + \mathrm{SUV}_2)]$	Test–retest difference expressed as percentage of mean
SD of *D*	DSD	SD of *D* over all subjects	SD of *d* would be used if analysis were performed in original units
Within-subject coefficient of variation	wCV	DSD/√2	Reflects repeatability of a single measurement
Repeatability coefficient	RC	1.96 × √2 × wCV	Reflects 95% limits of repeatability for difference between two measurements under assumption that *D* is normally distributed
Mean absolute percentage difference	MAPD	Mean of |*D*| over all subjects	|*D*| indicates absolute value of *D*

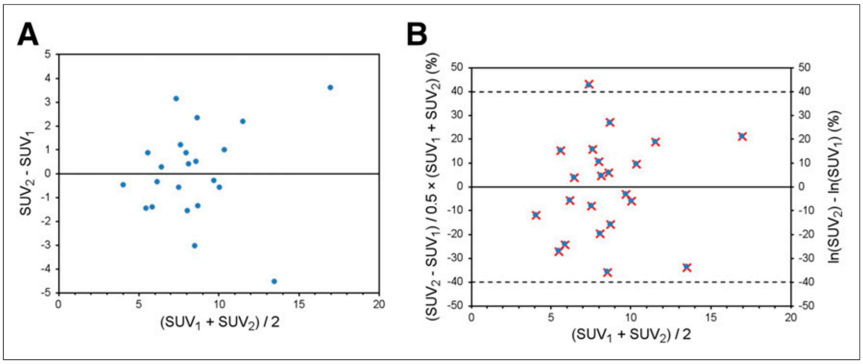

FIGURE 1. Bland–Altman plots showing SUV$_{max}$ difference data (●) in both original units (A) and relative units (B), as function of mean. Relative difference data are consistent with normal distribution, and Kendall τ (0.026, $P = 0.89$) indicates that magnitude is not proportional to mean. Dashed lines show limits of repeatability given by [−RC, +RC]. In B, data have also been plotted using secondary *y*-axis (×). Close agreement between red and blue data points illustrates that scales on left and right *y*-axes are substantially similar. (Data are from Heijmen et al. (*27*), kindly provided by the author.)

and DSD and is given by 1.96 × DSD. Under the assumption that *D* is normally distributed, RC represents the 95% limits of repeatability for the difference between 2 SUV measurements made under test–retest conditions. In other words, baseline and follow-up SUV measurements made on a perfectly stable tumor should be expected to differ by up to RC 95% of the time. Conversely, if the change in SUV were to exceed RC, it is reasonable to infer some real change in the tumor.

The relationship between MAPD and DSD was not stated in any of the papers included in this review. However, it can be shown that MAPD can be related to DSD under certain assumptions. The Bland–Altman approach, and the associated 95% limits of repeatability, require that the difference data *D* be normally distributed. For the purpose of comparing reports, it is reasonable to make this same assumption for the data that were originally analyzed in terms of MAPD. If we further assume that the difference data have a mean of zero, which is reasonable for test–retest data, it can be shown (*41*) that

$$\mathrm{MAPD} = \sqrt{\frac{2}{\pi}} \times \mathrm{DSD}. \qquad \text{Eq. 5}$$

The applicability of this relationship can be illustrated using data from the article of Nakamoto et al. (*22*). DSD was calculated from the tabulated SUV$_{max}$ data to be 13.44%. According to Equation 5, this corresponds to an MAPD of 10.72%, which is in close agreement with the published value of 11.30%, calculated using Equation 4. This relationship and the other relationships shown in Table 2 allow the data from the different reports to be directly compared.

ORIGINAL UNITS OR RELATIVE UNITS?

One issue that arises in test–retest studies of this kind is whether to analyze the data in the units of the original measurement (*d* expressed in SUV units) or in relative units (*D* expressed as a percentage). Relative units are integral to the calculation of MAPD, but RC can be expressed either in SUV units or as a percentage. The appropriate choice depends on the characteristics of the data and is an important consideration. Figure 1 shows an example (*27*) that illustrates the typical dependence of the difference data on the magnitude of the SUV. The absolute difference in the original units ($|d|$) was usually found to be proportional to the average ($\overline{\mathrm{SUV}}$), and as a result, limits of repeatability expressed in SUV units would not be applicable over the full range of SUVs. Relative units appear to be a better way to express SUV repeatability, because the magnitude of the relative difference ($|D|$) was generally independent of $\overline{\mathrm{SUV}}$. Most but not all (*24*) papers addressing SUV repeatability expressed their results in dimensionless relative units.

Characterizing repeatability in relative units is well suited to the way SUV is used in response assessment studies, which commonly quote percentage change in SUV relative to a baseline measurement. In addition to being easily interpreted, relative units are helpful when one is comparing literature reports that use different SUV formulations. SUV data derived using lean body mass as opposed to total body mass normalization have different ranges and are not directly comparable. However, the use of the relative difference *D* to characterize repeatability allows comparison of data from different reports irrespective of the SUV normalization schemes.

An important contribution was made by de Langen et al. (*28*), who investigated the relationship between SUV variability and tumor uptake. By combining data from multiple studies, they showed that test–retest differences expressed in relative units ($|D|$) were not, in fact, independent of the level of uptake ($\overline{\mathrm{SUV}}$) as assumed in most other studies. Even when expressed in percentage terms, repeatability improved with higher uptake, and it may not be correct to assume that fixed limits of repeatability are applicable across the full range of SUVs. A practical concern is for low-uptake tumors that have poorer repeatability than the wider group. To account for these low-uptake tumors, de Langen et al. recommended that minimal changes in both relative and absolute SUVs be required for tumor response assessment studies.

Although not yet resolved, it seems that relative units may be more appropriate than original units but that neither is entirely adequate. The most complete way to characterize repeatability, including the most appropriate units, remains a subject of ongoing interest.

LOG TRANSFORMATION

Closely related to the use of relative units is the use of log transformation. The fact that only a subset of papers (*25,32,33*) used log transformation would seem to complicate comparison of reports, but in fact, log-transformed data can readily be compared with relative difference data. Log transformation is a way of accounting for the proportionality that was usually found between the absolute difference ($|d|$) and the average ($\overline{\mathrm{SUV}}$). Natural log transformation is recommended, as opposed to other log transforms, because the difference in natural logs has a very intuitive interpretation. $\ln(\mathrm{SUV}_2) - \ln(\mathrm{SUV}_1)$ is approximately equal to the relative difference, $(\mathrm{SUV}_2 - \mathrm{SUV}_1)/\overline{\mathrm{SUV}}$. For example, if SUV$_1$ and SUV$_2$ are assumed to be 9 and 10, respectively, $(\mathrm{SUV}_2 - \mathrm{SUV}_1)/\overline{\mathrm{SUV}} = 0.105$ and $\ln(\mathrm{SUV}_2) - \ln(\mathrm{SUV}_1) = 0.105$. The applicability of this close approximation has been confirmed for

PET repeatability data (*42*) and is illustrated in Figure 1B. It can be seen that difference data on the natural log scale can be directly interpreted as relative differences without the need for back-transformation. The SD of difference data on the log scale (20.5% for the data in Fig. 1B) is largely equivalent to the DSD derived from relative units (20.3% for the data in Fig. 1B). This relationship greatly simplifies interpretation of log-transformed data and allows a direct comparison of reports that use relative difference data (*D*) and natural log transformation.

SYMMETRIC OR ASYMMETRIC LIMITS OF REPEATABILITY?

Some differences exist in the literature regarding interpretation of RC. If the test–retest difference data can be assumed to be normally distributed, with zero mean and a variability that is constant over the range of measurements, the 95% limits of repeatability are given by [–RC, +RC]. In the test–retest setting, SUV differences are as likely to be in one direction as in the other, and the limits of repeatability are symmetric about zero. This interpretation is frequently adopted in the SUV repeatability literature and is consistent with the general framework of Bland and Altman (*40*). However, two notable PET papers (*25,33*) include the use of asymmetric limits of repeatability in which the lower and upper RCs differ. For example, Weber et al. (*33*) reported that a decrease in $\mathrm{SUV}_{\max}$ by more than 28% would be required to indicate tumor response, whereas tumor progression would require an increase by more than 39%. These asymmetric limits are not so much due to an inadequate number of samples in the test–retest data, nor are they due to a systematic bias between the first and second scans. Asymmetric limits of repeatability were introduced in order to account for SUV changes relative to a baseline value (*33*).

In a test–retest setting, relative difference data would typically be expressed with respect to the average of two measurements, according to Equation 2. However, this situation differs from the typical clinical situation, in which the difference between baseline (SUV_1) and follow-up (SUV_2) is usually expressed relative to a single baseline measurement:

$$\Delta\mathrm{SUV} = \frac{(\mathrm{SUV}_2 - \mathrm{SUV}_1)}{\mathrm{SUV}_1} \times 100\%. \qquad \text{Eq. 6}$$

For example, if baseline and follow-up SUVs were 18 and 25, respectively, ΔSUV would be approximately +39%. However, if the same two SUVs were considered in reverse (baseline SUV of 25, follow-up SUV of 18), ΔSUV would be −28%. The use of a single baseline SUV as the reference leads to a skewing of the data that necessitates the asymmetric RCs.

Figure 2 attempts to illustrate the situation. Two random samples were drawn from a normal distribution with a coefficient of variation of 12%. This procedure simulated an idealized test–retest setting and was chosen to match the $\mathrm{SUV}_{\max}$ data of Weber et al. (*33*). The sampling process was repeated 1,000 times, and Figure 2A shows the SUV differences divided by their average (Eq. 2). With this particular set of samples, DSD was measured to be 16.7%, corresponding to an RC of 33%, which is shown as symmetric limits in Figure 2A. In Figure 2B the same SUV difference data were divided by a single baseline SUV (Eq. 6), and an asymmetric distribution is clear. For example, notice that there are no data points below −40% but many above +40%. Asymmetric RCs can be determined following the approach of Velasquez et al. (*25*) and Weber et al. (*33*):

$$\mathrm{LRC} = (\exp(-1.96 \cdot \mathrm{SD}_{\mathrm{dln}}) - 1) \times 100\% \qquad \text{Eq. 7}$$

$$\mathrm{URC} = (\exp(+1.96 \cdot \mathrm{SD}_{\mathrm{dln}}) - 1) \times 100\%, \qquad \text{Eq. 8}$$

where LRC is the lower RC, URC is the upper RC, and $\mathrm{SD}_{\mathrm{dln}}$ is the SD of the difference on the log scale. Similar asymmetric limits can be obtained by converting the symmetric RC limits in the units of Equation 2 (relative to the average of two measurements) to their equivalent using the units shown in Equation 6 (relative to a single baseline measurement). It can be shown that

$$\mathrm{LRC} = \frac{-\mathrm{RC}}{1 + (\mathrm{RC}/200\%)} \qquad \text{Eq. 9}$$

$$\mathrm{URC} = \frac{\mathrm{RC}}{1 - (\mathrm{RC}/200\%)}, \qquad \text{Eq. 10}$$

where LRC, URC, and RC (the symmetric limit defined in Table 2) are all in percentage terms. Figure 2B shows LRC and URC limits at [–28%, +39%], and it can be seen that 50 data points lie outside this range, indicating that 95% of the 1,000 data points are within these asymmetric limits. Asymmetric RCs are thus seen to be appropriate for changes relative to a baseline measurement, which is the way SUV is currently used in the response assessment setting.

SUMMARY OF REPEATABILITY RESULTS

This section compares the results from the different studies, with the caveat that such a comparison inevitably involves data acquired under slightly different conditions. For example, the following analysis includes repeatability data from studies that analyzed multiple tumors per patient as well as studies that assessed only one tumor per patient. To compare results, the different statistical metrics were converted to a common parameter, wCV. For the papers that used the Bland–Altman methodology, wCV could be readily inferred using the relationships summarized in Table 2 even if not explicitly reported in the original article. For the papers that reported MAPD, Equation 5 was also used. For example, Nakamoto et al. (*22*) reported the MAPD for $\mathrm{SUV}_{\max}$ to be 11.30%. Using Equation 5, we can infer a DSD of 14.16% and a wCV of 10.01%.

Table 3 shows how the $\mathrm{SUV}_{\max}$ results from each paper were converted to an inferred wCV using the procedure described above. Similar analyses were performed for $\mathrm{SUV}_{\mathrm{mean}}$ and $\mathrm{SUV}_{\mathrm{peak}}$ and are shown in Tables 4 and 5, respectively. Inferred wCV values for all 3 SUV metrics are shown graphically in Figure 3. The mean wCV over all relevant papers was 10.96% (SD, 3.32), 9.98% (SD, 3.06), and 9.60% (SD, 3.40) for $\mathrm{SUV}_{\max}$, $\mathrm{SUV}_{\mathrm{mean}}$, and $\mathrm{SUV}_{\mathrm{peak}}$, respectively. The differences between these means were not statistically significant ($P > 0.05$), and the overall average wCV, combining all 3 SUV metrics, was 10.27% (SD, 3.20).

DISCUSSION

In this paper, the literature on the repeatability of SUV in ^{18}F-FDG oncologic PET has been reviewed. Differences and shared aspects of methodology were identified, in particular with regard to statistical analysis. By converting different statistical measures to a common index, we were able to directly compare results from multiple reports. Over all the publications, which included tumors

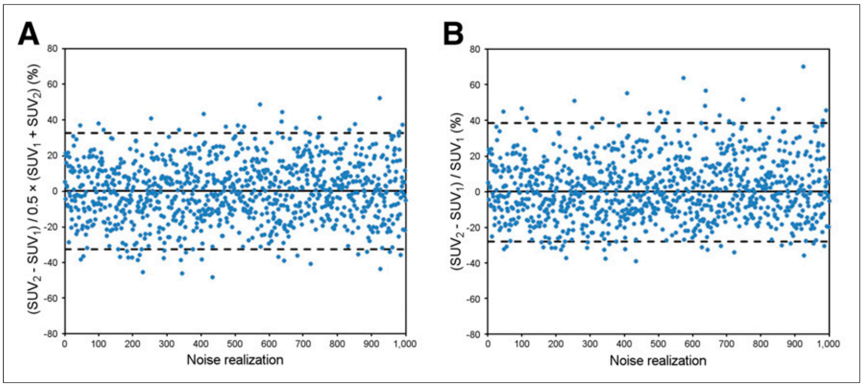

FIGURE 2. Simulated test–retest difference data. SUV$_1$ and SUV$_2$ were random samples drawn from normal distribution with coefficient of variation of 12%. One thousand pairs of random samples were generated, corresponding to the different noise realizations shown on *x*-axis. (A) Differences between SUV$_1$ and SUV$_2$ are shown relative to their average. DSD was 16.7% (MAPD, 13.5%). Dashed lines indicate 95% limits of repeatability that are symmetric about zero [−33%, +33%]. (B) Differences are shown relative to single baseline value (SUV$_1$). Asymmetric limits of repeatability are marked at [−28%, +39%].

with a wide range of SUVs, the average wCV was approximately 10% irrespective of the VOI type.

Although differences were noted between the various publications, the consistency between reports was striking. Only a few papers reported a wCV of over 12%. The relatively poor repeatability observed in the study by Kumar et al. (*30*) can probably be attributed to the high variability in uptake periods, low average tumor uptake, and nonstandard definition of relative difference. Unlike other publications, the relative difference data were not calculated relative to the average (Eq. 2) but instead were expressed relative to a single baseline value (Eq. 6). Heijmen et al. (*27*) also reported a wCV of over 12%. In this case, the particular patient population could have played a role because a subset of patients received chemotherapy within 1–3 mo of PET data acquisition. When the study population was divided into those who had chemotherapy 1–3 mo before PET and those who had it more than 3 mo before PET, RC for SUV$_{max}$ dropped from 47.0% (wCV, 16.96%) to 33.3% (wCV, 12.01%). In general, the importance of standardized patient preparation (*43*) should be emphasized, including particular attention to consistent uptake times.

On the other side of the repeatability range, Rasmussen et al. (*34*) reported remarkably low variability (wCV, 4.8% for SUV$_{max}$). A possible explanation is the unusually high tumor uptake in this patient population (average SUV$_{max}$, 15.0). De Langen et al. (*28*) have shown that SUV repeatability improves with increasing tumor uptake, possibly because of a higher signal-to-noise ratio in these high-uptake regions of the image. Most of the papers included in this review did not directly address this issue, and their results reflect the average repeatability over a broad range of tumor uptake values. Neglecting potential trends within their data was understandable given the small number of data points that were typically available in each study, but a more involved analysis will probably be required to better characterize repeatability over the full range of SUVs. De Langen et al. proposed a combination of absolute and relative difference thresholds to characterize limits of repeatability. The method is flexible in that it allows for multiple combinations of absolute and relative difference cutoffs, one of which is consistent with published guidelines for tumor response assessment (*1*). Another approach involving relative difference thresholds that vary as a function of baseline SUV has also been proposed (*38*).

Interestingly, there was no clear trend toward improved repeatability as scanner technology evolved. This is perhaps surprising given the substantial improvements in PET technology that have been introduced over the past 20 y. For example, Rasmussen et al. (*34*) compared PET reconstruction with and without advanced algorithms (time of flight in combination with point spread

TABLE 3
SUV$_{max}$ Repeatability Estimates

Publication	Repeatability parameter	Parameter value	Where in original article	Inferred wCV (%)
Nakamoto (*22*)	MAPD	11.3	Table 4	10.01
Krak (*23*)	MAPD	13	Table 2	11.52
Velasquez (*25*)	wCV	11.9	Table 5	11.90
Hatt (*26*)	DSD	16.7	Page 1371	11.81
Heijmen (*27*)	RC	39	Table 1	14.08
Hoang (*29*)	MAPD	12.6	Table 2	11.17
Kumar (*30*)	RC	49	Page 177	17.69
van Velden (*31*)	MAPD	12.1	Page 17	10.72
Rockall (*32*)	RC	17.3	Table 2	6.25
Weber (*33*)	DSD	17	Table 2	12.02
Rasmussen (*34*)	wCV	4.8	Table 5	4.80
Kramer (*35*)	RC	26.6	Table 3	9.60
Mean				10.96 (SD, 3.32)

TABLE 4
SUV_{mean} Repeatability Estimates

Publication	Repeatability parameter	Parameter value	Where in original article	Inferred wCV (%)
Weber (*21*)	DSD	9.1	Table 2	6.43
Krak (*23*)	MAPD	12	Table 2	10.63
Velasquez (*25*)	wCV	11.8	Table 5	11.80
Hatt (*26*)	DSD	15.6	Table 2	11.03
Heijmen (*27*)	RC	31.2	Table 1	11.26
Hoang (*29*)	MAPD	11.4	Table 2	10.10
Kumar (*30*)	RC	44	Page 177	15.87
van Velden (*31*)	MAPD	11.8	Page 17	10.46
Rasmussen (*34*)	wCV	5.7	Table 5	5.70
Kramer (*35*)	RC	18.1	Table 3	6.53
Mean				9.98 (SD, 3.06)

function modeling) and found no improvement with the more sophisticated algorithm. They also compared repeatability between PET/CT and PET/MR—the first report to do so—and found no significant difference. Various factors are likely at play. SUV variability is greatly influenced by biologic factors that would be expected to remain unchanged irrespective of the scanner system. Also some of the high-performing early work involved dynamic data acquisition that allowed for highly controlled uptake periods and extended data acquisition, compared with the whole-body studies used in more recent studies.

In general, repeatability was similar for the various SUV types (SUV_{max}, SUV_{mean}, and SUV_{peak}) despite involving very different approaches to tumor sampling. SUV_{mean} includes much greater volume averaging than SUV_{max} but requires consistent delineation of potentially heterogeneous tumors. SUV_{peak} might appear to offer an advantageous compromise between SUV_{max} and SUV_{mean}, but the literature was not consistent on this issue. Some studies found that SUV_{peak} offered no improvement over SUV_{max} (*25,33,34*), whereas others did show an improvement (*22,23,35*). In the latter group, the use of automated software for identifying the peak region, as opposed to centering a fixed-size VOI over the maximum pixel, may have contributed to the improved repeatability. A separate issue regarding the handling of multiple tumors per patient was similarly inconclusive. Weber et al. (*33*) and Velasquez et al. (*25*) found that repeatability was similar irrespective of whether SUV was derived from a single tumor or from the average of multiple tumors. In contrast, Kramer et al. (*35*) found substantially improved repeatability when averaging the SUV from multiple tumors, albeit in a small, single-center study.

Over all the studies included in this review, tumor SUV had an average wCV of approximately 10% (10.27%), which corresponds to symmetric RCs of ±28%. These limits are in close agreement with the ±30% criterion that was previously recommended for PET tumor response classification (PERCIST (*1*)). Asymmetric limits of repeatability had not been introduced in the PET literature at the time this recommendation was published and even now have not been fully established. Nevertheless, they would seem to be appropriate for tumor response assessment with respect to a baseline measurement and should be considered for future iterations of PERCIST. Under this assumption, a wCV of 10.27% would correspond to lower and upper RCs of −25 and +33%. Of course, many of the studies included in this review had poorer repeatability than the group average, but most achieved a wCV of under 12%, which corresponds to RC limits of [−28%, +39%] (*33*).

Although these repeatability data provide a useful context for interpreting small changes in tumor SUV, broader considerations

TABLE 5
SUV_{peak} Repeatability Estimates

Publication	Repeatability parameter	Parameter value	Where in original article	Inferred wCV (%)
Minn (*20*)	MAPD	10	Table 4	8.86
Krak (*23*)	MAPD	10	Table 2	8.86
Velasquez (*25*)	wCV	12.8	Table 5	12.80
Heijmen (*27*)	RC	37.0	Table 1	13.35
Rockall (*32*)	RC	16.3	Table 2	5.88
Weber (*33*)	DSD	20	Table 2	14.14
Rasmussen (*34*)	wCV	5.7	Table 5	5.70
Kramer (*35*)	RC	19.9	Table 3	7.18
Mean				9.60 (SD, 3.40)

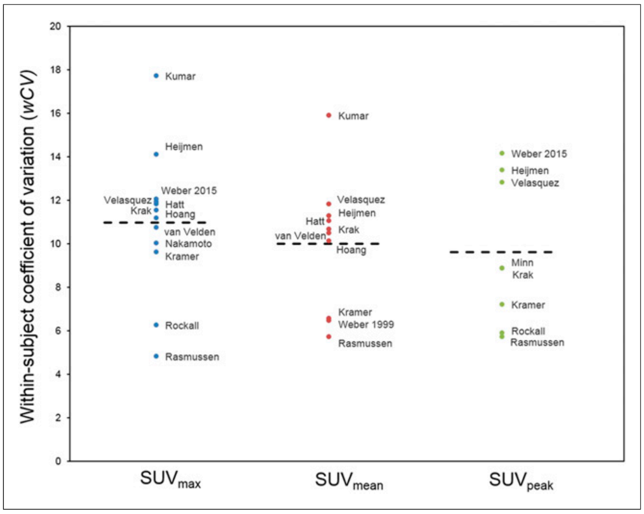

FIGURE 3. Summary of SUV repeatability results. wCV was inferred from published data and is shown separately for SUV$_{max}$, SUV$_{mean}$, and SUV$_{peak}$. Labels by each data point refer to publications noted in Table 1. Dashed horizontal lines indicate mean wCV for each SUV type: 10.96%, 9.98%, and 9.60% for SUV$_{max}$, SUV$_{mean}$, and SUV$_{peak}$, respectively.

are involved when predicting clinical outcome. For example, a tumor SUV decrease only slightly more than the limits of repeatability indicates a small treatment effect that may not be sufficient to cure the disease. The optimum change in SUV for differentiating between patients with good and bad prognoses is likely much greater than the limits of repeatability of the SUV measurement. Meignan et al. (*44*) found a 66% decrease in SUV$_{max}$ to be the optimum cutoff for identifying responders in the setting of diffuse large B-cell lymphoma after 2 cycles of chemotherapy. So although SUV repeatability limits can help distinguish between real tumor changes and measurement variability, a higher threshold is needed to best predict a successful response to treatment.

CONCLUSION

This review confirms that SUV is a highly repeatable metric for quantifying ^{18}F-FDG uptake in oncologic PET. When acquired with careful attention to protocol, tumor SUV can be measured with a wCV of approximately 10%. In a response assessment setting, tumor SUV reductions of more than 25% and increases of more than 33% are unlikely to be due to measurement variability. Broader margins may be required for sites with less rigorous protocol compliance, but in general, SUV is a highly repeatable imaging biomarker that is ideally suited to monitoring tumor response to treatment in individual patients.

ACKNOWLEDGMENT

Helpful discussions with members of the Radiological Society of North America’s Quantitative Imaging Biomarkers Alliance, FDG PET Biomarker Committee, are gratefully acknowledged.

REFERENCES

1. 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.
2. 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. *Eur J Cancer.* 1999;35:1773–1782.
3. Lin C, Itti E, Haioun C, et al. Early ^{18}F-FDG PET for prediction of prognosis in patients with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis. *J Nucl Med.* 2007;48:1626–1632.
4. Itti E, Meignan M, Berriola-Riedinger A, et al. An international confirmatory study of the prognostic value of early PET/CT in diffuse large B-cell lymphoma: comparison between Deauville criteria and ΔSUVmax. *Eur J Nucl Med Mol Imaging.* 2013;40:1312–1320.
5. Keyes JW. SUV: standard uptake or silly useless value? *J Nucl Med.* 1995;36: 1836–1839.
6. Phelps ME, Huang SC, Hoffman EJ, Selin C, Sokoloff L, Kuhl DE. Tomographic measurement of local cerebral glucose metabolic rate in humans with (F-18)2-fluoro-2-deoxy-D-glucose: validation of method. *Ann Neurol.* 1979;6: 371–388.
7. Stroobants S, Goeminne J, Seegers M, et al. ^{18}FDG-positron emission tomography for the early prediction of response in advanced soft tissue sarcoma treated with imatinib mesylate (Glivec). *Eur J Cancer.* 2003;39:2012–2020.
8. Dose Schwarz J, Bader M, Jenicke L, Hemminger G, Janicke F, Avril N. Early prediction of response to chemotherapy in metastatic breast cancer using sequential ^{18}F-FDG PET. *J Nucl Med.* 2005;46:1144–1150.
9. Avril N, Sassen S, Schmalfeldt B, et al. Prediction of response to neoadjuvant chemotherapy by sequential F-18-fluorodeoxyglucose positron emission tomography in patients with advanced-stage ovarian cancer. *J Clin Oncol.* 2005;23: 7445–7453.
10. Hutchings M, Loft A, Hansen M, et al. FDG-PET after two cycles of chemotherapy predicts treatment failure and progression-free survival in Hodgkin lymphoma. *Blood.* 2006;107:52–59.
11. Lordick F, Ott K, Krause BJ, et al. PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: the MUNICON phase II trial. *Lancet Oncol.* 2007;8:797–805.
12. Capirci C, Rampin L, Erba PA, et al. Sequential FDG-PET/CT reliably predicts response of locally advanced rectal cancer to neo-adjuvant chemo-radiation therapy. *Eur J Nucl Med Mol Imaging.* 2007;34:1583–1593.
13. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. *AJR.* 2010;195:310–320.
14. Kurland BF, Muzi M, Peterson LM, et al. Multicenter clinical trials using ^{18}F-FDG PET to measure early response to oncologic therapy: effects of injection-to-acquisition time variability on required sample size. *J Nucl Med.* 2016;57: 226–230.
15. Schwartz J, Humm JL, Gonen M, et al. Repeatability of SUV measurements in serial PET. *Med Phys.* 2011;38:2629–2638.
16. 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.
17. 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.
18. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake values. *J Nucl Med.* 2012;53:1041–1047.
19. 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.
20. Minn H, Zasadny KR, Quint LE, Wahl RL. Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET. *Radiology.* 1995;196:167–173.
21. 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.
22. Nakamoto Y, Zasadny KR, Minn H, Wahl RL. Reproducibility of common semi-quantitative parameters for evaluating lung cancer glucose metabolism with positron emission tomography using 2-deoxy-2-[^{18}F]fluoro-D-glucose. *Mol Imaging Biol.* 2002;4:171–178.
23. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA. 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.
24. Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by ^{18}F-FDG PET in malignant tumors. *J Nucl Med.* 2008;49: 1804–1808.
25. Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of ^{18}F-FDG PET in a multicenter phase 1 study of patients with advanced gastrointestinal malignancies. *J Nucl Med.* 2009;50:1646–1654.

26. Hatt M, Cheze-Le Rest C, Aboagye EO, et al. Reproducibility of ^{18}F-FDG and 3′-deoxy-3′-^{18}F-fluorothymidine PET tumor volume measurements. *J Nucl Med.* 2010;51:1368–1376.

27. Heijmen L, de Geus-Oei LF, de Wilt JH, et al. Reproducibility of functional volume and activity concentration in ^{18}F-FDG PET/CT of liver metastases in colorectal cancer. *Eur J Nucl Med Mol Imaging.* 2012;39:1858–1867.

28. de Langen AJ, Vincent A, Velasquez LM, et al. Repeatability of ^{18}F-FDG uptake measurements in tumors: a metaanalysis. *J Nucl Med.* 2012;53:701–708.

29. Hoang JK, Das SK, Choudhury KR, Yoo DS, Brizel DM. Using FDG-PET to measure early treatment response in head and neck squamous cell carcinoma: quantifying intrinsic variability in order to understand treatment-induced change. *Am J Neuroradiol.* 2013;34:1428–1433.

30. Kumar V, Nath K, Berman CG, et al. Variance of SUVs for FDG-PET/CT is greater in clinical practice than under ideal study settings. *Clin Nucl Med.* 2013;38:175–182.

31. van Velden FH, Nissen IA, Jongsma F, et al. Test-retest variability of various quantitative measures to characterize tracer uptake and/or tracer uptake heterogeneity in metastasized liver for patients with colorectal carcinoma. *Mol Imaging Biol.* 2014;16:13–18.

32. Rockall AG, Avril N, Lam R, et al. Repeatability of quantitative FDG-PET/CT and contrast enhanced CT on recurrent ovarian carcinoma: test-retest measurements for tumor FDG uptake, diameter and volume. *Clin Cancer Res.* 2014;20:2751–2760.

33. Weber WA, Gatsonis CA, Mozley PD, et al. Repeatability of ^{18}F-FDG PET/CT in advanced non-small cell lung cancer: prospective assessment in 2 multicenter trials. *J Nucl Med.* 2015;56:1137–1143.

34. Rasmussen JH, Fischer BM, Aznar MC, et al. Reproducibility of ^{18}F-FDG PET uptake measurements in head and neck squamous cell carcinoma on both PET/CT and PET/MR. *Br J Radiol.* 2015;88:20140655.

35. Kramer GM, Frings V, Hoetjes N, et al. Repeatability of quantitative whole body ^{18}F-FDG PET/CT uptake measures as function of uptake interval and lesion selection in non-small cell lung cancer patients. *J Nucl Med.* 2016;57:1343–1349.

36. 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.

37. Kamibayashi T, Tsuchida T, Demura Y, et al. Reproducibility of semi-quantitative parameters in FDG-PET using two different PET scanners: influence of attenuation correction method and examination interval. *Mol Imaging Biol.* 2008;10:162–166.

38. Bengtsson T, Sanabria-Bohorquez SM, McCarthy TJ, Binns DS, Hicks RJ, de Crespigny AJ. Statistically assigned response criteria in solid tumors (STARCIST). *Cancer Imaging.* 2015;15:9.

39. Hristova I, Boellaard R, Vogel W, et al. Retrospective quality control review of FDG scans in the imaging sub-study of PALETTE EORTC 62072/VEG110727: a randomized, double-blind, placebo-controlled phase III trial. *Eur J Nucl Med Mol Imaging.* 2015;42:848–857.

40. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135–160.

41. Geary RC. The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika.* 1935;27:310–332.

42. Lodge MA, Holdhoff M, Leal JP, et al. Repeatability of ^{18}F-FLT PET in a multicenter study of patients with high grade glioma. *J Nucl Med.* 2017;58:393–398.

43. 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.

44. Meignan M, Itti E, Gallamini A, Haioun C. Interim ^{18}F-fluorodeoxyglucose positron emission tomography in diffuse large B-cell lymphoma: qualitative or quantitative interpretation—where do we stand? *Leuk Lymphoma.* 2009;50:1753–1756.