Instrumentation factors affecting variance and bias of quantifying tracer uptake with PET/CT

R. K. Doot^{a)}

Department of Radiology, Box 357987, University of Washington, Seattle, Washington 98195

J. S. Scheuermann

Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

P. E. Christian

Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah 84112

J. S. Karp

Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

P. E. Kinahan

Department of Radiology, Box 357987, University of Washington, Seattle, Washington 98195

(Received 22 April 2010; revised 20 August 2010; accepted for publication 8 September 2010; published 29 October 2010)

Purpose: The variances and biases inherent in quantifying PET tracer uptake from instrumentation factors are needed to ascertain the significance of any measured differences such as in quantifying response to therapy. The authors studied the repeatability and reproducibility of serial PET measures of activity as a function of object size, acquisition, reconstruction, and analysis method on one scanner and at three PET centers using a single protocol with long half-life phantoms.

Methods: The authors assessed standard deviations (SDs) and mean biases of consecutive measures of PET activity concentrations in a uniform phantom and a NEMA NU-2 image quality (IQ) phantom filled with 9 months half-life ^{68}Ge in an epoxy matrix. Activity measurements were normalized by dividing by a common decay corrected true value and reported as recovery coefficients (RCs). Each experimental set consisted of 20 consecutive PET scans of either a stationary phantom to evaluate repeatability or a repositioned phantom to assess reproducibility. One site conducted a comprehensive series of repeatability and reproducibility experiments, while two other sites repeated the reproducibility experiments using the same IQ phantom. An equation was derived to estimate the SD of a new PET measure from a known SD based on the ratios of available coincident counts between the two PET measures.

Results: For stationary uniform phantom scans, the SDs of maximum RCs were three to five times less than predicted for uncorrelated pixels within circular regions of interest (ROIs) with diameters ranging from 1 to 15 cm. For stationary IQ phantom scans from 1 cm diameter ROIs, the average SDs of mean and maximum RCs ranged from 1.4% to 8.0%, depending on the methods of acquisition and reconstruction (coefficients of variation range 2.5% to 9.8%). Similar SDs were observed for both analytic and iterative reconstruction methods ($p \geq 0.08$). SDs of RCs for 2D acquisitions were significantly higher than for 3D acquisitions ($p \leq 0.008$) for same acquisition and processing parameters. SDs of maximum RCs were larger than corresponding mean values for stationary IQ phantom scans ($p \leq 0.02$), although the magnitude of difference is reduced due to noise correlations in the image. Increased smoothing decreased SDs ($p \leq 0.045$) and decreased maximum and mean RCs ($p \leq 0.02$). Reproducibility of GE DSTe, Philips Gemini TF, and Siemens Biograph Hi-REZ PET/CT scans of the same IQ phantom, with similar acquisition, reconstruction, and repositioning among 20 scans, were, in general, similar (mean and maximum RC SD range 2.5% to 4.8%).

Conclusions: Short-term scanner variability is low compared to other sources of error. There are tradeoffs in noise and bias depending on acquisition, processing, and analysis methods. The SD of a new PET measure can be estimated from a known SD if the ratios of available coincident counts between the two PET scanner acquisitions are known and both employ the same ROI definition. Results suggest it is feasible to use PET/CTs from different vendors and sites in clinical trials if they are properly cross-calibrated. © 2010 American Association of Physicists in Medicine.

[DOI: [10.1118/1.3499298](https://doi.org/10.1118/1.3499298)]

Key words: reproducibility, bias, PET, image analysis, image reconstruction

I. INTRODUCTION

Response of disease to treatment can be quantified using serial positron emission tomography (PET) scans to measure changes in concentrations of tracer uptake by diseased tissue.¹⁻⁵ PET evaluations of response have been reported to predict outcome in nonsmall cell lung cancer,^{6,7} esophageal squamous cell carcinoma,⁸ ovarian cancer,⁹ metastatic breast cancer,¹⁰ locally advanced adenocarcinoma,¹¹ and neoadjuvant locally advanced breast cancer.¹² Reported changes in uptake are due to several factors including response to therapy, natural (i.e., biologic) variability, inconsistency of patient preparation,^{1,5,13} and systemic variability in measurement from differences in hardware, reconstruction, region of interest (ROI) analysis, and subject positioning.¹³ The importance of consistent time between the injection of PET tracer and patient scan,¹⁴ and the relative merits of different reconstruction methods¹⁵⁻¹⁷ and ROI analysis techniques^{15,16,18} have also been published.

Based on the reported correlations between changes in PET measures and both pathologic response and patient outcome, a growing number of clinical trials use quantitative ¹⁸F-fluorodeoxyglucose (FDG)-PET measurements as end points.^{1,19} This increases the importance of understanding the variance and bias intrinsic to serial PET scan measurements and their impact on multicenter clinical trials. The reproducibility of PET quantification must be assessed to enable the determination of a threshold change required for classifying patient response to therapy and to aid the estimation of the impact of this variance on clinical trial design. While the test-retest reproducibility of patient imaging has been evaluated,^{15,16,20-23} the combined impacts of PET/CT hardware and subject positioning in data processing and analysis on the reproducibility of tracer uptake quantification have not been as well studied.

A NEMA NU-2 image quality (IQ) phantom study comparing the abilities of maximum standard uptake values (SUV_{max}), which is based on 1 pixel, with other activity uptake measures using multiple pixels reported the SUV_{max} frequency of errors in assessing the direction of change of 1.3% was among the lowest for “clinically likely” scenarios (range 1.2%–7.8%).²⁴ A similar result was also found from the analysis of 62 repeat baseline patient scans, where the repeatability of SUV measures were found to be similar whether based on 1 pixel (SUV_{max}) or multiple pixels (SUV_{mean} , etc.).²⁵ These reported cases of less than expected variance for SUV_{max} deserves further study.

The Netherlands protocol for standardization of FDG-PET scans in multicenter trials¹³ reported that differences in PET quantification methodology prevents combining PET measures from different centers unless every center agrees to some form of protocol standardization and cross-calibration of PET measures through common phantom experiments. Our study presented here with PET scans of the same long half-life phantom at multiple sites is one method for the suggested cross-calibration of PET measures between sites.

A set of PET/CT scans was acquired of a uniform phantom to validate a proposed method for estimating standard

deviations (SDs) of PET measures due to instrumentation factors based on differences in available coincident counts and number of pixels. To evaluate PET measurement error due only to instrumental factors, i.e., independent of dose calibrator and filling procedures, we also assessed the reproducibility of measurements of the same long half-life, non-uniform phantom on modern PET/CT scanners from each of the three manufacturers. Each scanner was located at a different PET imaging center and used local imaging protocols, except a similar reconstruction smoothing level was selected for between scanner comparisons. Five sets of PET/CT scans were acquired, reconstructed, and analyzed to determine the measurement variance and bias in repeat scans of the NEMA IQ phantom for a total of six PET/CT scan sets.

II. MATERIALS AND METHODS

Our studies used two phantoms filled with PET tracer epoxy with a 9 month half-life. A uniform phantom was scanned only at one site to validate a relationship between measurement variance and available coincident counts. A modified NEMA NU-2 IQ PET phantom was imaged repeatedly at three different PET centers. The NEMA-specified procedure is to fill the IQ phantom with a predetermined concentration of F-18 (or FDG) in water, with a concentration ratio for the small spheres to background of 4:1 and 8:1 and relatively short half-life of 110 min. This process is subject to intracenter and intercenter errors in activity measurement filling variability. To circumvent this potential source of additional bias and variability and to allow repeated scans with only minor changes in coincident count rates, we used a phantom with long-lived isotope of known concentrations for spheres and background. Thus the identically *filled* phantom was repeatedly imaged at each center. With this IQ phantom, two types of variability were studied under the hypothesis they would show increasing levels of variability:

- (1) *Repeated studies*: The phantom was in the same position on the patient table. This evaluates the variance and bias independent of any positioning effects, i.e., variations are due to instrumentation only, including the statistical noise from photon counting, and should have the minimum measured variance.
- (2) *Repositioned studies*: The phantom was repositioned between scans. These evaluated the incremental effect of random sphere positioning in relation to scanner detectors. These studies are comparable to the positioning effects (if any) that occur in patient test-retest studies.

The above studies were evaluated in detail on one scanner to evaluate the effect of independent parameters such as scan duration, number of pixels in ROIs, 2D versus 3D acquisition, reconstruction method, and smoothing. The repositioned study was then repeated on PET/CT scanners from the other two major manufacturers to evaluate intermanufacturer effects with the three repositioned studies completed within 187 d between October 2007 and April 2008.

TABLE I. NEMA NU-2 IQ phantom parameters.

Parameter	Value
Base model	NEMA NU-2 IQ phantom
Nominal interior volume (empty)	9.7 L
Nominal interior length of phantom	180 mm
Nominal sphere inner diameters (mm)	10, 13, 17, 22, 28, 37
Nominal sphere volumes (cc)	0.52, 1.15, 2.57, 5.58, 11.49, 26.52
Active epoxy volumes in spheres (cc)	0.49, 1.10, 2.39, 5.43, 11.23, 26.34
Radioactive material	$^{68}\text{Ge}/^{68}\text{Ga}$ (270.8 day ^{68}Ge half-life)
Branching ratio for ^{68}Ga	89%
Total activity assay (stated)	167.64 MBq (4.531 mCi) $\pm 10\%$
Date of assay	22 January 2007
Background concentration	16.169 kBq (0.437 μCi) $\pm 10\%$
Hot spheres concentration	64.713 kBq (1.749 μCi) $\pm 10\%$
True T/BG ratio (stated)	4.00
Net weight of epoxy/ ^{68}Ge matrix	24 lb (11 kg)

II.A. Solid source uniform phantom

A standard 20 cm diameter uniform phantom was filled by Sanders Medical (www.sandersmedical.com) with 6100 mL of $^{68}\text{Ge}/^{68}\text{Ga}$ (270.8 d half-life) in an epoxy matrix and sealed with an activity concentration of 1.7 kBq/mL at the time of the repeated PET study.

II.B. Modified NEMA NU-2 IQ solid source phantom

Under direction of the Society of Nuclear Medicine (SNM) standards validation task force, a phantom was filled by Sanders Medical with $^{68}\text{Ge}/^{68}\text{Ga}$ in an epoxy matrix and sealed.²⁶ The phantom was based on a NEMA NU-2 IQ phantom (manufactured by Data Spectrum, Durham, NC) with the central 5 cm diameter “lung” cylinder removed. In addition, the two larger hollow spheres were changed to hot spheres, as opposed to the cold spheres specified in the NEMA NU-2 instructions. Hot sphere diameters were 10, 13, 17, 22, 28, and 37 mm. The target/background ratio was 4:1. Sanders Medical measured the mass of the radioactive epoxy added to each sphere and divided by a density of 1.058 g/cc to determine the active epoxy volume present in each sphere. A summary of IQ phantom parameters is provided in Table I.

II.C. General Electric Discovery-STE PET/CT studies description

A comprehensive set of phantom studies were performed at the University of Washington Medical Center (Seattle, WA) using a Discovery-STE (DSTE) PET/CT scanner [General Electric (GE) Healthcare Technologies, Milwaukee, WI] with a 16-slice CT and a 156 mm axial PET field of view in both 2D and 3D acquisition modes.²⁷ Each PET scan was reconstructed with a slice thickness of 3.27 mm and a 128×128 reconstruction matrix grid. The transaxial field of view reconstruction diameter was 350 mm for repeated scans of a stationary phantom (pixel size 2.73 mm $\times$ 2.73 mm). For multiscanner repositioned phantom studies, the field of

view reconstruction diameter was changed to 550 mm (pixel size 4.30 mm $\times$ 4.30 mm) to minimize the differences in pixel areas used by the three scanners.

In the first set of repeated studies with a uniform phantom, three sequential dynamic scans of a stationary uniform ^{68}Ge phantom were performed on the DSTE with the duration of uniform scan time bins ($n=20$) increasing by a factor of 4 between dynamic scans from 5 to 20 to 80 min to evaluate the impact of available coincident counts on measurement variance and bias.

In the second set of repeated studies with a nonuniform phantom, one dynamic scan in 2D acquisition mode and one scan in 3D mode of the stationary IQ phantom were conducted using 20 consecutive 5 min scans. The effects of changing acquisition mode (2D or 3D) and image reconstruction method [analytic filtered back projection (FBP) or iterative ordered subsets expectation maximization (OSEM)] on image analysis variance were evaluated. The images were reconstructed using three levels of apodization (smoothing) with 7, 10, and 13 mm Hann windows for the analytical reconstructions and 7, 10, and 13 mm postreconstruction Gaussian smoothing for the OSEM reconstructions that consisted of eight iterations and 28 subsets. The smoothing filter ranges were chosen to include typical settings used clinically for whole body PET scans.

For the repositioned studies, the incremental effect of random positioning was assessed by scanning the IQ phantom 20 times (each scan for 5 min in 3D mode), with the phantom removed from the scanner and repositioned in the center of the transverse scan field of view to approximately ± 2 mm using the laser guide lines and manual adjustment of three angles of rotation. The 5 min scans were reconstructed using 3D-OSEM with eight iterations and 28 subsets. The OSEM iterations and subsets setting were selected to approximately match settings used clinically by the Siemens Biograph Hi-REZ PET/CT, described in Sec. II D, to aid comparison of results for the two scanners.²⁶ The repositioned study on the DSTE was conducted 40 d after completion of the repeated IQ phantom study.

TABLE II. PET/CT scanner, acquisition, and reconstruction (recon) parameters.

Parameter	Discovery-STE PET/CT (DSTE)	Gemini TruFlight PET/CT (TF)	Biograph Hi-REZ PET/CT (Hi-REZ)
PET trial site	University of Washington Medical Center	Hospital of the University of Pennsylvania	Huntsman Cancer Institute, University of Utah
Scanner	General Electric	Philips Healthcare,	Siemens Medical
Manufacturer	Healthcare Milwaukee, WI	Eindhoven, Netherlands	Solutions USA, Knoxville, TN
Number of CT slices	16	16	16
PET mode	3D or 2D	3D only	3D only
PET slice thickness	3.27 mm	4.075 mm	2 mm
Scan time	5 min	6 min	3 min
Image recon grid	128 × 128	144 × 144	168 × 168
Field of view diameter	550 mm, 350 mm ^a	576 mm	683 mm
Pixel size	4.30 mm, 2.73 mm ^a	4.00 mm	4.06 mm
Analytic recon	2D-FBP or 3DRP	n/a	n/a
Analytic Hann filters	7, 10, 13 mm	n/a	n/a
Iterative recon	3D-OSEM ^b	Iterative (TOF)	FORE+OSEM
Iterations, subsets	8, 28 ^b	6, 33	8, 24
Iterative recon filter	Gaussian 7, ^b 10, 13 mm	Sharp with $\lambda=1.0$	Gaussian 7 mm

^aDSTE repeated studies of the uniform and IQ phantoms were only reconstructed using a diameter of 350 and 2.73 mm square pixel.

^bDSTE repeated studies of the uniform phantom and repositioned studies of the IQ phantom were only reconstructed using 3D-OSEM with a 7 mm Gaussian filter.

II.D. Siemens Biograph Hi-REZ PET/CT study description

The repositioned studies were repeated at the Huntsman Cancer Institute at the University of Utah in Salt Lake City using a Biograph Hi-REZ PET/CT scanner with a 16-slice CT. Twenty 6 min scans were acquired in 3D mode with the position of the PET scan window varied between scans. The PET scan window was positioned over the area of interest on the CT scout scan of the phantom. The PET scan window could not be positioned in exactly the same location of the CT scout scan due to constraints in the acquisition system and the control of patient bed positioning. As a result, the phantom was not manually repositioned between scans. Instead, the variability due to combined random axial repositioning and repeat scanning was assessed. Each 6 min PET scan was reconstructed using Fourier rebinning (FORE) +OSEM with eight iterations and 24 subsets, a slice thickness of 2.0 mm, 683 mm transaxial field of view diameter, and a 168 × 168 reconstruction grid (pixel size 4.06 mm × 4.06 mm). The PET parameters, including scan duration, were selected to match current clinical settings at the Huntsman Cancer Institute.

II.E. Philips Gemini TF PET/CT studies description

The repositioned studies were repeated at the University of Pennsylvania Hospital in Philadelphia using a Gemini-16 TruFlight (TF) PET/CT scanner (Philips Healthcare, Eindhoven, The Netherlands) with a 16-slice CT. The phantom scans consisted of 20 consecutive 3 min fully 3D acquisitions. Each 3 min PET scan was reconstructed using an iterative time of flight algorithm with Philips's "sharp" reconstruction setting and with six iterations, 33 subsets, lambda equal to 1.0, a slice thickness of 4.0 mm, a 576 mm transaxial field of view diameter, and a 144 × 144 reconstruction

grid (pixel size 4.0 mm × 4.0 mm). The PET parameters, including scan duration, were selected to be similar to the clinical settings at the Hospital of the University of Pennsylvania; the only difference is that clinical studies are reconstructed with three instead of six iterations in an effort to decrease the reconstruction time.

A summary of the PET/CT scanner acquisition and reconstruction parameters for the three scanners is given in Table II. All acquisitions were corrected for attenuation, isotope decay, dead time, scattered and random coincidences, variations in detector and geometrical efficiency variations and global scale factor.

II.F. Measurements and data analysis

To study the impact of available coincident counts on measure variance, activity measurements of the uniform ⁶⁸Ge phantom from the same six PET slices were collected using concentrically centered circular ROIs with diameters of 1, 8, and 15 cm for the three dynamic scans with a factor of 4 difference in available counts between sequential scans.

To evaluate the values for the hot spheres, we manually centered 10 mm diameter ROIs on sphere centers and recorded the maximum and average activity concentrations of the central axial slice of each sphere using the corresponding manufacturer's analysis software: GE Advantage Workstation, Siemens Leonardo e.soft, and Philips Extended Brilliance Workspace. Maximum and mean values from ROIs are defined as ROI_{max} and ROI_{mean}, respectively. Measurements of hot sphere activity concentrations were normalized by dividing by the known activity concentrations (accounting for decay of ⁶⁸Ge) and the ratios were reported as recovery coefficients (RC). Maximum and mean RCs were measured using ROI_{max} and ROI_{mean}, respectively.

To check background SUV values (weight normalized), SUVs were measured using circular ROIs with 10 mm diameters located in the center of the IQ phantom in the plane of hot spheres and near the edge of the phantom at the intersection of lines bisecting the centers of the 13 and 17 mm spheres and centers of 22 and 28 mm spheres in the plane of the hot spheres.

Our figure of merit for variance is the sample SD, which is calculated by taking the square root of the sample variance. The dispersion measurement of SD was normalized in selected cases by dividing by the mean estimate and reported as coefficient of variation (COV). Assuming a random sampling model for independent, identically distributed random observations with no correlations, the variance of an average sample is equal to the ratio of the population variance to the sample count.²⁸ Taking the square root of both sides of the aforementioned equality yields the sample standard deviation equal to the ratio of the standard deviation of the population to the square root of sample count. The resulting equality suggests that the sample standard deviation is proportional to the reciprocal square root of the number of counts used for each PET measure calculation and any proportional change in the reciprocal square root of sample count will yield a proportional change in a measured sample standard deviation. If the random sampling model assumptions are valid and there are no correlations, then the ratios of SDs from two independent repeat PET measurements with similar population variances can be estimated by the reciprocal square root of the ratio of available counts from the two sets of observations to provide an estimate of changes in PET measurement error due to common user-defined PET parameters

$$\frac{s_1}{s_2} = \left(\frac{x_1 d_1 a_1 p_1}{x_2 d_2 a_2 p_2} \right)^{-1/2}, \quad (1)$$

where $s_1:s_2$ is the ratio of the first SD to the second SD, $x_1:x_2$ is the ratio of thicknesses of the transaxial PET plane, $d_1:d_2$ is the ratio of scan durations, $a_1:a_2$ is the ratio of injected activities, and $p_1:p_2$ is the ratio of the number of same-sized pixels in the same type ROIs used for the two cases. For example, the ratios of SDs from measuring the same object on the same scanner using two different injected activities could be estimated by substituting the ratio of injected activities for $a_1:a_2$ and substituting one for the other ratios in Eq. (1). The above equation does not account for the differences in individual scanner sensitivities due to factors such as crystal dead times and scatter correction. Equation (1) is not expected to be valid for predicting the proportional change in PET measurement error from the selection of either maximum or mean ROI measurements from the same ROI due to correlations between the two measures.

To validate our method for estimating SDs by available counts, we conducted three sequential dynamic scans of a stationary uniform long half-life phantom with the duration of the identical scan time bins increasing by a factor of 4 between dynamic scans. Measurements were collected from the same six contiguous PET slices for all three dynamic scans to allow paired t-test comparisons. Differences be-

tween measurement averages or standard deviations for each experimental set including either six contiguous slices for the uniform phantom or all six spheres for the IQ phantom were tested using two-tailed, paired student t-tests, which require the differences to be normally distributed. Measurements were made using the same six PET slices from the uniform phantom in each dynamic scan set to create the same number of paired measurement objects as found in the IQ phantom. This allowed the same paired t-test analysis method to be applied to both the uniform and IQ phantom studies with similar power. The assumption of normal distributions was tested using a Shapiro–Wilk test for 12 IQ phantom data sets corresponding to maximum and mean activity concentrations for the six spheres from the DSTE repeat scans ($n=20$) with analytic 3D-FBP method of 3D reprojection (3DRP) image reconstruction and a 7 mm Hann window. Statistical analyses were conducted using a combination of Excel spreadsheets (Microsoft) and JMP 5.0 (SAS Institute). Results were considered significant for p values less than 0.05.

III. RESULTS

III.A. NEMA NU-2 IQ phantom construction

Air gaps in the phantom were noticed during initial imaging, likely caused by degassing and shrinkage of the epoxy material during exothermic curing. CT and PET images of the phantom are shown in Fig. 1. The air voids located in the spheres near the filling stems likely contributed to the vendor calculated active volumes all being lower than the corresponding nominal sphere volumes in Table I. The presence of air bubbles in the smallest spheres restricted our phantom study to one PET layer using the minimum PET thickness of each scanner to avoid including confounding volumes of zero activity in measurements. The background activity in the IQ phantom ranged from 5.1 to 8.2 kBq/mL over the course of experiments at three PET centers. These phantom background activity concentrations approximate patient background concentrations in a 70 kg patient 1 h after injection by a range of 313–505 MBq (8.5–13.6 mCi) (assuming water constitutes 60% of bodyweight) for the repeated and repositioned IQ phantom studies.

III.B. Test for normal distributions

The assumption of normality was based on Shapiro–Wilk tests²⁹ with an average W statistic of 0.96 ($p \geq 0.06$) for 12 data sets corresponding to maximum and mean activity concentrations for the six spheres from the DSTE repeat scans ($n=20$) with 3DRP image reconstruction and a 7 mm Hann window. The Shapiro–Wilk tests failed to prove that the data are not from a normal distribution. This indicated that an assumption of normality was reasonable and justified the use of student's paired t-test for subsequent analyses.

III.C. Variance and bias dependence on available counts using uniform phantom

The bias and variance of the RC results are in Fig. 2, where bias [Fig. 2(a)] is the difference between the average

FIG. 1. Modified NEMA NU-2 IQ phantom. Air gaps in the main chamber, stems, and partially in the spheres themselves are evident on the magnified sagittal phantom CT images in the top row of images and less apparent in bottom row of phantom PET images with 10 cm long white and black scale bars.

measurement RC and 100% RC. Average RCs from ROI_{max} exhibited strongly increasing positive bias with increasing ROI size and decreasing scan duration while average RCs from ROI_{mean} displayed a slightly increasing negative bias with increasing scan durations in Fig. 2(a). Average RC SDs ranged from 10.8% to 0.2% (from corresponding 5 min scans from ROI_{max} with 1 cm diameters to 80 min scans from ROI_{mean} with 15 cm diameters). Note that using a 15 cm diameter ROI_{mean} from a 20 min scan of the uniform phantom corresponds approximately to the DSTE scanner calibration procedure³⁰ and yielded a SD of 0.3%. The plots of SDs versus the logarithmic scale of the scan duration appear linear in Fig. 2(b) with high correlation coefficients ($0.94 \leq r^2 \leq 0.98$) for the six data sets as predicted by Eq. (1).

Average COVs from ROI_{mean} ranged from 6.7% to 0.2% (from corresponding 5 min scans from 1 cm diameter ROIs and 80 min scans from 15 cm diameter ROIs). We did not calculate COVs from ROI_{max} due to the confounding bias effect [Fig. 2(a)].

The average number of coincident counts per PET image reconstruction of uniform phantom scans were 2.0×10^7 , 8.0×10^7 , and 3.2×10^8 for dynamic time bin durations of 5, 20, and 80 min, respectively. Based on the observed factor of 4 or 16 between the numbers of coincident counts, Eq. (1) predicts the ratio between SDs for the 5 and 20 min scans or 20 and 80 min scans is two while the ratio between SDs for the 5 min and 80 min scans is four. The average SD ratios using the same ROI type ($ROI_{max}:ROI_{max}$ or

FIG. 2. Impact of relative number of counts and circular ROI diameter and type (ROI_{max} or ROI_{mean}) on recovery coefficient (a) bias with standard deviation error bars and (b) variance only from 3D-OSEM reconstructions of stationary uniform phantom scans ($n=20$). The scan durations were shifted slightly horizontally and only one linear fit was shown in (b) for clarity.

TABLE III. Impact of ROI_{max} or ROI_{mean} definition on SDs of RCs from 80 min duration PET/CT scans of a uniform long half-life phantom (n=20).

Circular ROI diameter (cm)	Max:mean pixel ratios (unitless)	Predicted SD ratio ^a	Measured SD ratio ^b	Predicted: measured ratio
1	1:19	4.36	1.72 ^c	2.5
8	1:730	27.0	6.21 ^c	4.4
15	1:2474	49.7	10.0 ^c	5.0

^aPredicted SD ratios were calculated using Eq. (1) and were unitless.^bMeasured SD ratio is unitless mean of measured SD ratios from six contiguous slices.^cMeasured SD ratios were significantly lower than predicted SD ratios ($p < 0.0001$).

ROI_{mean}:ROI_{mean}) and size for the 5 and 20 min scans or 20 and 80 min scans for all six slices for all three ROI diameters ranged from 1.9 to 2.3, while the average SD ratio between the 5 and 80 min scans ranged from 3.8 to 4.5 with lowest and highest ratios calculated from ROI_{max} measurements. These SD ratios between the dynamic scans with different time bin duration were not significantly different from the ratios predicted by Eq. (1) ($p \geq 0.29$). Predicted and measured SD ratios from repeated scans of 80 min duration were similar when comparing SDs from concentric ROI_{mean} that had different a number of pixels ($p \geq 0.35$) but were signifi-

cantly different when using concentric ROI_{max} ($p < 0.0001$). However, Table III shows that the SDs from ROI_{max} were three to five times lower than expected from Eq. (1) from ROI_{mean} SDs from the 80 min scans based on the ratio of pixels used in ROI_{max} and ROI_{mean} calculations ($p < 0.0001$). Measured SD ratios were similar to SD ratios predicted by Eq. (1) for all scenarios ($p \geq 0.29$) except for ROI_{max}:ROI_{max} SD ratios from concentric ROIs of different diameters and for ROI_{max}:ROI_{mean} SD ratios measured from the same ROIs ($p < 0.0001$).

III.D. Variance and bias in repeated scan studies from nonuniform IQ phantom

Figure 3(a) show the maximum and mean RC from 3D-FBP and 3D-OSEM reconstructions (10 mm smoothing) versus sphere diameter for the stationary phantom imaged 20 times on the DSTE PET/CT. Figure 3(b) plots RC from ROI_{max} using 3D-OSEM with three levels of smoothing. Results using FBP and/or 2D acquisitions were similar and so numerical results for these cases are not shown.

The SDs are indicated by the error bars in Fig. 3. Numerical values of SDs and COVs are plotted in Fig. 4, which show the impact of the choice of smoothing and analysis method (e.g., ROI_{max} versus ROI_{mean}). These results have the following trends:

FIG. 3. (a) Recovery coefficients of ROI_{max} and ROI_{mean} from 3D-FBP and 3D-OSEM reconstructions versus sphere diameter for a stationary IQ phantom imaged 20 times on a DSTE PET/CT using 10 mm smoothing (n=20). (b) Maximum recovery coefficients from 3D-OSEM reconstructions using 7, 10, and 13 mm smoothing. The sphere diameters were shifted slightly for clarity.

FIG. 4. (a) Standard deviations and (b) coefficients of variation of recovery coefficients versus sphere diameter for 5 min acquisitions (n=20) of the stationary IQ phantom reconstructed with 3D-OSEM using 7, 10, and 13 mm smoothing.

TABLE IV. Standard deviations of RCs averaged over all IQ phantom spheres.

ROI analysis	Smoothing (mm)	2D-FBP (%)	2D-OSEM (%)	3D-FBP (%)	3D-OSEM (%)
Max	7	8.0	7.5	3.9	3.5
Max	10	5.7	5.1	2.7	2.4
Max	13	4.1	3.8	1.9	1.8
Mean	7	5.5	5.1	2.3	2.2
Mean	10	4.5	4.0	1.9	1.8
Mean	13	3.3	3.2	1.5	1.4

- SDs were not linearly correlated with sphere size in Fig. 4(a) ($0.05 \leq r^2 \leq 0.50$, $p > 0.11$).
- SDs increased with reduced smoothing.
- SDs from $ROI_{\max}$ were approximately 1.5–2 times higher than from ROI_{mean} .
- SDs ranged from 11.2% (2D acquisition, sharpest filter, $ROI_{\max}$) to 1.2% (3D acquisition, smoothest filter, ROI_{mean}) for individual spheres.
- Coefficients of variation [Fig. 4(b)] increased for the smaller spheres due to the reduction of the RC (Fig. 3) and ranged from 2.5% to 9.8% when COVs were averaged for all six sphere diameters with the same acquisition and reconstruction.

Since the SDs were not linearly correlated with sphere size, we computed the average standard deviation for all six spheres from 20 repeated scans for each combination of acquisition mode (2D versus 3D), reconstruction method (FBP versus OSEM), smoothing (7, 10, and 13 mm), and analysis method ($ROI_{\max}$ versus ROI_{mean}). These results are presented in Table IV, which show that:

- SDs for FBP reconstructions were slightly higher than those for OSEM, but the differences were not statistically significant (2D $p \geq 0.08$ and 3D $p \geq 0.24$).
- SDs for reconstructions of data acquired via 2D mode were significantly higher than from 3D acquisitions with the same processing parameters ($p \leq 0.008$).
- Increasing the amount of smoothing by 3 mm FWHM significantly reduced the SDs ($p \leq 0.045$).
- SDs for RCs from $ROI_{\max}$ were higher than corresponding SDs for RCs from ROI_{mean} ($p \leq 0.02$).

The ratio of relative noise for measures from $ROI_{\max}$ versus ROI_{mean} averaged for all four algorithms in Table IV decreases from 1.55 to 1.32 to 1.25 with increasing smoothing, which shows the relative noise of $ROI_{\max}$ measures decreases with smoothing.

Two general bias trends are apparent in Fig. 3 between RCs from the repeated DSTE studies. For all 12 reconstructions of the repeated DSTE scans in Table IV,

- Maximum RCs were significantly higher than mean RCs ($p \leq 0.004$) acquired using the same acquisition (2D or 3D) and reconstructed using the same method (OSEM or FBP) and smoothing (7, 10, or 13 mm).
- Incremental increases in reconstruction smoothing of 3 mm FWHM significantly decreased both maximum and mean RCs ($p \leq 0.02$) acquired using the same acquisition

(2D or 3D) and reconstructed using the same method (OSEM or FBP).

Other observed bias trends for the repeated DSTE studies were less obvious. Differences in corresponding maximum and mean RCs from OSEM and FBP RCs were statistically significantly different for the same level of smoothing for 3D acquisitions ($p \leq 0.002$) and exhibited a trend toward a difference for 2D acquisitions ($p \leq 0.07$). However, the bias between OSEM and FBP RCs may not be important in practice as observed in the similar OSEM and FBP RC curves in Fig. 3(a). There were no significant differences found between corresponding maximum RCs from OSEM reconstructions of 2D and 3D acquisitions ($p \geq 0.20$), while some significant differences were found for maximum FBP reconstructions of 2D and 3D acquisitions (range $0.003 \leq p \leq 0.14$). There were no significant differences found between corresponding mean RCs from OSEM reconstructions of 2D and 3D acquisitions or FBP reconstructions from the two acquisition modes ($p \geq 0.36$).

III.E. Variance and bias in repositioned nonuniform IQ phantom studies

The measurement of the background SUVs (ideally 1.0) for all three scanners are given in Table V for repositioned studies. The hot spheres results for the three scanners are presented in Fig. 5 as a function of sphere diameter. The RCs' behaviors as a function of diameter for the three scanners were similar. However, the corresponding maximum and mean RC curves from the repositioned studies for each scanner in Fig. 5 were almost all significantly different ($p \leq 0.02$) from each other, with the exception of the maximum RC curves for scanners TF and Hi-REZ ($p = 0.94$).

Average SDs for all sphere diameters are shown in Table VI for each of the three PET/CT scanners. The measured SDs for the three scanners were similar and, in most cases, the SDs were not significantly different for comparable met-

TABLE V. Mean $\pm$ SD SUV (g/ml) measures of background activity from 20 PET/CT scans of repositioned IQ phantom.

Scanner	Center ROI	Edge ROI	Center/edge	Average
TF	0.94 ± 0.03	0.99 ± 0.07	0.95 ± 0.05	0.96 ± 0.06
Hi-REZ	1.04 ± 0.12	1.00 ± 0.10	1.04 ± 0.14	1.02 ± 0.11
DSTE	1.03 ± 0.09	1.08 ± 0.07	0.96 ± 0.12	1.06 ± 0.09

FIG. 5. Recovery coefficients with standard deviation error bars from (a) ROI_{max} and (b) ROI_{mean} versus sphere diameter for the same IQ phantom repositioned ($n=20$) using PET/CTs from three vendors and reconstructed via 3D iterative algorithms with 7 mm postreconstruction Gaussian smoothing for the GE and Siemens scanners and Philips sharp smoothing for the Philips scanner. The sphere diameters were shifted slightly for clarity.

rics. The maximum and mean SDs for scanner DSTE measures of the repositioned IQ phantom (3.6% and 2.5%) in Table VI were not significantly higher than for corresponding values for stationary phantom scans (3.5% and 2.2%) in Table IV ($p \geq 0.39$).

IV. DISCUSSION

Using the uniform phantom, measured SD ratios were predicted by Eq. (1) for all scenarios ($p \geq 0.29$) except when comparing $ROI_{max}:ROI_{max}$ SD ratios from concentric ROIs of different diameters and when comparing $ROI_{max}:ROI_{mean}$ SD ratios from the same ROIs ($p < 0.0001$). Table III shows the SDs from ROI_{max} were three to five times lower than expected Eq. (1) when compared to SDs from ROI_{mean} from the same ROIs, suggesting there may be correlations between the maximum value pixel and other pixels within a ROI. This indicates that Eq. (1) is a valid method for estimating an unknown PET measure SD from a known SD as long as ratios of available counts between two PET scanner acquisitions, reconstructions, and image analyses are known, with the exception that SDs from ROI_{max} cannot be used to estimate SDs from ROI_{mean} and estimate the impact of changing ROI_{max} dimensions. It is interesting to note in Fig. 2(a) that the strong positive bias from ROI_{max} decreases with ROI size and the negative bias from ROI_{mean} increases slightly with scan duration. Full explanation of these behaviors, however, is beyond the scope of this study.

The modified NEMA NU-2 IQ phantom allows the measurement of key instrumentation and analysis effects in quantitative PET/CT imaging: Global scaling for activity concen-

tration and SUV, background uniformity, recovery coefficient losses due to partial volume and smoothing effects, method of analysis (e.g., ROI_{max} versus ROI_{mean}), and, in particular, the variance of these ROI measures. Knowledge of this reproducibility is essential for the proper use of PET as a quantitative assessment of change in response to therapy.

Air gaps in the spheres prevented effective calculation of the resolution losses for true spherical objects. However, it was decided the phantom could still be used to perform relative comparisons between serial images and different scanners. It should be noted that sphere wall thickness also impacts recovery coefficients. Hamill *et al.*³¹ have presented an approach using wall-less spherical sources. There are other limitations to the phantom. For example, it does not assess the impact on quantitative accuracy of scatter correction in the lung and does not include effects of cross-calibration with the dose calibrator.³⁰ The decision to add activity to the largest spheres allowed measurement of RC for larger diameters. In the NEMA NU-2 standard, these are used as cold spheres, but we believe this can lead to an underestimation, or underappreciation, of quantitation bias due to resolution loss, even for large objects.

The iterative image reconstruction parameters were chosen to have a relatively large number of iterations and subsets ((iterations $\times$ subsets ~ 200)). This allowed control of the noise/resolution tradeoffs to be dominated by the smoothing filter. It is well-known that controlling image noise by the number of subsets leads to object-dependent behavior. For example, with iterations $\times$ subsets ~ 50 , a small lesion will have a different apparent uptake next to another source of

TABLE VI. Reproducibility of SUV measures of repositioned IQ phantom.

Scanner	SD of max RC (%)	SD of mean RC (%)	Slice thickness (mm)	Scan time (min)	Background (kBq/mL)
TF	4.3	3.2	4.08	3	6.9
Hi-REZ	4.8	3.9 ^a	2.00	6	8.2
DSTE	3.6	2.5 ^a	3.27	5	5.1

^aSignificantly different ($p=0.01$).

tracer uptake than when it is by itself. Thus a limitation of this comparison is that the results will change in an object-dependent manner if the iterative algorithms are underiterated,³² as is commonly done in clinical practice.

The measurement of the background SUV accuracy for all three scanners (Table V) is consistent with the much larger study of Scheuermann *et al.*³³ The sphere RCs (Figs. 3 and 5) show the expected bias as a function of sphere diameter, reconstruction method, and analysis method. We found that the mean RCs were not significantly different for 2D versus 3D acquisitions. Subsets of these, or similar, results have been presented elsewhere, characterizing the well-known size/resolution effects on RCs.^{15,24,32,34} In this study, we were also able to systematically vary the acquisition, reconstruction, and analysis parameters, as well as obtain estimates of the SDs based on 20 repeated scans.

The roughly constant SD values, i.e., not correlated with sphere diameter, were consistent with the use of a fixed 10 mm diameter ROI. Because of this, the SDs shown in Table IV (and subsequent tables) were averaged for all sphere diameters. As a reference, a 5 min 3D-OSEM image of SUVs with 10 mm smoothing will have a noise level of approximately 1.8% for the mean SUV value and 2.4% for the maximum SUV value. Other noise values can then be estimated using Eq. (1) if the amount of smoothing is not changed. For example, if the scan is shortened to 2.5 min, the SUV noise levels will increase to approximately 2.5% (mean) and 3.4% (maximum). If SD data such as presented in Tables IV and VI are available for every PET scanner model in a multicenter imaging trial, then a designer of clinical trials could use Eq. (1) to estimate the PET measurement variance due to instrumentation factors.

While it is recognized that the noise of $SUV_{\max}$ versus SUV_{mean} as region of interest metrics are not the same, the current arguments in the PET literature often argue against the use of $SUV_{\max}$ due to a perception of excessive noise levels.^{35,36} The ratio of maximum:mean SUV noise is dependent on several factors, including ROI size, smoothing level, and noise correlations introduced by the image reconstruction algorithm. For the 10 mm ROI diameter used here, the average ratio of maximum/mean RC noise calculated from Table IV ranged from 1.25 for 13 mm smoothing to 1.55 for 7 mm smoothing. The increases in noise for the maximum pixel measurement method is relatively modest, given that there are 19 pixels in the ROI used here. This reduction in noise is most likely due to the noise correlations introduced by the reconstruction method.³² These results, plus the predicted to measured SD ratios in Table III, imply that the maximum pixel measurement method is a reasonable approach, especially given the simplicity and robustness of implementation. However, this relationship may be different for heterogeneous tumors where $ROI_{\max}$ analyses may be even more appropriate if the goal is to measure the tumor portion with the highest level of tracer uptake.

The multiscanner comparison demonstrated differences in both RCs and SDs from $ROI_{\max}$ and ROI_{mean} (Fig. 5). The significant differences in mean RCs were likely not due to the small differences in pixel size, but may be influenced by

the manufacturer's reconstruction algorithm and/or differences in more fine-grained parameter settings, such as the manner of selecting subsets. In addition, it was not possible to accurately match the reconstructed resolutions for the three scanners. Differences in SD, in general, were not statistically significant and what differences there were may be due to the residual differences in reconstructed image resolution for the three scanners. The method for repositioning the phantom between Hi-REZ scans only shifted the location in the axial direction so it is possible that the SDs reported for Hi-REZ may be slightly lower than they could have been had the phantom been manually repositioned between scans as done for the other two scanners. We note that the results shown in Figs. 3 and 5 indicate that there is no direct connection between NEMA NU-2 specified scanner resolution and the quantitative accuracy for objects even up to 37 mm diameter (26.5 cc), depending on the image processing and analysis methods.

The multiscanner comparison in Fig. 2 of Boellaard *et al.*¹³ and Kinahan *et al.*³⁷ use the same phantom design as used here although in the Boellaard study the phantom was filled with aqueous F-18. In addition, the more recent study by Fahey *et al.*³⁸ used a compromise approach with a different ACR phantom by adding Ge-68 to the hot targets and F-18 to the background. All three of those studies, however, used single measurements at multiple sites and thus the variances reported were due a combination of bias and variance from each site. In other words, the values were equivalent to a root mean square error measurement. The results presented here, employing repeated measurements of the same test object, accurately measure the three-site variability on the order of 3%–5% (Table VI), substantially lower than the values found in the multisite studies discussed above.

Studies of single-site SUV variability in patients report a test-retest SD of approximately 10%.^{20,21} Since the single-site SD due to instrumentation is on the order of 3%–5% (Table VI), the additional variability is likely a result of biological variability and imaging protocol variations. Many of these additional sources of error are reviewed elsewhere.^{4,5,39,40} Finally, we note that the increased instrumentation variability found in multicenter calibration studies^{37,38} compared to the results presented here imply that multicenter patient test-retest variances can be much higher than for single-site studies, as was recently demonstrated in the study by Velasquez *et al.*²⁵ The observed biases between the RC curves from the three scanners is a likely major contributor to the reported larger variances for multicenter imaging of the same phantom.^{37,38}

A limitation of this study, and the multicenter and patient test-retest studies discussed above, is that there is no evaluation of the long-term variability of scanner calibration. Therefore, the relative biases between the RC curves in Fig. 5 may change with every quarterly global calibration of the individual scanners and the relative absolute RC biases should not be used to compare the performance of the scanners without also considering other metrics such as the target RC divided by the background RC. A recent presentation by Lockhart *et al.*³⁰ shows that this may be an additional source

of variability of approximately 5%–10%. Our $ROI_{\max}:ROI_{\text{mean}}$ ratio relationship from phantoms may also be different for the heterogeneous tumors often observed in the clinic.

V. CONCLUSIONS

Long half-life PET calibration phantoms allow for direct comparison of quantitative results, exclusive of patient related factors, from sites using different PET scanners, acquisition protocols, and processing methods. In this study we used such phantoms to assess the variance and bias of PET measures as a function of object size, ROI definition, scan duration, acquisition mode, and image reconstruction parameters. We found that there were differences in the size-dependent biases and that even for 37 mm diameter objects (26.5 cc) there were appreciable biases, depending on the image processing and analysis methods. Conversely, there were smaller differences in the measured SDs among the three different PET/CT scanners when using approximately matched reconstruction.

More detailed evaluations on one of the scanners revealed the following main points:

- (1) $ROI_{\max}$ values have more bias than ROI_{mean} values, especially for shorter scans.
- (2) There were no significant differences in SDs for objects ≥ 10 mm in diameter when measuring a stationary or repositioned phantom when using the same reconstruction settings ($p \geq 0.39$).
- (3) The SD of a new PET measure can be estimated from a known SD using Eq. (1) as long as the ratios of available coincident counts between the two PET scanner acquisitions, reconstructions, and image analyses are known with the exception of estimating SD from $ROI_{\max}$ using known SD from ROI_{mean} or vice versa.
- (4) Measures from $ROI_{\max}$ were three to five times less noisy than might be expected from the ratio of total available coincident counts when compared to measures from ROI_{mean} , likely due to noise correlations introduced by image reconstruction. This suggests maximum PET measures may be suitable for clinical trials where the higher measure SD may be less important than the ease of making maximum PET measures provided same size ROIs with similar number of pixels are used for all maximum measurements.
- (5) SDs for analytical FBP reconstructions with matched resolution were not significantly different from those for iterative OSEM reconstructions ($p \geq 0.08$) with a reasonably large number of iterations and subsets.
- (6) The SDs due to single-site instrumentation effects over short time periods is substantially lower than reported for multisite instrumentation differences possibly due to long-term scanner calibration drift and patient and imaging protocol variability.

Knowledge and appropriate use of these instrumentation factors will allow more accurate assessment of the signifi-

cance of changes in PET tracer uptake, whether for clinical trials or patient specific evaluation of response to therapy.

ACKNOWLEDGMENTS

This work was supported in part by a SNM Student Fellowship awarded to Robert Doot, the U.S. NCI Contract No. 24XS036-004 (RIDER), and U.S. NIH under Grant Nos. CA74135, CA113941, CA115870, and CA124573. Paul Kinahan has a research contract with GE Healthcare Technologies. Joel Karp has a sponsored research agreement with Philips Healthcare. The authors also appreciate the support of Larry Clarke and Barbara Croft from the NCI Cancer Imaging Program, Alexander McEwan from the Society of Nuclear Medicine, Robert Zimmerman from the Harvard Joint Program for Nuclear Medicine, and Janet Reddin from the University of Pennsylvania. In addition, the authors thank the members of the ACRIN PET core laboratory and AAPM Task Group 145 (Quantitative PET/CT Imaging) and the RSNA Quantitative Imaging Biomarkers Alliance (QIBA) for many insightful and helpful discussions.

^{a)} Author to whom correspondence should be addressed. Electronic mail: bdoot@uw.edu

¹L. K. Shankar, J. M. Hoffman, S. Bacharach, M. M. Graham, J. Karp, A. A. Lammertsma, S. Larson, D. A. Mankoff, B. A. Siegel, A. Van den Abbeele, J. Yap, and D. Sullivan, "Consensus recommendations for the use of 18F-FDG PET as an indicator of therapeutic response in patients in National Cancer Institute Trials," *J. Nucl. Med.* **47**, 1059–1066 (2006).

²A. A. Lammertsma, C. J. Hoekstra, G. Giaccone, and O. S. Hoekstra, "How should we analyse FDG PET studies for monitoring tumour response?," *Eur. J. Nucl. Med. Mol. Imaging* **33**, 16–21 (2006).

³W. A. Weber and H. Wieder, "Monitoring chemotherapy and radiotherapy of solid tumors," *Eur. J. Nucl. Med. Mol. Imaging* **33**, 27–37 (2006).

⁴R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, "From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors," *J. Nucl. Med.* **50**, 122S–150S (2009).

⁵R. Boellaard *et al.*, "FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: Version 1.0," *Eur. J. Nucl. Med. Mol. Imaging* **37**, 181–200 (2010).

⁶M. R. MacManus, R. Hicks, R. Fisher, D. Rischin, M. Michael, A. Wirth, and D. L. Ball, "FDG-PET-detected extracranial metastasis in patients with non-small cell lung cancer undergoing staging for surgery or radical radiotherapy-survival correlates with metastatic disease burden," *Acta Oncol.* **42**, 48–54 (2003).

⁷W. A. Weber, V. Petersen, B. Schmidt, L. Tyndale-Hines, T. Link, C. Peschel, and M. Schwaiger, "Positron emission tomography in non-small-cell lung cancer: Prediction of response to chemotherapy by quantitative assessment of glucose use," *J. Clin. Oncol.* **21**, 2651–2657 (2003).

⁸H. A. Wieder, B. L. Brucher, F. Zimmermann, K. Becker, F. Lordick, A. Beer, M. Schwaiger, U. Fink, J. R. Siewert, H. J. Stein, and W. A. Weber, "Time course of tumor metabolic activity during chemoradiotherapy of esophageal squamous cell carcinoma and response to treatment," *J. Clin. Oncol.* **22**, 900–908 (2004).

⁹N. Avril, S. Sassen, B. Schmalfeldt, J. Naehrig, S. Rutke, W. A. Weber, M. Werner, H. Graeff, M. Schwaiger, and W. Kuhn, "Prediction of response to neoadjuvant chemotherapy by sequential F-18-fluorodeoxyglucose positron emission tomography in patients with advanced-stage ovarian cancer," *J. Clin. Oncol.* **23**, 7445–7453 (2005).

¹⁰F. Cachin, H. M. Prince, A. Hogg, R. E. Ware, and R. J. Hicks, "Powerful prognostic stratification by [18F] fluorodeoxyglucose positron emission tomography in patients with metastatic breast cancer treated with high-dose chemotherapy," *J. Clin. Oncol.* **24**, 3026–3031 (2006).

¹¹F. Lordick *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.* **8**, 797–805 (2007).

¹²L. K. Dunnwald, J. R. Gralow, G. K. Ellis, R. B. Livingston, H. M. Linden, J. M. Specht, R. K. Doot, T. J. Lawton, W. E. Barlow, B. F.

- Kurland, E. K. Schubert, and D. A. Mankoff, "Tumor metabolism and blood flow changes by positron emission tomography: Relation to survival in patients treated with neoadjuvant chemotherapy for locally advanced breast cancer," *J. Clin. Oncol.* **26**, 4449–4457 (2008).
- ¹³R. Boellaard, W. J. Oyen, C. J. Hoekstra, O. S. Hoekstra, E. P. Visser, A. T. Willemsen, B. Arends, F. J. Verzijlbergen, J. Zijlstra, A. M. Paans, E. F. Comans, and J. Pruim, "The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials," *Eur. J. Nucl. Med. Mol. Imaging* **35**, 2320–2333 (2008).
- ¹⁴S. Beaulieu, P. Kinahan, J. Tseng, L. K. Dunnwald, E. K. Schubert, P. Pham, B. Lewellen, and D. A. Mankoff, "SUV varies with time after injection in (18)F-FDG PET of breast cancer: Characterization and method to adjust for time differences," *J. Nucl. Med.* **44**, 1044–1050 (2003).
- ¹⁵M. Westerterp, J. Pruim, W. Oyen, O. Hoekstra, A. Paans, E. Visser, J. Van Lanschot, G. Sloof, and R. Boellaard, "Quantification of FDG PET studies using standardised uptake values in multi-centre trials: Effects of image reconstruction, resolution and ROI definition parameters," *Eur. J. Nucl. Med. Mol. Imaging* **34**, 392–404 (2007).
- ¹⁶N. Krak, R. Boellaard, O. Hoekstra, J. Twisk, C. Hoekstra, and A. Lammertsma, "Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial," *Eur. J. Nucl. Med. Mol. Imaging* **32**, 294–301 (2005).
- ¹⁷V. Bettinardi, P. Mancosu, M. Danna, G. Giovacchini, C. Landoni, M. Picchio, M. C. Gilardi, A. Savi, I. Castiglioni, M. Lecchi, and F. Fazio, "Two-dimensional vs three-dimensional imaging in whole body oncologic PET/CT: A Discovery-STE phantom and patient study," *Q. J. Nucl. Med. Mol. Imaging* **51**, 214–223 (2007).
- ¹⁸S. Larson *et al.*, "Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis," *Clinical Positron Imaging* **2**, 159–171 (1999).
- ¹⁹M. E. Juweid and B. D. Cheson, "Positron-emission tomography and assessment of cancer therapy," *N. Engl. J. Med.* **354**, 496–507 (2006).
- ²⁰H. Minn, K. Zasadny, L. Quint, and R. Wahl, "Lung cancer: Reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET," *Radiology* **196**, 167–173 (1995).
- ²¹W. A. Weber, S. I. Ziegler, R. Thodtmann, A. R. Hanauske, and M. Schwaiger, "Reproducibility of metabolic measurements in malignant tumors using FDG PET," *J. Nucl. Med.* **40**, 1771–1777 (1999).
- ²²T. Kamibayashi, T. Tsuchida, Y. Demura, T. Tsujikawa, H. Okazawa, T. Kudoh, and H. Kimura, "Reproducibility of semi-quantitative parameters in FDG-PET using two different PET scanners: Influence of attenuation correction method and examination interval," *Mol. Imaging Biol.* **10**, 162–166 (2008).
- ²³C. Nahmias and L. Wahl, "Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumors," *J. Nucl. Med.* **49**, 1804–1808 (2008).
- ²⁴J. A. Boucek, R. J. Francis, C. G. Jones, N. Khan, B. A. Turlach, and A. J. Green, "Assessment of tumour response with (18)F-fluorodeoxyglucose positron emission tomography using three-dimensional measures compared to SUV_{max}—A phantom study," *Phys. Med. Biol.* **53**, 4213–4230 (2008).
- ²⁵L. M. Velasquez, R. Boellaard, G. Kolli, W. Hayes, O. S. Hoekstra, A. A. Lammertsma, and S. M. Galbraith, "Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies," *J. Nucl. Med.* **50**, 1646–1654 (2009).
- ²⁶R. K. Doot, P. E. Christian, D. A. Mankoff, and P. E. Kinahan, "Reproducibility of quantifying tracer uptake with PET/CT for Evaluation of response to therapy," in IEEE Nuclear Science Symposium Conference Record, Honolulu, HI, 2007, Vol. M12-8, pp. 2833–2837.
- ²⁷P. Kinahan, H. Vesselle, J. Williams, C. Stearns, R. Schmitz, A. Alessio, L. MacDonald, O. Malawi, T. Turkington, S. Kohlmyer, and T. Lewellen, "Performance evaluation of an integrated PET/CT scanner: Discovery STE," *J. Nucl. Med.* **47** (Supplement 1), 392P (2006).
- ²⁸G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters* (Wiley, New York, 1978), p. 88.
- ²⁹S. Shapiro and M. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika* **52**, 591–611 (1965).
- ³⁰C. Lockhart, L. MacDonald, A. Alessio, W. McDougald, R. Doot, T. Lewellen, and P. Kinahan, "Minimizing instrument calibration error to reduce the effect of variability on PET/CT SUV measurements," *J. Nucl. Med.* **50** (Supplement 2), 235P (2009).
- ³¹J. J. Hamill, C. E. Arnsdorff, M. E. Casey, X. Liu, and W. J. Raulston, "A 68Ge PET hot-sphere phantom with no cold shells," in IEEE Nuclear Science Symposium Conference Record, 2005, 5 pp.
- ³²S. Tong, A. M. Alessio, and P. E. Kinahan, "Noise and signal properties in PSF-based fully 3D PET image reconstruction: An experimental evaluation," *Phys. Med. Biol.* **55**, 1453–1473 (2010).
- ³³J. S. Scheuermann, J. R. Saffer, J. S. Karp, A. M. Levering, and B. A. Siegel, "Qualification of PET scanners for use in multicenter cancer clinical trials: The American College of Radiology Imaging Network experience," *J. Nucl. Med.* **50**, 1187–1193 (2009).
- ³⁴E. J. Hoffman, S. C. Huang, and M. E. Phelps, "Quantitation in positron emission computed tomography: 1. Effect of object size," *J. Comput. Assist. Tomogr.* **3**, 299–308 (1979).
- ³⁵M. Soret, S. L. Bacharach, and I. Buvat, "Partial-volume effect in PET tumor imaging," *J. Nucl. Med.* **48**, 932–945 (2007).
- ³⁶M. R. Benz, V. Evilevitch, M. S. Allen-Auerbach, F. C. Eilber, M. E. Phelps, J. Czernin, and W. A. Weber, "Treatment monitoring by 18F-FDG PET/CT in patients with sarcomas: Interobserver variability of quantitative parameters in treatment-induced changes in histopathologically responding and nonresponding tumors," *J. Nucl. Med.* **49**, 1038–1046 (2008).
- ³⁷P. Kinahan, R. Doot, P. Christian, J. Karp, J. Scheuermann, R. Zimmerman, J. Saffer, and A. McEwan, "Multi-center comparison of a PET/CT calibration phantom for imaging trials," *J. Nucl. Med.* **49** (Supplement 1), 63P (2008).
- ³⁸F. H. Fahey, P. E. Kinahan, R. K. Doot, M. Kocak, H. Thurston, and T. Y. Poussaint, "Variability in PET quantitation within a multicenter consortium," *Med. Phys.* **37**, 3660–3666 (2010).
- ³⁹L. K. Shankar and D. C. Sullivan, "PET/CT in cancer patient management. Commentary," *J. Nucl. Med.* **48**, No. 1 (Supplement 1S) (2007).
- ⁴⁰R. Boellaard, "Standards for PET image acquisition and quantitative data analysis," *J. Nucl. Med.* **50**, 11S–20S (2009).