Longitudinal monitoring of reconstructed activity concentration on a clinical time-of-flight PET/CT scanner

Lawrence R. MacDonald
Amy E. Perkins
Chi-Hua Tung

Longitudinal monitoring of reconstructed activity concentration on a clinical time-of-flight PET/CT scanner

Lawrence R. MacDonald,^{a,*} Amy E. Perkins,^b and Chi-Hua Tung^b

^aUniversity of Washington, Radiology Department, 1715 Northeast Columbia Road, Box 357987, Seattle, Washington 98195-7987, United States

^bPhilips Healthcare, 595 Miner Road, Highland Heights, Ohio 44143, United States

Abstract. Positron emission tomography (PET) images are potential quantitative biomarkers. Understanding long-term (months/years) biomarker variability is important for establishing confidence intervals on studies using such biomarkers over these time frames. PET biomarkers are derived from activity concentration (ρ) extracted from PET images. Over 30 months, we measured the stability of decay-normalized counts (N_n) and ρ by scanning the same 4.5-cm-diameter Ge-68 cylinder weekly, the same Na-22 point source daily, and a refilled 20-cm F-18 cylinder phantom monthly on a clinical TOF-PET/CT scanner. Longitudinal and adjacent-measurement variability was characterized. We found no drift in ρ or N_n for properly calibrated images over 24 months. During this time, ρ_{mean} ranged $\pm 5\%$ to 6% for count-matched Ge-68 and F-18 images, with coefficient of variation (COV) across time of 2.3% (Ge-68, 81 scans) and 3.2% (F-18, 24 scans). At typical patient image count levels the Ge-68 ρ_{mean} (ρ_{max}) COV across time was 6.9% (9.6%). Changes in ρ_{mean} between adjacent F-18 scans ($\Delta\text{days} = 34$) ranged between $\pm 10\%$, with corresponding date-matched changes in Ge-68 ρ_{mean} ranging $\pm 2\%$. We recommend (1) tracking trends in N_n with image ρ as a check of quantitative data corrections/calibrations and (2) tracking both mean and COV of ρ (single time point measures) to hundredths precision using standardized uptake values. © 2016 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.4.1.011004]

Keywords: PET/CT; quantitative imaging biomarker; quality control procedure; SUV variance; quantitative imaging biomarker alliance.

Paper 16093SSR received Jun. 1, 2016; accepted for publication Oct. 28, 2016; published online Nov. 23, 2016.

1 Introduction

Quantitative accuracy and precision in positron emission tomography (PET) images are important when considering the use of PET for longitudinal and cohort studies. This is especially true for clinical trials in which multiple patients are assessed for change in disease status over time.¹ Such studies seek to minimize data variability from interexamination processes, inter/intrasubject preparation, inter/intraplatform performance, and interinstitutional methodology. Several studies have examined contributions to PET image data variability from both instrumentation and patient scanning.^{2–12} Investigators have found quantification of activity concentration measured by PET to vary by 3% or less for immediately repeated phantom scans,⁴ and by 4% to 5% over a month for a properly calibrated scanner.⁷ Proper PET scanner calibration and consistent scan protocoling require dozens of steps that can potentially affect the quantitative accuracy of image data. Recommended standards for 18F-fluorodeoxyglucose (FDG) PET scanning for oncology patients have recently been published by working groups such as the uniform protocols for imaging in clinical trials (UPICT)^{1,13} and associated quantitative imaging biomarker alliance (QIBA).¹⁴ The UPICT FDG-PET/CT protocol and QIBA FDG-PET/CT profile provide guidance for obtaining the highest quantitative accuracy and minimum variance in FDG PET/CT oncology scanning based on the existing body of knowledge drawn from the literature and expert experience. Based on

this body of knowledge, the QIBA profile contains a claim on measured changes in FDG maximum standardized uptake value (SUVmax) that specifies a within subject coefficient of variation (COV) of 10% to 12% if the profile criteria are met. The profile points to a number of important considerations that qualify applicability of the claim, including the use of SUVmax measured in “evaluable” tumors on the same PET/CT scanner. The claim is a consensus claim that has not been substantiated by studies strictly conforming to the protocol and requires reassessment for recent technology and methods that were not used in the reports on which the claim is based, such as time-of-flight (TOF) and point-spread-function modeling image reconstruction, and PET/MRI systems.

PET scanner performance is more directly and easily characterized with phantom scans than with patient scans due to the inherent biological variability of patient scanning. While ultimately it is the variability in patient scanning that is important for clinical trials, it is also important to understand the underlying variability of the scanner alone. Previous studies characterizing the stability of PET scanner image quantification have looked at calibration factors and evaluated the effects of image acquisition and reconstruction parameters.^{4,7} The use of long half-life Ga-68/Ge-68 (Ge-68) phantoms has been adopted to characterize scanner variability over time in order to eliminate the variability of refilled F-18 phantoms. Lockhart et al. reported measured activity concentration from a clinical PET/CT scanner over a period of 1 year;⁷ however, that study focused on calibration factors and did not report detailed statistics regarding

*Address all correspondence to: Lawrence R. MacDonald, E-mail: macdon@uw.edu

the consistency of the measured activity concentration; there were only 10 measurements over the 1-year test period and correlations between measured activity concentration and other parameters from the raw data, calibration factors, or ambient conditions were not studied.

The goal of this study was to monitor PET image quantification measured from both F-18 and Ge-68 phantoms on a clinical TOF PET/CT scanner over multiple years and correlate quantitative consistency with other parameters that potentially influence quantification including the number of acquired coincidence counts, calibration parameters, and ambient conditions. The image quantification in this study represents all aspects of the scanner operation, including intrinsic sensitivity, normalization and other quantitative corrections, calibrations, and image reconstruction effects.

2 Materials and Methods

In clinical PET scanning, the primary quantitative image metric is the SUV_{max} given by the maximum activity concentration measured within a region of interest (ROI) on the image, divided by the ratio of injected activity to patient mass. The injected activity is decay-corrected to the time of the image data, and patient mass is the most common of several alternative surrogates used for the patient volume into which the injected activity is distributed. Thus, errors in clinical SUV_{max} may arise from errors in measuring the injected activity, its decay correction, and patient mass (or its alternative), in addition to errors in the image activity concentration. In this work, we focused on the stability of the measured activity concentration from repeated images of the same Ge-68 phantom so that injected activity and object mass were not variable parameters and contributed no error to the measurement. During this study, all system maintenance and calibrations were performed according to manufacturer recommendations without special attention to this stability study. We correlated findings from the fixed Ge-68 phantom with similar measurements from images of refilled F-18 phantoms. Using a separate small fixed Ge-68 source, we also tracked the consistency of the dose calibrator used to assay the F-18 phantom sources.

Over a span of 30 months (2/26/13 to 9/17/15), we performed weekly scans of a Ge-68 phantom and monthly scans of an F-18 phantom. The F-18 phantom we used was a 20-cm diameter by 30-cm long fillable cylinder with a nominal injection of 74 MBq of F-18. The Ge-68 phantom was a 4.5-cm-diameter by 4.5-cm-long cylinder of isotope-embedded epoxy ("PET X-Cal System," RadQual Global Sources, Weare, New Hampshire).¹⁵ The PET X-Cal system 4.5 cm cylinder is carefully mixed for high radioisotope uniformity, and it is calibrated with an NIST-traceable source to obtain $\pm 1\%$ activity concentration value accuracy.¹⁶ This phantom began with an activity of 208 kBq/mL (on 2/26/13). For PET scanning, the 4.5-cm X-Cal cylinder was mounted inside a 20-cm diameter by 20-cm long fillable cylinder phantom that was filled with water with no radioactivity (Fig. 1); this was to provide scatter/attenuation medium to the phantom object.

The PET X-Cal system also includes a smaller 1.5 cm $\times$ 3 cm test-tube source from the same batch of Ge-68 epoxy mixture as the 4.5-cm cylinder. The smaller test-tube source was used to monitor consistency of the dose calibrator used to assay F-18 sources (CRC-25W Capintec, Inc., Ramsey, New Jersey). This source comes from the vendor with a precalibrated F-18 equivalent activity, based on relative branching ratios and abundances of decay products for the different decay mechanisms

Fig. 1 The 4.5-cm diameter by 4.5-cm long Ge-68 X-Cal phantom mounted inside a 20-cm water phantom. (a) Photograph of the phantom. (b) PET/CT scan of the phantom, coronal view. (c) Axial view and output from the XCaliper OsiriX software plugin used to automatically extract VOI data from the center of the Ge-68 phantom. VOI: $3 \times 3 \times 3$ voxels; 4 mm/voxel.

between Ge-68 and F-18. Our approximately monthly assay of this small Ge-68 source was done by the same person each time and used the F-18 setting on the dose calibrator to take a reading of activity over the course of this project.

The phantoms were scanned on a Gemini TF 64 TOF PET/CT scanner (Philips Healthcare, Cleveland, Ohio) and the data reconstructed using TOF, blob-basis ordered subset expectation maximization with 33 subsets, 3 iterations, and 14-cm TOF kernel. Resulting images had 4 mm $\times$ 4 mm $\times$ 4 mm voxels. An image analysis plugin developed at the University of Washington (XCaliper) was used with the OsiriX DICOM reader (v.5.5.2 Pixmeo, Bernex, Switzerland) to extract volume of interest (VOI) data from the Ge-68 X-Cal phantom image. This version of XCaliper was developed specifically to extract VOI data from the 4.5-cm diameter by 4.5-cm long Ge-68 phantom by automatically finding the central 27 voxels ($3 \times 3 \times 3$) in the center of the 4.5-cm object, thus, a 12 mm $\times$ 12 mm $\times$ 12 mm VOI. We extracted image data from the F-18 phantom scans using a manually drawn two-dimensional (2-D) ROI (15 cm diameter) on a center axial slice using OsiriX. 2-D ROI data from the Ge-68 X-Cal phantom scans were also extracted (32-mm diameter ROI on a central slice) using the Extended Brilliance Workstation software (version 4.0.3.5 EBW-NM 1.0 P, Philips Healthcare).

In addition to activity concentration (ρ), another measure of longitudinal stability we tracked was the number of coincidence counts (N_c) recorded by the system. This was done for the weekly Ge-68 X-Cal phantom scans and for a daily quality control (QC) scan of a Na-22 point source.

Stability analyses were performed on decay-normalized coincidence count data N_n

$$N_n(t) = \frac{N_c(t)}{N_c(0)} e^{t/\tau}, \quad (1)$$

where $N_c(t)$ represents the counts recorded from scans of the long-lived isotope sources at time t (measured in days in this study), and τ is the radioisotope decay constant. Similarly, the relative activity concentration (ρ_r) was defined as

$$\rho_r = (\rho \text{ measured from images}) / (\text{true } \rho \text{ in the phantom}). \quad (2)$$

Table 1 Scan frequency.

Source	Number of scans	Range of acq. counts	Number of days between scans				
			Min.	Max.	Median	Mean	Std. dev.
Na-22	543	6 M to 12 M	1	8	1	1.6	1.2
Ge-68	97	1.5 M to 6.4 M	1	16	7	7.6	3.0
F-18	24	67 to 80 M ^a	1	83	32	34.4	25.8

^aRange of F-18 phantom scan count data found in a subset of the scans (8 of 24). F-18 phantom scan count data were not routinely saved. We used a fixed F-18 phantom protocol of nominal 74 MBq and 3-min scan duration.

The numerator in Eq. (2) consisted of either the maximum ($\rho_{r-\max}$) or mean ($\rho_{r-\text{mean}}$) values in the ROIs.

Over the course of this study, we increased the duration of the Ge-68 X-Cal phantom scans to compensate for the decay. We edited the Ge-68 X-Cal phantom scan list-mode data to generate datasets with fixed numbers of sinogram counts and studied quantitative results from images with six count densities (100 k, 250 k, 500 k, 1 M, 2 M, 3 M counts, resulting in count densities given in Table 3). Each fractionated list-mode file was reconstructed in the same way as described above.

In addition to long-term quantitative trends, we also studied properties of changes between adjacent measurements. Examining changes between adjacent-measurement data provides information about changes that can be expected between two calibration or validation procedures. The scan frequency was different for different sources (Table 1). In comparing scan-to-scan changes between different sources, we matched the scan dates as close as possible. Thus, in comparing the change between two adjacent F-18 scans to the corresponding change in Ge-68 scans, we used the Ge-68 scan data from the days closest to the F-18 scans of interest. We used MATLAB[®] (MathWorks, Natick, Massachusetts) for the analyses of trends in longitudinal and adjacent-scan data [built-in MATLAB[®] tools for linear regression (least-squares fitting) and error metrics].

Additional parameters that we tracked included temperature from independent monitors, and several other system parameters from routine log files, such as detector amplifier settings, system timing resolution, and energy resolution. Temperature and humidity were tracked inside the scanner gantry and in the scanner room (outside of the gantry), and by independent temperature readings from 16 sensors on processing electronics boards within the gantry, recorded every 15 or 30 min depending on the sensor.

3 Results

Figure 1 shows the 4.5-cm Ge-68 X-Cal phantom inside the 20-cm cylinder filled with activity-free water. A PET/CT image of the phantom, and example output of the automatic VOI analyses software (XCaliper plugin to OsiriX) are also shown.

Table 1 gives the scan frequency statistics for each source, and the range of acquired counts of the long-lived isotopes. Figure 2 displays the values over time of relative activity concentration, $\rho_{r-\text{mean}}$, and normalized coincidence counts, N_n , for the weekly Ge-68 X-Cal phantom, daily Na-22 points source scan, and monthly F-18 phantom scan. Figure 2 annotations point to times of regular preventative maintenance (PM) and known miscalibration events. Notable events included jumps of ~3% to 6% in Ge-68 $\rho_{r-\text{mean}}$ due to an incorrect automated

quantitative adjustment applied by a particular QC protocol. This quantitative calibration factor did not affect the raw sinogram count data, and thus no jumps were observed in count data. When the proper quantitative conversion factor was applied to these data, the activity concentration in the resulting images was consistent with other scans. A data archive failure resulted in the loss of list-mode data for scans prior to October 2013, thus we could not rereconstruct those scans with the correct factors. Our analysis of quantitative stability on the scanner excluded the jumps in $\rho_{r-\text{mean}}$ that were due to this known error. Table 2 shows statistics of changes between adjacent scans, and the range and standard deviation of N_n and $\rho_{r-\text{mean}}$ values as a percentage of the mean of each respective parameter across all measurements.

The minimum adjacent-scan change for F-18 $\rho_{r-\text{mean}}$ was 0.10% (measurements 26 days apart). For the Ge-68 N_n and $\rho_{r-\text{mean}}$, and Na-22 N_n , the minima of adjacent-scan changes were $\leq 0.01\%$.

The Na-22 decay-normalized counts jumped by 2% on August 14, 2013 (Fig. 2), and remained steady for a period of 4 months, at which point there was a month of higher than usual variability, before returning to be very steady at the original decay-normalized value. There was a routine PM procedure performed on August 13 to 14, 2013; however, there was no record of hardware or software upgrade or any other system change during this time. Without including these data, there were 383 Na-22 N_n measurements, and the Table 2 Na-22 entries would be: absolute mean = 1.001, adjacent-scan maximum change of 2.45%, standard deviation 0.32%, the median and mean were nearly unchanged (0.24% and 0.31%, respectively), and N_n relative range (COV) of 98.9% to 102.2% (0.34%).

Figure 2 and Table 2 present $\rho_{r-\text{mean}}$ results from the scans as measured, without adjusting for equal counts in the images: Ge-68 phantoms scans were all 2 min in duration until May 2015, at which point scan duration was increased to 6 min because of the decay of the phantom. Results from images with equal numbers of counts are presented later (Fig. 6 and Table 3).

Due to lost list-mode data that prevented image reconstruction with the proper quantitative calibration settings, we had fewer measures of Ge-68 $\rho_{r-\text{mean}}$ ($n = 81$) than for Ge-68 N_n ($n = 131$) for the X-Cal phantom scans. Data from the same scans are presented in Table 2. Results from the entire, expanded set of Ge-68 N_n count data are very similar: %relative range (COV) = 98.0% to 103% (0.73%), no. scans = 131, range of acquired counts = 1.5 M to 12 M.

Another notable annotation in Fig. 2 is the large error bars on one or two of the F-18 phantom $\rho_{r-\text{mean}}$ measurements (error bars representing $\pm$ one standard deviation of the pixel

Fig. 2 The $\rho_{r\text{-mean}}$ from the Ge-68 and F-18 phantoms, and N_n from daily Na-22 point-source scans and from the Ge-68 scans from which $\rho_{r\text{-mean}}$ were taken. Error bars are ± 1 standard deviation of the voxel values within the VOI (Ge-68) or ROI (F-18). See text discussion of discontinuities in Ge-68 $\rho_{r\text{-mean}}$ that were due to an erroneous calibration file. Quarterly PM: quarterly PM service. In October 2015, the weekly scans of the Ge-68 phantom ended, and Na-22 data were not saved (several measurements were made in December 2015 and have been included).

Table 2 Changes in N_n and $\rho_{r\text{-mean}}$.

Source		Absolute mean ^a	Adjacent-scan change magnitude				Relative range ^c (%COV) across all measurements
			Max (# days) ^b	Median	Mean	SD	
²² Na	N_n	1.003	5.55% (1)	0.20%	0.34%	0.51%	96.4% to 102.1% (0.84%)
	N_n	0.998	1.83% (6)	0.38%	0.48%	0.41%	98.6% to 101% (0.54%)
⁶⁸ Ge ^d	$\rho_{r\text{-Vol}}$	0.966	5.84% (5)	1.13%	1.39%	1.19%	97.2% to 104% (1.42%)
	$\rho_{r\text{-ROI}}$	0.955	4.06% (5)	0.69%	0.90%	0.75%	97.9% to 102% (0.89%)
¹⁸ F	$\rho_{r\text{-ROI}}$	1.043	9.52% (27)	2.26%	3.19%	2.87%	94.4% to 104% (3.19%)

Note: COV: coefficient of variation = (standard deviation across time)/(mean across time).

SD: standard deviation of all adjacent-scan changes.

^aAbsolute mean is with respect to the initial measurement for counts (N_n) [Eq. (1)], and with respect to the known activity concentration for $\rho_{r\text{-mean}}$ [Eq. (2)].

^b(# days): the number of days between the two adjacent measurements with the maximum observed change.

^cRelative range is the minimum-to-maximum relative to the absolute mean across time.

^d⁶⁸Ge N_n and both ρ_r data here are from the same 81 scans (see text). Table 1 gives the number of scans for ²²Na and ¹⁸F data.

values in the ROI used to measure $\rho_{r\text{-mean}}$). These were explained by insufficient mixing of the radioisotope within the phantom that resulted in a nonuniform activity concentration. Figure 3 shows a typical image slice from a uniform F-18 phantom, and an image of a phantom that was not well mixed (annotated in Fig. 2). An interesting phenomenon here is that the $\rho_{r\text{-mean}}$ successfully met the $\pm 10\%$ accuracy requirement imposed on the mean activity concentration measurement for this QC procedure.

The system average energy resolution, amplifier gains, and temperatures are shown in Fig. 4 as examples of system parameters that we tracked over the duration of the project. Other parameters from the QC log files showed similar behavior to those seen in Fig. 4: occasional small discontinuities in system

parameters were present, only sometimes correlated with scanner service/maintenance, but the discontinuities did not correlate with each other, or with quantitative image metrics (N_n , ρ_r).

The system temperature was tracked by monitors placed in the imaging suite (ambient-external) and inside the scanner gantry (ambient-internal), as well as by sensors placed directly on several of the acquisition and processing electronics circuit boards (DAQ boards¹). Ambient temperatures were very steady within a range of 2°C, and in-step between all sensors. Temperatures on the digitizer boards were 2°C to 3°C higher than the ambient temperatures and drifted upward $\sim 4^\circ\text{C}$ over 2+ years [Fig. 4(c)]. On February 11, 2015, the daily QC procedure reported a warning that one or more components did not meet specification limits. The components were identified as drifting low voltage power supplies and

Table 3 Mean of ρ_r across time (24 months, 81 scans) in the 4.5-cm Ge-68 X-Cal phantom.

Kcnt/mL	$\rho_{r\text{-mean}}$					$\rho_{r\text{-max}}$				
	Mean	CoV	Min ^a	Max ^a	Max change (#days)	Mean	CoV	Min ^a	Max ^a	Max change (#days)
1.4	0.967	6.9%	81%	119%	27.4% (6)	1.215	9.6%	76%	129%	39.1% (6)
3.5	0.973	4.1%	88%	111%	14.7% (12)	1.134	5.6%	89%	115%	27.6% (12)
6.9	0.969	3.0%	93%	106%	9.2% (7)	1.081	4.5%	90%	112%	16.7% (7)
13.8	0.968	2.3%	95%	106%	8.4% (6)	1.048	3.3%	93%	108%	10.9% (6)
27.1	0.968	1.9%	96%	106%	5.7% (5)	1.027	2.3%	95%	110%	11.2% (5)
42.7	0.968	1.6%	97%	105%	5.7% (5)	1.020	2.1%	96%	105%	5.6% (5)

Note: COV: coefficient of variation = (standard deviation across time)/(mean across time).
Max change: maximum absolute value change (%) in ρ_r between any two successive measurements; number of days between the max change is given in parentheses.
Data are from 12 mm $\times$ 12 mm $\times$ 12 mm VOIs.
^aMin and Max are given as a percentage of the mean.

Fig. 3 (a) Typical central slice image of a QC scan of the 20-cm-diameter uniform F-18 phantom. (b) The same QC phantom was scanned after insufficient mixing of the radioisotope within the phantom volume, leading to the nonuniform activity distribution seen here. Despite the nonuniform radioisotope distribution, the measured $\rho_{r\text{-mean}}$ was within acceptable range ($0.9 < \rho_{r\text{-mean}} < 1.1$) for this QC test. (c) COV ($\text{COV} = \text{SD}/\text{mean}$) for individual measurements within VOIs and ROIs for each Ge-68 and F-18 $\rho_{r\text{-mean}}$ shown as error bars in Fig. 2.

were corrected by the service engineer, which restored DAQ temperatures to the original values. Prior to performing this repair, while the DAQ board temperatures were elevated, we scanned the Ge-68 X-Cal phantom and obtained a $\rho_{r\text{-mean}}$ value that was within one standard deviation of the overall mean across all dates. A temperature spike on June 1 to 2, 2014 [Fig. 4(c) inset] was noticed retrospectively; counts from the daily QC Na-22 scan resulted in $N_n = 1.00$ on that day, but no phantom scan was performed on that day.

The results from assays of the smaller test-tube sized Ge-68 source are shown in Fig. 5. The dose calibrator Ge-68 source had a low activity relative to a patient dose because it is a very small volume drawn from the same batch of Ge-68 infused epoxy as the 4.5-cm PET scanner phantom. While the 4.5-cm PET scanner phantom has a high activity concentration relative to the activity concentration typically found in a patient, the small volume of the dose calibrator source results in a small total activity. The difference between measured and expected activity in this source can be seen to increase over time. During the last

months of the project, we noted the background baseline reading of the empty dose calibrator prior to each assay, but not in a systematic way. We observed the background to be <4 kBq.

In Fig. 2, results were presented as-as-measured, which were not normalized to have the same number of counts to control for Poisson noise. Results from Ge-68 phantom images reconstructed using six different fixed count density levels (sinogram counts per unit active volume) are shown in Fig. 6 and Table 3. For comparison, the count density in the F-18 phantom scans was 12 to 14 kcnt/mL, and, we estimate typical patient scan count density of ~ 1 to 2 kcnt/mL from this scanner.

Linear regression [Fig. 6(a); least-squares fitting] suggested no significant change of $\rho_{r\text{-mean}}$ over time. Regression results were similar for $\rho_{r\text{-max}}$, although closer to suggesting an increasing trend: at 42.7 kcnt/mL: $\Delta\rho_{r\text{-max}}/\Delta t = 0.83\%/yr$, 95% confidence interval (CI) = $[-0.08, 1.73]$, $p\text{-value} = 0.073$, $R^2 = 0.040$.

The adjacent-scan changes in N_n or $\rho_{r\text{-mean}}$ for matched dates were not correlated across the different sources [Fig. 7(a)].

Fig. 4 System and ambient conditions across time: (a) Average 511 keV% FWHM energy resolution. (b) Scanner detectors amplifier gain adjustments. (c) Temperature on scanner acquisition boards (DAQ) and two independent measurements of ambient temperature (internal and external to the system gantry). The inset shows a close correlation among the three independent temperature sensors during an abrupt increase of 5 °C over about 12 h on June 2, 2014. The dew point and relative humidity were also monitored in the imaging suite: dew point trends followed temperature very closely, and relative humidity was highly variable between 0% and 60% seemingly randomly.

Fig. 5 Dose calibrator measurements. The solid curve represents the theoretical Ge-68 activity as it decays (half-life = 270.8 days) relative to the first measurement. Square markers show assay measurements (right-axis). Diamond markers show the percentage error (left-axis) between the measured and expected activity.

Fig. 6 (a) Linear regression results for the highest count level X-Cal Ge-68 phantom $\rho_{r\text{-mean}}$ over time. (b) X-Cal Ge-68 phantom $\rho_{r\text{-mean}}$ and $\rho_{r\text{-max}}$, averaged across dates, as a function of count density [count density = (sinogram counts)/(phantom active volume)]. The $\rho_{r\text{-mean}}$ of 0.97 (Table 3) represents a 3% bias with respect to the activity concentration quoted by the phantom manufacturer; the PET scanner was calibrated to obtain $\rho_{r\text{-mean}} = 1.00$ in an F-18 phantom. CI, confidence interval.

Likewise, correlations were weak between matched adjacent-scan changes from Ge-68 N_n , $\rho_{r\text{-mean}}$ (VOI) and $\rho_{r\text{-mean}}$ (ROI). There was a moderate correlation between Ge-68 VOI $\rho_{r\text{-mean}}$ and $\rho_{r\text{-max}}$ (correlation coefficient = 0.57). Adjacent-scan changes for F-18 had a wide range of values [−10% to +8%, Fig. 7(b)]. The corresponding changes in the date-matched Ge-68 $\rho_{r\text{-mean}}$ and Na-22 N_n were much smaller, as shown in Fig. 7(b). While Fig. 7 compares changes of date-matched scan subsets from different sources, Table 2 gives statistics for adjacent-scan changes for all scans for each source.

Data in Table 3 are a subset of Tables 1 and 2 data due to insufficient counts, or the loss of certain list-mode data files that prevented us from performing sinogram count filtering. Further, Tables 1 and 2 data are a subset of all acquired data

(Fig. 2), again due to lost list-mode data that prevented image reconstruction with the proper quantitative calibration file.

4 Discussion

We monitored the stability of quantitative PET image metrics on a single clinical TOF-PET/CT scanner using fixed long-lived sources over the course of 30 months and found no drift in bias of either coincidence counts acquired (N_n) or mean activity concentration ($\rho_{r\text{-mean}}$). Data with known errors in the quantitative calibration were either retrospectively corrected or omitted from the stability analysis if correction was not possible. The variance of $\rho_{r\text{-mean}}$ depended on the count-density within the images, as did the bias of $\rho_{r\text{-max}}$ [Fig. 6(b), Table 3], but the bias of $\rho_{r\text{-mean}}$ was not count-dependent. The COV across time

Fig. 7 (a) The lack of correlation between changes in matched pair measurement from the three sources. For Na-22 versus Ge-68: Na-22 N_n data were taken from days with a Ge-68 phantom scan; for F-18 versus Ge-68: Ge-68 ROI data were taken from days closest to F-18 phantom scans. (b) Frequency histogram of the changes in adjacent measurements for F-18 $\rho_{r\text{-mean}}$, and corresponding changes in the date-matched Ge-68 $\rho_{r\text{-mean}}$ (ROI), and Na-22 N_n data. Different bar widths reflect different binning used to capture different data ranges.

of $\rho_{r-\max}$ was 2.1%, only slightly larger than that of $\rho_{r-\text{mean}}$ (COV = 1.6%) for the highest count densities (>25 kcnt/mL in this study, Table 3). We estimated the image count density in a typical oncology patient scanning to be ~1 to 2 kcnt/mL for this scanner. At these count levels, COV across time in Ge-68 phantom $\rho_{r-\text{mean}}$ and $\rho_{r-\max}$ was 6.9% and 9.6%, respectively. Further, at count-densities less than ~5 to 10 kcnt/mL, $\rho_{r-\max}$ bias increased considerably with decreased count-density, thus adding further variability to SUV_{max}, which is the most common image biomarker taken from PET scanning.

The overall range of $\rho_{r-\text{mean}}$ values for count-density-matched F-18 and Ge-68 images was about the same at $\pm 5\%$ to 6% each (Tables 2 and 3, count density in F-18 images was 12 to 14 kcnt/mL). The COV was smaller for Ge-68 (2.3%, 81 scans) than for F-18 (3.2%, 24 scans), and these relationships held for the subset of Ge-68 scans that were date-matched to the F-18 scans. Despite moderate overall range, we observed large scan-to-scan fluctuations in the monthly F-18 phantom scans, reaching 9.5% difference between adjacent scans, which was considerably larger than scan-to-scan variations in date-matched Ge-68 phantom scans (Fig. 7). The F-18 phantom data should capture typical variability of clinical QC procedures because they were performed as part of routine maintenance and service by the department staff (certified nuclear medicine technologist) and vendor field service engineer, without particular attention to this stability study.

The scanner stability results are encouraging for the use of PET in longitudinal studies in which patients may be scanned over the course of months or years. However, the use of $\rho_{r-\max}$ to calculate SUV is more susceptible to variability and bias than $\rho_{r-\text{mean}}$, especially at image count levels typical of current clinical PET scanning; we found stable $\rho_{r-\max}$ only for significantly higher image count densities [Fig. 6(b)]. These results were found for the typical clinical image reconstruction protocol used in this study. We would expect similar results for alternative image reconstruction parameters consistently applied to all data. Previous studies have shown the importance of using consistent image reconstruction parameters, as well as consistent image acquisition protocols, because absolute quantification values vary with these parameters.⁴

While no quantitative drifts were measured, we did see measurement outliers and acute events due to operator and system errors. PET scan quantification is susceptible to acute events because of the many steps required to maintain quantitative accuracy.¹⁷

One of the acute events we observed (operator error) resulted from insufficient mixing of F-18 radioisotope in the QC phantom. The $\rho_{r-\text{mean}}$ happened to meet specifications for this QC scan, but it was clear that the image was not acceptable from a uniformity standpoint [Fig. 3(b)]. Staff training and experience should be sufficient to catch such a qualitative error; however, it is also straightforward to record the standard deviation of pixel values within an ROI and impose quantitative “acceptance limits” on the standard deviation, as is done with the ROI mean (e.g., $0.9 < \text{SUV}_{\text{mean}} < 1.1$). Based on our results for the Gemini TF system tested here, a COV above 10% (5%) would suggest unusual nonuniformity for the F-18 (Ge-68) phantom scans [Fig. 3(c)]. A lower limit might also be imposed because an artificially low COV could indicate something wrong in the imaging process. Limits on COV would reduce qualitative subjectivity in assessing uniformity. In our case, the error was due to a clinical staff member with limited experience with PET

scanning, and in particular, filling and scanning a phantom. The use of a solid Ge-68 QC phantom is an alternative to refilled F-18 phantoms. Considerations in the choice of QC phantom include the cost of a solid Ge-68 phantom versus the cost of multiple F-18 doses over the useful lifetime of a Ge-68 phantom, practical phantom size, and the benefits of using larger phantoms for QC, the time and convenience of not filling a phantom, and the observed bias in Ge-68 phantom $\rho_{r-\text{mean}}$ [bias = -3% in this study (Tables 2 and 3)].

The second acute system error during our study was an erroneous quantitative correction leading to the ~3% to 5% jump in ρ_r (Fig. 2) due to an outdated calibration procedure, intended for a different detector configuration, that had not been removed from this system. A 3% to 5% change could easily be masked when using generally accepted SUV QC limits of $\pm 10\%$ in refillable F-18 phantoms. Our scan-to-scan variability results for properly calibrated F-18 scans showed larger changes than the 3% to 5% introduced by the miscalibration. This demonstrates an advantage of using the Ge-68 phantom because the scan-to-scan and overall long-term variances were lower, which would assist in identifying acute errors. Correlating N_n with $\rho_{r-\text{mean}}$ for Ge-68 allowed us to identify the erroneous calibration. A limitation of this study was that F-18 acquired count data were not retained, thus we could not correlate counts to ρ_r for the F-18 data.

The jump in Na-22 N_n on August 14, 2013, and subsequent return to initial value 4 months later remains unclear. Service records show no changes to the system, and no other daily QC parameter changed. The daily Na-22 QC procedure can be sensitive to the exact placement of the Na-22 point source in the scanner field of view. The system QC software checks the source location to within a finite tolerance of a small number of millimeters. It is possible that the daily placement that was used consistently until August 13, 2013, was adjusted slightly to a new daily position, and then readjusted back to the original position. These jumps were not excluded from Table 2 statistics because they could not be discounted as a known adjustment or error (although the abrupt increase and eventual return to original trend suggests something systematic). As given in the text below Table 2, the Na-22 N_n variability was considerably smaller when these data were excluded.

Another potential acute system error was avoided by the routine daily QC protocol. The temperature of the DAQ boards drifted slowly upward by 4°C over 24 months, which triggered an inspection of a potentially failing component (low-voltage power supply) that was readjusted in this case, restoring normal system parameters (Fig. 4). The warning came before any effect on image quantification was seen, as the $\rho_{r-\text{mean}}$ measured prior to corrective action on the day of the warning was within one standard deviation of the overall mean.

We tracked adjacent-scan changes to see if they were smaller than the overall range of observed changes, and thus able to provide a more precise QC target or requirement. For all three sources, we observed the maximum adjacent-scan change to be roughly equal to the full range of values across the entire period of the project (Table 2); e.g., the maximum adjacent-scan change in F-18 $\rho_{r-\text{mean}}$ was close to 10%, which amounts to the two adjacent scans spanning the $\pm 5\%$ (relative to overall mean) full range of F-18 $\rho_{r-\text{mean}}$ values observed. Thus, based on our observations, QC on $\rho_{r-\text{mean}}$ quantification must allow for changes between adjacent scans equal to the full range of values observed for the scanner. Quantitative changes between adjacent

scans for different sources were not correlated, and were weakly correlated for differing metrics from the same Ge-68 source (Fig. 7). This suggests that the changes were not systematic, but rather inherently random in nature, likely due to counting statistics for N_n , with slightly higher variability for Ge-68 ρ_r due to the added image reconstruction step, and even higher for refilled F-18 phantoms (Tables 2 and 3).

We recorded Ge-68 $\rho_{r\text{-mean}}$ using a 2-D ROI upon completion of the scan as an immediate check of consistency at a basic level. The three-dimensional (3-D) VOI data were extracted later in an off-line process. Our results showed a slight difference between 2-D-ROI and 3-D-VOI results (Table 2) that can be attributed to the differing region definition, and perhaps also the use of different software. This was not intended as a comparison between software packages. A systematic comparison between quantitative image metrics extracted from DICOM images using different software packages has been reported.¹⁸

It was not clear whether assaying the dose-calibrator Ge-68 test tube source from the PET X-Cal system provided quantitative consistency information beyond the routine QC for the dose calibrator. Figure 5 shows that the dose calibrator was very consistent over the course of this project. It is possible that the equivalent information could be found from the manufacturer-recommended daily QC. The low initial activity (<1 MBq) of the dose calibrator Ge-68 source shortened its valuable lifetime.

Based on the results of this study, we recommend the following steps for monitoring long-term quantitative consistency on PET/CT scanners:

1. Track trends for both acquired counts in a QC scan, as well as the activity concentration extracted from the corresponding reconstructed image: only small, random deviations from expected trends should be seen; anything otherwise could signify a miscalibration.
2. Track average SUV to within hundredths, and also track the standard deviation within the ROI used to calculate mean SUV. If the scanner software calculates SUV, it must be possible to enter accurate values of the active phantom mass and net activity in the phantom. Phantoms are typically much less massive than patients, and scanner software may not save lower phantom mass to sufficient precision for accurate tracking of SUV.
3. Use a Ge-68 phantom to perform quantification QC: our results and others⁷ show substantially better quantitative precision when using a fixed phantom in place of refilled F-18 phantoms. The increased precision could enable identification of smaller quantitative miscalibrations.

5 Conclusions

Correlating variations in phantom SUV with measured counts can help identify miscalibration of SUV. Extremely consistent quantitative stability, demonstrated by negligible quantitative drift over 2 years shown in this study using a standard clinical TOF-PET/CT protocol, suggests that the quantitative uncertainty due to the scanner alone can be significantly smaller than uncertainty expected from inherent biological variability in patients.⁹ However, for the image count-density typical of

oncology scans, and for the most common quantitative metric, SUVmax, we found COV across time of 9.6%, which approaches the 10% to 12% within subject COV found by the QIBA FDG-PET/CT technical committee.¹⁴ The many steps needed to obtain quantitative accuracy in PET/CT images requires careful monitoring of system performance; we recommend: (1) tracking trends in acquired counts with image activity concentration as a check of quantitative data corrections/calibrations and (2) tracking both the mean and standard deviation of ROI activity concentration to hundredths precision using standardized uptake values (SUV). Based on our consistency study, achieving SUV range limits smaller than $1.00 \pm 10\%$, ($\pm 5.0\%$) and an ROI standard deviation <10% (<5%) is obtainable in routine QC tests using a re-filled F-18 (fixed Ge-68) phantom with typical acquisition and processing protocols.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

We gratefully thank Hwi Kim of Philips Healthcare for his assistance during this study, and Thomas Laurence and Jerome Griesmer at Philips Healthcare for providing valuable guidance and feedback during the study. This study was supported by research grants from Philips Healthcare and the U.S. NIH Grant No. U01-CA148131.

References

1. M. M. Graham et al., "Summary of the UPICT protocol for 18F-FDG PET/CT imaging in oncology clinical trials," *J. Nucl. Med.* **56**, 955–961 (2015).
2. L. M. Velasquez et al., "Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies," *J. Nucl. Med.* **50**, 1646–1654 (2009).
3. C. R. Meyer et al., "Quantitative imaging to assess tumor response to therapy: common themes of measurement, truth data, and error sources," *Transl. Oncol.* **2**, 198–210 (2009).
4. R. K. Doot et al., "Instrumentation factors affecting variance and bias of quantifying tracer uptake with PET/CT," *Med. Phys.* **37**, 6035–6046 (2010).
5. V. Frings et al., "Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer," *J. Nucl. Med.* **51**, 1870–1877 (2010).
6. F. H. Fahey et al., "Variability in PET quantitation within a multicenter consortium," *Med. Phys.* **37**, 3660–3666 (2010).
7. C. M. Lockhart et al., "Quantifying and reducing the effect of calibration error on variability of PET/CT standardized uptake value measurements," *J. Nucl. Med.* **52**, 218–224 (2011).
8. J. Schwartz et al., "Repeatability of SUV measurements in serial PET," *Med. Phys.* **38**(5), 2629–2638 (2011).
9. V. Kumar et al., "Variance of SUVs for FDG-PET/CT is greater in clinical practice than under ideal study settings," *Clin. Nucl. Med.* **38**, 175–182 (2013).
10. N. E. Makris et al., "Evaluation of strategies towards harmonization of FDG PET/CT studies in multicenter trials: comparison of scanner validation phantoms and data analysis procedures," *Eur. J. Nucl. Med. Mol. Imaging* **40**(10), 1507–1515 (2013).
11. R. K. Doot et al., "Biases in multicenter longitudinal PET standardized uptake value measurements," *Transl. Oncol.* **7**(1), 48–54 (2014).
12. W. A. Weber et al., "Repeatability of 18F-FDG PET/CT in advanced non-small cell lung cancer: prospective assessment in 2 multicenter trials," *J. Nucl. Med.* **56**, 1137–1143 (2015).
13. FDG-PET/CT, UPICT, 2013, http://qibawiki.rsna.org/images/d/de/UPICT_Oncologic_FDG-PETCTProtocol_6-07-13.pdf (16 November 2016).

14. FDG-PET/CT Technical Committee, "FDG-PET/CT as an imaging biomarker measuring response to cancer therapy, quantitative imaging biomarkers alliance, version 1.05," publicly reviewed version, <http://www.rsna.org/qiba/> (11 December 2013).
15. R. K. Doot et al., "Early experiences in establishing a regional quantitative imaging network for PET/CT clinical trials," *Magn. Reson. Imaging* **30**(9), 1291–1300 (2012).
16. B. E. Zimmerman and J. T. Cessna, "Development of a traceable calibration methodology for solid (68)Ge/(68)Ga sources used as a calibration surrogate for (18)F in radionuclide activity calibrators," *J. Nucl. Med.* **51**(3), 448–453 (2010).
17. P. E. Kinahan et al., "PET/CT assessment of response to therapy: tumor change measurement, truth data, and error," *Transl. Oncol.* **2**, 223–230 (2009).
18. L. A. Pierce et al., "A digital reference object to analyze calculation accuracy of PET standardized uptake value," *Radiology* **277**(2), 538–545 (2015).

Lawrence R. MacDonald is a research associate professor in the Department of Radiology at the University of Washington. His areas of research involve high-resolution nuclear medicine technologies and applications. He has experience in academia and industry developing preclinical and application-specific SPECT/CT, PET, and imaging probe systems. His current research focuses on the use of PET imaging as a quantitative biomarker of disease. He received his PhD in physics at UCLA.

Biographies for the other authors are not available.