Measuring temporal stability of positron emission tomography standardized uptake value bias using long-lived sources in a multicenter network

Darrin Byrd
Rebecca Christopfel
Grae Arabasz
Ciprian Catana
Joel Karp
Martin A. Lodge
Charles Laymon
Eduardo G. Moros
Mikalai Budzevich
Sadek Nehmeh
Joshua Scheuermann
John Sunderland
Jun Zhang
Paul Kinahan

Measuring temporal stability of positron emission tomography standardized uptake value bias using long-lived sources in a multicenter network

Darrin Byrd,^a Rebecca Christopf,^a Grae Arabasz,^b Ciprian Catana,^b Joel Karp,^c Martin A. Lodge,^d Charles Laymon,^e Eduardo G. Moros,^f Mikalai Budzevich,^f Sadek Nehmeh,^g Joshua Scheuermann,^c John Sunderland,^h Jun Zhang,ⁱ and Paul Kinahan^{a,*}

^aUniversity of Washington, Department of Radiology, Seattle, Washington, United States

^bAthinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Charlestown, Massachusetts, United States

^cUniversity of Pennsylvania, Department of Radiology, Philadelphia, Pennsylvania, United States

^dJohns Hopkins University, Department of Radiology and Radiological Science, Baltimore, Maryland, United States

^eUniversity of Pittsburgh, Presbyterian University Hospital, Department of Radiology, Pittsburgh, Pennsylvania, United States

^fMoffitt Cancer Center, Tampa, Florida, United States

^gWeill Cornell Medical College, Department of Radiology, New York, United States

^hUniversity of Iowa, Department of Radiology, Iowa City, Iowa, United States

ⁱThe Ohio State University, Department of Radiology, Columbus, Ohio, United States

Abstract. Positron emission tomography (PET) is a quantitative imaging modality, but the computation of standardized uptake values (SUVs) requires several instruments to be correctly calibrated. Variability in the calibration process may lead to unreliable quantitation. Sealed source kits containing traceable amounts of $^{68}\text{Ge}/^{68}\text{Ga}$ were used to measure signal stability for 19 PET scanners at nine hospitals in the National Cancer Institute's Quantitative Imaging Network. Repeated measurements of the sources were performed on PET scanners and in dose calibrators. The measured scanner and dose calibrator signal biases were used to compute the bias in SUVs at multiple time points for each site over a 14-month period. Estimation of absolute SUV accuracy was confounded by bias from the solid phantoms' physical properties. On average, the intrascanner coefficient of variation for SUV measurements was 3.5%. Over the entire length of the study, single-scanner SUV values varied over a range of 11%. Dose calibrator bias was not correlated with scanner bias. Calibration factors from the image metadata were nearly as variable as scanner signal, and were correlated with signal for many scanners. SUVs often showed low intrascanner variability between successive measurements but were also prone to shifts in apparent bias, possibly in part due to scanner recalibrations that are part of regular scanner quality control. Biases of key factors in the computation of SUVs were not correlated and their temporal variations did not cancel out of the computation. Long-lived sources and image metadata may provide a check on the recalibration process. © 2018 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.5.1.011016]

Keywords: quantitative positron emission tomography imaging; calibration; standardized uptake value; clinical trials.

Paper 17202SSR received Jul. 7, 2017; accepted for publication Nov. 20, 2017; published online Jan. 4, 2018; corrected Mar. 23, 2019.

1 Introduction

The search for new cancer therapies requires costly clinical trials, many of which do not succeed. In particular, the efficacy of new pharmaceuticals must be demonstrated in patient populations, a process that consumes time and resources and frequently does not yield an effective product.¹ Positron emission tomography (PET) combined with x-ray computed tomography (PET/CT) can aid clinical trials because for many cancers, changes in PET image values such as the standardized uptake value (SUV) can predict therapy efficacy quickly after a treatment is administered.^{2–4} In principle, this allows more efficient discovery of effective treatments through more informed and adaptive trial management.

However, bias in PET images may reduce the predictive power of SUVs, especially if biases are not constant.⁵ Sources of bias in SUVs are numerous and well documented, potentially arising from software settings, hardware calibration, clinical protocols, patient variability, and human error.⁶ SUVs are

computed as shown in Eq. (1), where A is the radioactivity concentration measured in the image, d' is the decay-corrected quantity of radioactivity that was injected into the patient, and V is some surrogate for the dose volume, such as body weight

$$\text{SUV} = \frac{A}{d'V}. \quad (1)$$

The primary instrumentation factors affecting SUV accuracy are biases in A , which is measured by the PET scanner, and in d' , which is measured by a dose calibrator. It has been shown that biases due to instrumentation and clinical factors are more stable at single sites versus multicenter consortiums.^{7,8} However, while limiting measurements to a single site may reduce variability in SUVs, it also slows patient accrual and fails to demonstrate generalizability of the method or drug being tested.

Multicenter trials allow faster patient accrual but require more care in controlling biases owing to the differences in

*Address all correspondence to: Paul Kinahan, E-mail: kinahan@uw.edu

instrument characteristics and clinical protocols across sites. The need to control SUV bias has been partially addressed through consensus documents, which provide operating guidelines,^{9–11} and accreditation organizations, which validate instrument performance.^{12–14} Additionally, recent efforts have looked at the dependence of bias on reconstruction settings to find “harmonized” settings that match biases for small lesions.^{14,15} A key area that is still lacking, though, is a characterization of bias stability that is due not to user-selected settings and clinical workflows, but to hardware performance and calibration. While accredited sites undergo periodic quality assurance tests that confirm PET signal accuracy is within a specified range, these tests most often involve measurements of short-lived sources that cannot independently evaluate bias stability of dose calibrators and PET/CT scanners, nor can they separate true changes in measurement bias from variability that arises from the manual preparation of the test object. In contrast, long-lived sources provide very stable signal for calibration tests, up to a decay factor, but have not been widely used before to study longitudinal bias stability across networks of hospitals.

In the present work, we examine the variability of scanner and dose calibrator biases through measurements of long-lived sources made at nine hospitals. Two-hundred and thirty-six repeated PET measurements were made an average of 20 days apart using solid epoxy sources uniformly infused with ⁶⁸gallium/⁶⁸germanium (⁶⁸Ge). We report the results of a central analysis of scanner and dose calibrator data performed at the University of Washington.

2 Methods

2.1 Calibration Measurements

Calibration stability was assessed with long-lived cross-calibration kits (X-Cal kits) described in detail elsewhere^{16,17} and briefly summarized here. The kits contained two ⁶⁸Ge sealed sources, one to be assayed in a dose calibrator and the other to be measured by a PET/CT scanner. The dose calibrator standard contained ~0.8 MBq (20 μ Ci) of ⁶⁸Ge activity initially and approximated the geometry of a syringe containing a clinical dose of radiotracer [Fig. 1(a)]. All dose calibrator standards were calibrated to a well-characterized National Institute of Standard and Technology (NIST) ⁶⁸Ge standard.¹⁸ The scanner source, or X-Cal phantom [Fig. 1(b)], was cylindrical with an active region 45 mm in diameter by 45 mm in height. The initial activity concentration was ~250 kBq/ml (6.8 μ Ci/ml) for a total activity of 20 MBq (0.54 mCi). The X-Cal phantom was made from the same batch of epoxy as the dose calibrator source for optimal consistency from the dose calibrator measurement to the scanner measurement.

The X-Cal phantom was imaged with a site’s standard whole body oncology protocol. Imaging parameters such as scan duration and x-ray tube voltage therefore varied across sites, but it was verified from image metadata that all sites used the proper nuclide setting (⁶⁸Ge) for the scans. During the PET scan, the X-Cal phantom was affixed to the bottom of an American College of Radiology (ACR) flood phantom or equivalent (Data Spectrum Corporation, Durham, North Carolina) using a mounting bracket provided with the kits [Fig. 1(b)]. The flood phantom was filled with nonradioactive water. The dose calibrator standard was assayed as if it were a patient dose. A conversion factor supplied by NIST allowed either ⁶⁸Ge or ¹⁸F calibration settings to be used^{18,19} for the measurement, and sites used ¹⁸F

Fig. 1 X-Cal kits included two long-lived sources for use in (a) a site’s dose calibrator and (b) PET/CT scanner. The scanner source is shown mounted in a 20-cm American College of Radiology flood phantom, which is filled with water during scanning.

settings to best represent clinical bias. A background radiation reading for the dose calibrator was recorded and subtracted from source assay. The dose calibrator measurement and the PET/CT scan were typically completed on the same day.

Dose calibrator assay data and PET images were submitted to the University of Washington Imaging Research Laboratory for analysis. Images were stored in the digital imaging and communication in medicine (DICOM[®]) format.²⁰ Fused PET and CT images of the phantom assembly are shown in Fig. 2.

2.2 Network

All sites participating in this study (Table 1) were members of the National Cancer Institute’s Quantitative Imaging Network (QIN).²¹ Data were collected for 19 PET/CT scanners and 16 dose calibrators at the nine sites (some dose calibrators were used for more than one PET/CT scanner). Table 2 shows makes and models of scanners used in this study. One scanner used magnetic resonance imaging as its complementary modality and therefore did not have attenuation data from CT images. For this scanner, a CT image of the phantom from a different scanner was used to generate an attenuation map, which was then used in the reconstruction. Because no magnetic resonance data were used, we will refer to this method as PET/CT, just as for the other scanners of Table 2. For reconstruction, 10 scanners used three-dimensional (3-D)-ordered subset expectation maximization (OSEM),²² two of which used time-of-flight data. Three scanners used two-dimensional (2-D) OSEM, of which one used Fourier rebinning from 3-D data.²³ Five scanners used time-of-flight BLOB-OS-TF reconstruction.²⁴ One used the analytic method of 3-D reprojection.²⁵ The number of scans per scanner ranged from 3 to 43 (average of 13) and the duration over which scans were performed ranged from 39 to 412 days (average of 232). The average duration between scans for the entire set of data ($N = 236$) was 20 days. Table 3 shows makes and models of dose calibrators, which were tested in parallel with the scanners. Not all PET/CT scans had dose calibrator measurements accompanying them. A total of 161 dose calibrator measurements were made with an average separation of 24 days between measurements.

Fig. 2 Fused PET/CT images of an X-Cal phantom in a water-filled 20-cm diameter cylindrical phantom. PET signal is shown in orange in (a) transaxial and (b) coronal views.

Table 1 Participating QIN sites.

H. Lee Moffitt Cancer Center and Research Institute
Johns Hopkins University
Massachusetts General Hospital
Memorial Sloan Kettering Cancer Center
The Ohio State University
University of Iowa
University of Pennsylvania
University of Pittsburgh
University of Washington

Table 3 Dose calibrators at participating QIN sites. Some dose calibrators were used for multiple PET/CT scanners.

make and model	<i>n</i>
Atom Lab 100	1
Capintec 127R	1
Capintec CRC-15R	1
Capintec CRC-15W	1
Capintec CRC-25R	5
Capintec CRC-25W	2
Capintec CRC-712	1

Table 2 PET/CT scanners at participating QIN sites.

Make	Model	<i>n</i>
General electric	Discovery STE	5
	Discovery 690	1
	Discovery RX	1
	Discovery LS	1
Philips	Gemini TF	2
	Ingenuity TF	2
	Gemini TF big bore	1
Siemens	Biograph mCT	2
	Biograph Duo	1
	Biograph 40	1
	Biograph mMR	1
	ECAT	1

2.3 Image Analysis

The signal from each reconstructed image was computed as the mean of voxel values within a region of interest (ROI) in the center of the phantom. ROIs were drawn via an automated algorithm, XCaliper, which was implemented in a plug-in developed at the University of Washington for the OsiriX DICOM[®] viewer.²⁶ XCaliper computes the centroid of the phantom signal and draws a whole-voxel ROI in the phantom's uniform central region based on user-specified maximum ROI dimensions and the voxel dimensions listed in the DICOM[®] header. For this study, the maximum ROI size was 15 mm in each dimension. This size was chosen to exclude voxels near the edge of the phantom, where signal intensity is lessened due to partial-voxel effects and finite image resolution. A trans-axial section of an XCaliper ROI, which spans several trans-axial slices, is shown for a representative reconstruction in Fig. 3. The plug-in reports the mean signal λ within the ROI.

2.4 Metrics

The term “recovery coefficient” was originally defined by Hoffman et al.²⁷ as “...the apparent isotope concentration in the image divided by the true isotope concentration. Its value is 1.0 for large objects in which image concentration is equal

Fig. 3 PET image with XCaliper ROI marked with a gray square. The black bounding box is 15 mm in each dimension (including axially) and is centered on the phantom signal's centroid, which is marked by the plus sign. The ROI includes all voxels that are completely contained in the bounding box.

to the true object concentration.” Here we use the central region of the X-Cal phantom to avoid the effects of object size, and instead focus on global bias in SUVs. As shown in Table 4, the scanner recovery ρ is computed with the XCaliper mean signal and known phantom activity concentration. The dose calibrator recovery r is computed using the ^{18}F equivalent activity, which is the true ^{68}Ge activity times a scale factor that accounts for calibrator nuclide dependence.^{18,19} The recovery for the SUV, which we call σ , is computed as the ratio of scanner and dose calibrator recoveries,²⁸ as suggested by Eq. (1). For the three recovery metrics (ρ , r , and σ) of each scanner-calibrator pair, we characterize measurement repeatability by reporting the relative percent range (maximum value minus minimum,

Table 4 Metrics.

Metric	Symbol
XCaliper signal mean	λ
Known phantom activity concentration	L
Scanner recovery	$\rho = \lambda/L$
Dose calibrator standard assay	α
Known dose calibrator standard activity (F-18 equivalent)	G
Dose calibrator recovery	$r = \alpha/G$
SUV recovery	$\sigma = \rho/r$
Relative percent range (magnitude of range divided by mean)	Φ
Magnitude of maximum change in successive measurements (per scanner or calibrator)	$ \Delta_{\text{MAX}} $
Number of measurements (per scanner or calibrator)	N

Table 5 Extracted image metadata and associated DICOM[®] tag.

DICOM [®] tag	Field	Comments
0054,1322	Dose calibration factor ^a	GE and Siemens only
7053,1009	Activity scale factor ^a	Philips only
0054,1323	Scatter fraction	
0054,1324	Dead time factor	
0009,1071	Total prompts	
0009,10bb	Filter parameter	
0017,1004	Date of last calibration	

^aNumbers from the fields of rows 1 and 2 are referred to as “image scale factors” below.

divided by overall mean), the mean and standard deviation, and the largest change seen in successive measurements $|\Delta_{\text{max}}|$.

2.5 Image Metadata

For each PET/CT scan, the metadata from the DICOM[®] images were extracted. The metadata are saved in each image slice produced by the scanners and contain many imaging parameters. The parameters that were deemed relevant for the present study are shown in Table 5. Of particular note, we computed Pearson correlation coefficients for the factors of rows 1 and 2 with ρ and r , using the appropriate field for a given scanner. Because of their similar function,^{11,29–31} we will group together those numbers from rows 1 and 2 of Table 5, and will refer to them collectively as “image scale factors.”

3 Results

3.1 Scanners

Figure 4 shows that scanner recovery ρ was variable in time and between sites. Table 6 shows more characteristics of ρ on a per-scanner basis. The average ρ across scanners was 0.924. The standard deviation of ρ ranged from 0.011 to 0.065. The coefficient of variation (COV) for ρ was 3.5%. Six scanners had $\Phi > 0.1$. The average $|\Delta_{\text{max}}|$ was 0.057.

Table 7 shows the statistics of scanner recovery from subsets of scanners in which the make and model are similar. Note that because different instances of scanners are grouped together in Table 7, differences in average scanner bias can contribute to dispersion metrics, unlike in Table 6.

3.2 Dose Calibrators

Figure 5 shows that dose calibrator recovery, like scanner recovery, was prone to vary unpredictably in time. Table 8 shows the statistics across calibrators. Where single dose calibrators have been used for multiple scanners, each instance of a scanner-calibrator pair has been included in the table. The average r was 1.01 across all calibrators. Standard deviations of r ranged from 0.004 to 0.048. Three scanners were affected by dose calibrators with $\Phi > 0.1$. The average Φ for dose calibrators was 0.065. The average $|\Delta_{\text{max}}|$ was 0.051.

Fig. 4 Scanner recovery ρ versus time. Lines with matched color represent different scanners at a given site. Scans were performed between June 2013 and March 2015.

3.3 Standardized Uptake Values

Figure 6 shows that the calculated SUV recovery σ (Table 4) versus time showed no discernible trend. Table 9 shows the values of σ across scanners. Average σ across scanner-calibrator

pairs was 0.910. Per-pair standard deviation ranged from 0.0044 to 0.067. The standard deviation and $|\Delta_{\max}|$ of σ were slightly larger on average than those of ρ and r .

Figure 7 shows the recovery for all scanners, ρ , plotted against the recovery from dose calibrators, r . Three pairs of

Table 6 Statistics of scanner recovery ρ . Rows do not correspond to Table 1. The bottom row shows distribution of site means and not the dataset as a whole.

N	Days	Mean	Std deviation	[min–max]	Φ	$ \Delta_{\max} $
12	323	0.861	0.023	[0.809–0.881]	0.084	0.041
11	290	0.867	0.019	[0.828–0.891]	0.073	0.039
7	389	0.865	0.009	[0.856–0.879]	0.027	0.023
3	39	0.893	0.003	[0.890–0.894]	0.005	0.005
3	66	0.973	0.003	[0.970–0.977]	0.007	0.006
10	124	0.865	0.005	[0.859–0.872]	0.016	0.014
20	258	0.975	0.041	[0.884–1.063]	0.184	0.089
15	397	0.888	0.035	[0.796–0.944]	0.167	0.085
14	437	0.928	0.029	[0.856–0.977]	0.130	0.121
12	389	0.965	0.024	[0.927–0.995]	0.070	0.052
6	391	0.983	0.037	[0.945–1.031]	0.088	0.062
43	412	0.858	0.065	[0.752–0.944]	0.224	0.100
14	185	0.943	0.018	[0.885–0.965]	0.085	0.067
11	122	0.991	0.011	[0.975–1.011]	0.036	0.034
13	185	0.930	0.016	[0.889–0.953]	0.069	0.043
13	179	0.948	0.045	[0.887–1.084]	0.208	0.145
9	265	0.951	0.033	[0.896–1.005]	0.114	0.060
8	213	0.908	0.018	[0.889–0.934]	0.049	0.032
12	401	0.963	0.030	[0.923–1.004]	0.084	0.058
12.4 ± 8.3	267 ± 127	0.924 ± 0.047	0.025 ± 0.016		0.091 ± 0.066	0.057 ± 0.038

Table 7 Scanner recovery ρ for subsets of scanners.

Make and model	N	Mean	Std dev	Cov
GE DSTE	57	0.907	0.046	0.051
Siemens biograph	46	0.938	0.051	0.055
Philips TF	34	0.927	0.053	0.057
All scans	236	0.916	0.060	0.065

scanners and dose calibrators had recovery values with correlation coefficients $r^2 > 0.75$. For most pairs, correlation coefficients were below 0.50.

3.4 Image Metadata

For many of the scanners, the scanner recovery ρ was correlated with scale factors from the DICOM® headers. Table 10 shows the correlation coefficients and COVs for ρ and the image scale factor. Correlations between dose calibrator recovery r and the image scale factor were less significant and are not shown.

Fig. 5 Dose calibrator recovery r versus time. Colors are matched to Fig. 4.**Table 8** Statistics of dose calibrator recovery r . Rows do not correspond to Table 1. The bottom row shows the distribution of site means and not the dataset as a whole.

N	Days	Mean	Std deviation	[min-max]	Φ	$ \Delta_{\max} $
12	323	0.985	0.048	[0.851–1.018]	0.169	0.106
11	323	0.987	0.048	[0.851–1.018]	0.169	0.106
6	389	1.011	0.020	[0.998–1.051]	0.052	0.043
2	357	1.004	0.005	[1.001–1.008]	0.007	0.007
2	228	1.029	0.030	[1.008–1.051]	0.042	0.043
15	397	0.997	0.046	[0.934–1.085]	0.151	0.093
4	524	0.985	0.011	[0.970–0.996]	0.026	0.025
6	542	1.003	0.007	[0.992–1.013]	0.020	0.011
38	412	0.992	0.013	[0.963–1.033]	0.070	0.048
14	294	1.058	0.010	[1.045–1.079]	0.033	0.018
11	294	1.063	0.022	[1.050–1.127]	0.072	0.076
13	294	1.063	0.020	[1.050–1.127]	0.072	0.076
10	286	1.070	0.004	[1.065–1.078]	0.012	0.010
3	262	0.979	0.027	[0.949–1.001]	0.053	0.052
8	213	0.995	0.030	[0.922–1.011]	0.089	0.089
6	213	1.005	0.004	[0.997–1.009]	0.011	0.010
10.1 ± 8.6	334 ± 99	1.01 ± 0.03	0.022 ± 0.015		0.065 ± 0.055	0.051 ± 0.036

Fig. 6 SUV recovery σ versus time. Colors are matched to Figs. 4 and 5.

Table 9 Statistics of SUV recovery σ . Rows do not correspond to Table 1. The bottom row shows the distributions of site average and not of the dataset taken as a whole.

N	Days	Mean	Std deviation	[min–max]	Φ	$ \Delta_{\max} $
12	323	0.876	0.033	[0.837–0.970]	0.152	0.114
11	323	0.880	0.036	[0.837–0.978]	0.160	0.085
6	389	0.854	0.016	[0.823–0.869]	0.054	0.046
2	357	0.890	0.004	[0.887–0.893]	0.007	0.006
2	228	0.946	0.033	[0.922–0.969]	0.049	0.047
15	397	0.892	0.057	[0.785–1.010]	0.252	0.158
4	524	0.940	0.029	[0.908–0.976]	0.073	0.069
6	542	0.980	0.040	[0.936–1.031]	0.097	0.054
38	412	0.875	0.067	[0.763–0.947]	0.210	0.095
14	294	0.889	0.020	[0.826–0.906]	0.090	0.064
11	294	0.930	0.024	[0.866–0.956]	0.096	0.064
13	294	0.874	0.023	[0.822–0.900]	0.090	0.078
10	286	0.888	0.048	[0.822–1.010]	0.211	0.134
3	262	0.948	0.006	[0.943–0.955]	0.013	0.012
8	213	0.914	0.044	[0.882–1.015]	0.146	0.111
6	213	0.981	0.024	[0.939–1.007]	0.069	0.068
10.1 ± 8.6	334 ± 99	0.910 ± 0.039	0.032 ± 0.017		0.111 ± 0.072	0.075 ± 0.041

For General Electric, Philips and Siemens scanners, Fig. 8 shows phantom signal recovery and image scale factors. Each pane shows two scanners for a given manufacturer, representing the largest and smallest magnitude of the correlation in the intramanufacturer dataset for scanners having at least five measurements. It can be seen that correlations (and lack of correlations) were found in all three types of scanners.

Figure 9 shows the time series of image scale factors and scanner recovery for scanners having a readable “date of last

calibration” field in their DICOMs[®]. Data have been rescaled and translated to represent percent changes over time relative to their means. It can be seen that dashed lines (i.e., scanner recalibrations) often fall between changes in (a) the image scale factor. This pattern was not as evident in the scanner recovery data (b), which exhibit more randomness.

No notable changes or trends were seen in the other DICOM[®] header fields of Table 5, which, respectively, characterized scatter fraction, detector deadtime, detected events, and image smoothing.

Fig. 7 Dose calibrator recovery versus scanner recovery.

Table 10 Statistics of the scanner recovery and the image scale factors at each site for which they were computable. The second column shows Pearson correlation coefficients r^2 for scanner recovery ρ and scale factors from rows 1 and 2 of Table 3. Third and fourth columns show the COV across repeat scans of scanner recovery ρ and the image scale factor from the DICOMs®.

Measurements	Correlation coefficient r^2	COV of ρ	COV of image scale factor
12	0.934	0.031	0.022
12	0.915	0.027	0.026
10	0.872	0.023	0.022
7	0.752	0.010	0.005
10	0.340	0.005	0.004
20	0.985	0.036	0.035
15	0.725	0.040	0.040
8	0.749	0.009	0.012
15	0.143	0.024	0.011
6	0.953	0.038	0.038
8	0.353	0.013	0.010
3	0.150	0.003	0.001
8	0.892	0.020	0.021
11	0.002	0.011	1.47×10^{-7}
10.4 ± 4.2	0.626 ± 0.337	0.021 ± 0.012	0.017 ± 0.013

4 Discussion

4.1 Scanner Bias

Scanner recovery ρ was consistently below one, with average recovery around 90% of known signal for many scanners as shown in Fig. 4. Similar bias has been previously observed in well-calibrated scanners when these solid phantoms were used.¹⁷ The likely causes are incorrect modeling in the reconstructions for the scatter and attenuation of photons by the

epoxy.³² The differing physics between the phantom and the patients means that these biases are likely not equal to the biases present in patient scans and we are thus prevented from using this data to predict the bias of clinical SUVs. Still, the contribution of these physical effects is expected to be stable from scan to scan, and therefore scanner stability may be evaluated.

Table 6 shows that scanner recovery values were not stable over the course of this study. It was common for scanners to have good stability over some number of successive measurements followed by a larger shift in apparent bias, as shown in Fig. 4. Changes in scanner bias could be due to a number of factors. Previously, changes to the applied x-ray tube voltage in the CT measurements used for attenuation correction have been shown to affect bias,³² but that effect was smaller in magnitude than the changes seen here, and changes in the CT voltage are not expected in these data. Corrections for random and scattered counts could also introduce biases that would change as the source decayed; however, these effects would likely vary smoothly with activity, and are unlikely to produce the jagged nonmonotonic behavior shown in Fig. 4.

A more likely explanation is that scanner bias for these measurements is actually variable even when no user error occurs. The causes of this variability are likely numerous, but Fig. 8 suggests that for some scanners, it is related to the image scale factor.

The reason we have examined the image scale factor is that for many scanners, the sensitivity, which is the ratio of detected events in the reconstructed image to the true number of decays in the object, is determined empirically at regular time intervals.³³ Although we cannot confirm the precise implementation of this process for all scanners in this study, our data suggest that recalibrations affect the image scale factors (Fig. 9), which in turn affect the bias of image values (Table 10 and Fig. 8). We note that recalibration frequently involves short-lived nuclides and manual phantom filling, as well as measurements that are prone to human error. One possible interpretation of our results is that true scanner sensitivity is relatively stable in time, while the measurement of sensitivity is not. In particular, Table 10 shows that the image scale factor is nearly as variable as scanner recovery. Recalibration may, therefore, contribute to bias instead of compensating for it. However, not all scanners showed good correlation between ρ and the image scale factor, and the disparity in the two panes of Fig. 9 suggests that other sources of measurement noise also affect scanner bias. More investigation is needed, but DICOM® header data appear to have value for quality control in that instability of the scale factor may indicate variable scanner bias.

Table 7 shows that for subsets of scanners whose make and model are similar, the COV of recovery is reduced slightly versus the entire scanner population. This is expected, since it was observed that different scanners had different biases on average. However, the intramodel variability shown in Table 7 is still larger than that seen at most single sites, suggesting that even with similar scanners, signal variability will tend to increase when multiple sites are involved.

4.2 Dose Calibrator Bias

Dose calibrator biases were generally smaller than scanner bias though they still varied in time and between sites. One possible explanation for this variability is that dose calibrators have a sensitivity adjustment factor that is set by the user and that must be changed depending on the nuclide being assayed.

Fig. 8 For each of three scanner manufacturers (one per pane), representative plots of scanner signal recovery and image scale factors, shown with solid and dashed lines, respectively, for the most- and least-correlated cases. To emphasize correlations, each variable was scaled and translated to equal fractional fluctuation about its mean. In each plot, a given line color represents one scanner.

Fig. 9 Percent deviation from mean of (a) the image scale factor and (b) scanner recovery for all scanners for which “date of last calibration” was readable in the DICOM® headers. Where a calibration was designated as having taken place between scans, lines are dashed. Line colors and markers are matched between panes.

Differences in these settings between sites, as well as human error in changing settings, could be responsible for the variability seen in Fig. 5 and Table 8. For one of the sites in this study, it was determined (by the site) that an incorrect calibration setting had been used, and data were retrospectively corrected. We note that although the use of ^{18}F equivalent activity meant sites could use the ^{18}F calibration setting for their dose calibrators (i.e., their standard clinical settings), it does not ensure that a correct calibration setting will be used. Further, to compute ^{18}F equivalent activity, we used a single conversion factor of 1.024 for all dose calibrators, and this may contribute to the differences in bias between sites having different models of calibrators.¹⁹

For reasons discussed below, the dose calibrator standards assayed in this study were on the order of 370 kBq (10 μCi), which is ~ 1000 times smaller than the quantity that would be injected into patients. The low activity of the standards meant factors such as background radiation or measurement noise would have greater influence on the recovery values calculated for these measurements than they would on patient dose assays. However, we have attempted to compensate for changes in background by recording and subtracting background readings, and we believe the calibrators are linear (that is, accurate) in their readings. Checks of instrument linearity are in fact part of quality control procedures for dose calibrators, though the range of activity levels used may not extend to the low values tested here.³⁴ Still, the lack of trends in the data does

suggest a linearity of the devices over the tested range, as an activity-dependent sensitivity would lead to a time dependence of the biases in this experiment, which was not observed. While the bias and variability in our dose calibrator measurements are not readily attributable to known causes, they are similar to those seen in previous studies of dose calibrators using ^{68}Ge sources at this activity level^{28,35} and at activity levels closer to clinical values.³⁶

4.3 Standardized Uptake Value Bias

The formula used to compute SUVs [Eq. (1)] suggests that the ratio of recoveries from the scanner and the dose calibrator should be equal to the recovery of SUVs. Variability of SUV recovery σ , computed as in Table 4, was higher than that of scanner or dose calibrator measurements as gauged by the standard deviation and $|\Delta_{\text{max}}|$. If scanner and dose calibrator biases were highly correlated, SUVs could in principle be stable even with varying biases in both instruments. Significant correlations were not seen in our dataset (Fig. 7), and σ consequently was not less variable than the scanner or dose calibrator recovery values.

We note that instrument calibration is just one source of SUV variability. The test–retest repeatability of clinical SUVs has been shown to be 10% to 12% in single sites studies with careful quality control. In multicenter studies it may be several times larger,³⁷ suggesting that the contributions from other sources

of variability are significant. Calibration bias is, therefore, one factor of many that should be monitored in clinical trials.

4.4 Previous Studies

Our metric σ can be readily compared with results from previous studies that measured SUV bias with ^{18}F . This is because it is common practice to test scanner and dose calibrator accuracy together in the following way: a dose is assayed in the onsite dose calibrator, injected into a phantom of known volume, and imaged on the PET scanner. The ratio of the computed activity concentration to reconstructed signal is then used to characterize SUV accuracy. Using this method, Scheuermann found that 12% of over 100 scans submitted as part of ACR accreditation had errors $>10\%$.¹² With a similar approach, Fahey reported a 33% COV for signal recovery across a network of hospitals. Geworski found SUV errors near 10% in 5 of 19 scanners.³⁶ However, using ^{18}F limits the information obtained on scanner and dose calibrator stability because the observed biases could be due to a number of causes including incorrect decay correction, incomplete dose injection, miscalculation of activity concentration, or real changes in the bias of the scanner or dose calibrator, none of which can be detected individually. Additionally, it has been shown that measurements of PET signal variability across sites can depend on whether long- or short-lived sources are used.³⁸

The primary advantage of using long-lived sources is that with proper decay corrections the known signal is extremely stable and allows data taken at different time points to be compared. A previous study with long-lived sources showed that the variability of scanner calibration was $\sim 5\%$ for a single scanner when quality control procedures were in place to catch errors introduced by periodic scanner recalibrations.³³ In the course of that study, several recalibration errors were observed. Using the X-Cal kits of the present study, Doot et al.²⁸ showed that recalibration introduced changes in scanner bias of $\sim 10\%$ at a majority of sites in a small ($N = 6$) network of hospitals.

The present study is consistent with previous reports using aqueous and long-lived phantoms. Figure 4 shows that at any given time point, a scanner is liable to be within some finite range on the axis of signal recovery that spans $\sim 10\%$ of the signal value. Therefore, a sampling of scanner biases with short-lived nuclides would be expected to show a similar or larger range of biases, as has been observed.

4.5 Limitations

Some design features of the sources imposed limitations on our study. We chose to use the same batch of epoxy in the manufacturing of the phantoms and dose calibrator standards, which led to the latter being much lower in activity than a clinical patient dose. While this design choice has intuitive appeal in that comparison of biases between the scanner and dose calibrator has one fewer confounding factor (i.e., it assures that manufacturing variability is minimized between sources), it does constrain the activity of the dose calibrator standard. Because PET scanners cannot handle arbitrarily high coincidence detection rates, the activity concentration of the epoxy was limited. The amount of epoxy that could be used in the dose calibrator standard was also limited by the need to match the source geometry to that of a patient dose in a syringe. The resulting low activity means that our dose calibrator data may not reflect the signal properties of clinical radiotracer doses. Using separate batches

of epoxy with higher activity concentration for the dose calibrator standards in future studies would provide measurements closer to the clinical range.

The use of solid epoxy also makes comparison of scanners difficult, as the relationship between bias values in the phantoms and bias in patients is not only unknown but also potentially different for each scanner due different implementations of data corrections for physical effects. The use of an epoxy whose physical properties are similar to human tissues would allow not only comparisons between scanners but also evaluation of clinical SUV bias. We are considering this for future versions of the phantom.

5 Conclusions

Our phantom images suffered some signal bias due to the epoxy used in their construction, and consequently we have focused on signal stability rather than absolute accuracy. SUV bias from instrument calibration was often stable over successive measurements, and on average had a modest COV of 3.5%. However, over the course of our 14-month study, shifts in bias were apparent for many scanners, and on average SUV recovery varied over an intra-scanner range of 11%. The biases of scanners and dose calibrators were not correlated, and estimated SUV variability was not smaller than the variability of either instrument. Information saved in the DICOM[®] headers appears to show that scanner recalibration influences PET scanner bias and potentially contributes to changes in SUV bias. The variability of scale factors saved in the DICOMs[®] was nearly as large as that of scanner signal. This supports the conclusion of Lockhart et al.³³ that a long-lived source should be used as an independent check on the calibration process to reduce potential recalibration errors.

Disclosures

The $^{68}\text{Ge}/^{68}\text{Ga}$ PET dose calibrator and scanner crosscalibration kits used in this article are sold by RadQual, LLC., and protected by United States Patent 7,858,925. Author P.K. is one of the inventors of the aforementioned patent and may receive compensation from the patent licensing agreement between the University of Washington and RadQual, LCC. No other relevant financial interests were disclosed.

Acknowledgments

This work was supported by NIH Grant U01-CA148131. The authors wish to thank technologists Trang Marquez, and Patrick Clark, and Dr. Lawrence MacDonald, for their collection of phantom data, and Dr. Larry Clarke for his work in organizing the Quantitative Imaging Network.

References

1. R. L. Wahl et al., "From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors," *J. Nucl. Med.* **50**, 122S–150S (2009).
2. R. Frank and R. Hargreaves, "Clinical biomarkers in drug discovery and development," *Nat. Rev. Drug Discovery* **2**, 566–580 (2003).
3. D. A. Mankoff, D. A. Pryma, and A. S. Clark, "Molecular imaging biomarkers for oncology clinical trials," *J. Nucl. Med.* **55**, 525–528 (2014).
4. W. A. Weber, "Assessing tumor response to therapy," *J. Nucl. Med.* **50**(Suppl 1), 10S (2009).
5. R. K. Doot et al., "Design considerations for using PET as a response measure in single site and multicenter clinical trials," *Acad. Radiol.* **19**, 184–190 (2012).
6. R. Boellaard, "Standards for PET image acquisition and quantitative data analysis," *J. Nucl. Med.* **50**, 11S–20S (2009).

7. F. H. Fahey et al., "Variability in PET quantitation within a multicenter consortium," *Med. Phys.* **37**, 3660–3666 (2010).
8. L. M. Velasquez et al., "Repeatability of ^{18}F -FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies," *J. Nucl. Med.* **50**, 1646–1654 (2009).
9. M. M. Graham et al., "Summary of the UPECT protocol for ^{18}F -FDG PET/CT imaging in oncology clinical trials," *J. Nucl. Med.* **56**, 955–961 (2015).
10. R. Boellaard et al., "FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0," *Eur. J. Nucl. Med. Mol. Imaging* **42**, 328–354 (2015).
11. FDG-PET/CT Technical Committee, "FDG-PET/CT as an imaging biomarker measuring response to cancer therapy profile version 1.05," 2013, RSNA.ORG/QIBA (30 November 2017).
12. J. S. Scheuermann et al., "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).
13. J. S. Scheuermann et al., "Qualification of national cancer institute-designated cancer centers for quantitative PET/CT imaging in clinical trials," *J. Nucl. Med.* **58**, 1065–1071 (2017).
14. J. Sunderland et al., "Assessment of PET/CT reconstruction harmonization through Gaussian post-filtration," *J. Nucl. Med.* **57**, 1917–1917 (2016).
15. D. Byrd et al., "A post-hoc methodology for harmonizing PET image reconstruction protocols," *J. Nucl. Med.* **55**, 372–372 (2014).
16. B. E. Zimmerman et al., "Development of a calibration methodology for large-volume, solid (68)Ge/(68)Ga phantoms for traceable measurements in positron emission tomography," *Appl. Radiat. Isot.* **87**, 5–9 (2014).
17. D. W. Byrd et al., "Evaluation of cross-calibrated (68)Ge/(68)Ga phantoms for assessing PET/CT measurement bias in oncology imaging for single- and multicenter trials," *Tomography* **2**, 353–360 (2016).
18. 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**, 448–453 (2010).
19. B. Zimmerman, D. Bergeron, and J. Cessna, "Impact of recent change in the NIST standard for ^{18}F on the relative response of ^{68}Ge -based mock syringe dose calibrator standards," *J. Nucl. Med.* **115**, 159384 (2015).
20. <http://www.dicomstandard.org/current/> (1 December 2017).
21. L. P. Clarke et al., "The quantitative imaging network: NCI's historical perspective and planned goals," *Transl. Oncol.* **7**, 1–4 (2014).
22. H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Trans. Med. Imaging* **13**(4), 601–609 (1994).
23. M. Defrise et al., "Exact and approximate rebinning algorithms for 3-D PET data," *IEEE Trans. Med. Imaging* **16**, 145–158 (1997).
24. M. E. Daube-Witherspoon et al., "Application of the row action maximum likelihood algorithm with spherical basis functions to clinical PET imaging," *IEEE Trans. Nucl. Sci.* **48**, 24–30 (2001).
25. P. E. Kinahan and J. G. Rogers, "Analytic 3D image reconstruction using all detected events," *IEEE Trans. Nucl. Sci.* **36**, 964–968 (1989).
26. A. Rosset, L. Spadola, and O. Ratib, "OsiriX: an open-source software for navigating in multidimensional DICOM images," *J. Digital Imaging* **17**, 205–216 (2004).
27. E. J. Hoffman, S.-C. Huang, and M. E. Phelps, "Quantitation in positron emission computed tomography: 1. Effect of object size," *J. Comput. Assisted Tomogr.* **3**, 299–308 (1979).
28. R. K. Doot et al., "Biases in multicenter longitudinal PET standardized uptake value measurements," *Transl. Oncol.* **7**, 48–54 (2014).
29. "DICOM[®] conformance statements—molecular imaging," <https://www.healthcare.siemens.com/services/it-standards/dicom-conformance-statements-molecular-imaging> (3 July 2017).
30. "Nuclear medicine connectivity DICOM[®] conformance statements," <http://www.usa.philips.com/healthcare/resources/support-documentation/dicom-nuclear-medicine> (3 July 2017).
31. "ACR-NEMA digital imaging and communications in medicine (DICOM[®]) supplement 12 positron emission tomography image objects," ftp://medical.nema.org/medical/dicom/final/sup12_ft.pdf (4 July 2017).
32. D. Byrd et al., "Attenuation effects in solid phantoms used for PET/CT scanner calibration," *J. Nucl. Med.* **55**(Suppl. 1), 2095 (2014).
33. 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).
34. P. Zanzonico, "Routine quality control of clinical nuclear medicine instrumentation: a brief review," *J. Nucl. Med.* **49**, 1114–1131 (2008).
35. R. K. Doot et al., "Instrumentation factors affecting variance and bias of quantifying tracer uptake with PET/CT," *Med. Phys.* **37**, 6035–6046 (2010).
36. L. Geworski et al., "Multicenter comparison of calibration and cross calibration of PET scanners," *J. Nucl. Med.* **43**, 635–639 (2002).
37. V. Kumar et al., "Variance of standardized uptake values for FDG-PET/CT greater in clinical practice than under ideal study settings," *Clin. Nucl. Med.* **38**, 175–182 (2013).
38. S. Chauvie et al., "The ^{68}Ge phantom-based FDG-PET site qualification program for clinical trials adopted by FIL (Italian Foundation on lymphoma)," *Phys. Med.* **32**, 651–656 (2016).

Darrin Byrd received his BA degree in physics from the University of California, Berkeley, and his MS degree in physics from the University of Wisconsin, Milwaukee. He has worked in the Imaging Research Laboratory at the University of Washington since 2011. Currently, he is involved in numerous projects related to improving quality control for PET/CT imaging.

Rebecca Christopf received her bachelor's degree in physics and participated in an internship at Stanford Linear Accelerator Center as an undergraduate. Her professional expertise involves management and facilitation of multimodality clinical imaging studies as well as longitudinal observational studies.

Grae Arabasz is the lead PET/MR technologist at the Martinos Center for Biomedical Imaging. He has experience in combined PET/MRI since 2009 and is currently working with dozens of researchers and physician-scientists on developing neurological and whole-body oncological applications.

Ciprian Catana is the director of integrated MR-PET imaging at the Martinos Center, MGH and an associate professor in radiology at Harvard Medical School. He has extensive experience in combining PET and MRI and is currently focusing on developing quantitative MR-PET for human use. Working closely with dozens of basic science researchers and physician-scientists, he is applying this novel technology to the study of normal brain and neuropsychiatric diseases as well as exploring its clinical potential for whole-body oncological applications.

Joel Karp is a professor of radiologic physics in the Department of Radiology and in the Department of Physics and Astronomy at the University of Pennsylvania. He is the chief of the Physics and Instrumentation Research Group in radiology and directs nuclear medicine/PET physics and QC in the clinic, as well as the Small Animal Imaging Facility Nuclear Medicine (PET/SPECT/CT) core. His research has focused on investigations to improve and characterize the performance of PET technology and has resulted in the development of fully 3-D PET scanners, some of which have been implemented commercially for human and animal imaging.

Martin A. Lodge is a medical physicist with interest in diagnostic imaging, particularly PET/CT. He has a PhD in nuclear medicine physics from the University of London and has worked in PET since 1996. Recent research interests have focused mainly on PET instrumentation, data processing, and image analysis. These have included data acquisition, quantitative image analysis, development of image-based biomarkers, statistical analysis of image data, optimizing clinical protocols, image artifacts, and quality assurance of PET/CT instrumentation.

Charles Laymon is an imaging physicist in the PET research facility at the University of Pittsburgh.

Eduardo G. Moros is the chief of medical physics at the Moffitt Cancer Center. He is an active member of the AAPM and ASTRO, an associate editor of the *Journal of Applied Clinical Medical Physics* and the *Journal of Radiation Oncology*. He has published over 165 peer-reviewed papers and has been a principal investigator/coinvestigator on numerous NIH and industry research grants. His current interest is radiomics applications in radiation oncology.

Mikalai Budzevich received his MS degree in nuclear physics from Belarusian State University, Belarus, in 2001. After moving to the US, he completed his PhD in applied physics at the University of South Florida in 2011. He has worked in the field of diagnostics imaging, nuclear medicine, and image processing as a postdoctoral fellow at Radiation Oncology Department at H. Lee Moffitt Cancer Center. In 2015, he joined the Small Animal Imaging Laboratory at Moffitt as a staff scientist.

Sadek Nehmeh is an associate professor of physics in radiology and chief of PET and nuclear physics at Cornell Medical College, New York, New York. His research interest focuses on the use of PET imaging as a means to monitor response and predict outcome. His work involves imaging of tumor metabolism, hypoxia, proliferation, as well as prostate ^{68}Ga -PSMA and cardiac $^{13}\text{NH}_3$ imaging.

Joshua Scheuermann is an ABR certified physicist working at the University of Pennsylvania in the Nuclear Medicine and Radiation Oncology Departments. His primary focus in nuclear medicine has been on developing and maintaining the quality control program for clinical SPECT and PET imaging. His areas of research include quantitative accuracy in PET imaging and internal radiation dosimetry of novel radiotracers.

John Sunderland is an associate professor of radiology and the director of the PET Imaging Center at the University of Iowa. He is a PET Imaging physicist and currently co chairs the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network.

Jun Zhang is currently an assistant professor in Department of Radiology and the senior medical physicist in Wright Center of Innovation in Biomedical Imaging at The Ohio State University. He serves as the director of the Operations for Imaging and Radiation Oncology Core Group (IROC) located in Ohio and the co chair (imaging) of IROC Service Standards under the NCI National Clinical Trials Network (NCTN). His recent research is mainly in the field of solid-state digital photon counting PET/CT and its clinical translation.

Paul Kinahan received his BASc and MASc degree in engineering physics from the University of British Columbia and his PhD in bioengineering in 1994 from the University of Pennsylvania. He is a professor and vice chair for research in the Department of Radiology, University of Washington, with joint appointments in physics, radiation oncology, and bioengineering. He is the director of UWMC PET/CT imaging physics and head of the Imaging Research Laboratory.