

# Qualification of PET Scanners for Use in Multicenter Cancer Clinical Trials: The American College of Radiology Imaging Network Experience

Joshua S. Scheuermann<sup>1</sup>, Janet R. Saffer<sup>1</sup>, Joel S. Karp<sup>1</sup>, Anthony M. Levering<sup>2</sup>, and Barry A. Siegel<sup>3</sup>

<sup>1</sup>Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania; <sup>2</sup>American College of Radiology, Philadelphia, Pennsylvania; and <sup>3</sup>Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, Missouri

The PET Core Laboratory of the American College of Radiology Imaging Network (ACRIN) qualifies sites to participate in multicenter research trials by quantitatively reviewing submitted PET scans of uniform cylinders to verify the accuracy of scanner standardized uptake value (SUV) calibration and qualitatively reviewing clinical PET images from each site. To date, cylinder and patient data from 169 PET scanners have been reviewed, and 146 have been qualified. **Methods:** Each site is required to submit data from 1 uniform cylinder and 2 patient test cases. Submitted phantom data are analyzed by drawing a circular region of interest that encompasses approximately 90% of the diameter of the interior of the phantom and then recording the mean SUV and SD of each transverse slice. In addition, average SUVs are measured in the liver of submitted patient scans. These data illustrate variations of SUVs across PET scanners and across institutions, and comparison of results with values submitted by the site indicate the level of experience of PET camera operators in calculating SUVs. **Results:** Of 101 scanner applications for which detailed records of the qualification process were available, 12 (12%) failed because of incorrect SUV or normalization calibrations. For sites to pass, the average cylinder SUV is required to be  $1.0 \pm 0.1$ . The average SUVs for uniform cylinder images for the most common scanners evaluated—Siemens Biograph PET/CT ( $n = 43$ ), GE Discovery LS PET/CT ( $n = 15$ ), GE Discovery ST PET/CT ( $n = 34$ ), Philips Allegro PET ( $n = 5$ ), and Philips Gemini PET/CT ( $n = 11$ )—were 0.99, 1.01, 1.00, 0.98, and 0.95, respectively, and the average liver SUVs for submitted test cases were 2.34, 2.13, 2.27, 1.73, and 1.92, respectively. **Conclusion:** Minimizing errors in SUV measurement is critical to achieving accurate quantification in clinical trials. The experience of the ACRIN PET Core Laboratory shows that many sites are unable to maintain accurate SUV calibrations without additional training or supervision. This raises concerns about using SUVs to quantify patient data without verification.

**Key Words:** positron emission tomography; PET quantification; scanner calibration; standardized uptake value; multicenter clinical trials

**J Nucl Med 2009; 50:1187–1193**  
DOI: 10.2967/jnumed.108.057455

**T**he emerging role of imaging endpoints as in vivo biomarkers requires imaging studies to produce reliable quantitative, semiquantitative, and qualitative results that can be used to assess disease status. Ensuring such consistency in multicenter clinical trials that involve imaging is problematic, however, because data acquisition and reconstruction are performed in many different settings and often with different types of instrumentation. Accordingly, standardization of image acquisition protocols is one important approach for addressing this problem. In the specific case of PET performed for cancer imaging with <sup>18</sup>F-FDG, consensus recommendations have been promulgated by the European Organization for Research and Treatment of Cancer (1) and, more recently, by the National Cancer Institute in the United States (2). These guidelines have proven to be quite valuable in protocol development for multicenter trials involving <sup>18</sup>F-FDG PET.

Another important process for ensuring the reliability of imaging endpoints is verification that the imaging instruments themselves are performing according to specifications. The American College of Radiology Imaging Network (ACRIN) PET Core Laboratory was initially developed to ensure that individual PET scanners were properly calibrated (and were being operated in accordance with manufacturer's recommendations and the above-mentioned consensus guidelines) before being used to obtain data for ACRIN multicenter trials. To this end, ACRIN developed a qualification procedure for PET scanners at participating institutions designed to check the basic calibrations of the system and to verify that clinical image quality would be adequate for semiquantitative analysis of PET data. The purpose of

Received Aug. 25, 2008; revision accepted Mar. 4, 2009.

For correspondence or reprints contact: Joshua S. Scheuermann, HUP Nuclear Medicine, 402 Blockley Hall, 423 Guardian Dr., Philadelphia, PA 19104.

E-mail: jscheu@mail.med.upenn.edu

COPYRIGHT © 2009 by the Society of Nuclear Medicine, Inc.

this article was to report our experience with this scanner qualification process over the past 3 y.

## MATERIALS AND METHODS

### Scanner Qualification Procedures

The ACRIN PET Core Laboratory requires that a site applying for PET scanner qualification submit the following image datasets: a uniform phantom (usually cylindrical) and 2 typical clinical whole-body  $^{18}\text{F}$ -FDG PET studies (3), with patient identifying information removed from the images. The patient studies should be representative of scanner image quality and quantification but are deliberately not part of the clinical trial. For each image dataset, sites are required to submit the non-attenuation-corrected PET image series, the fully corrected PET image series, and the transmission image series used for attenuation correction (this may be a CT series or a  $^{137}\text{Cs}$  or  $^{68}\text{Ge}$  transmission series).

The uniform phantom images can be obtained either with a fillable phantom containing  $^{18}\text{F}$  diluted in water or with a  $^{68}\text{Ge}/^{68}\text{Ga}$  solid phantom (used most commonly by sites with Siemens scanners). For a phantom filled with diluted  $^{18}\text{F}$ , the concentration is similar to that present in patient studies. The phantom submission instructions specify injecting 37–55.5 MBq (1–1.5 mCi) in a 6,283-mL phantom (supplied with Siemens scanners) or 74 MBq (2 mCi) in a 9,293-mL phantom (supplied with Philips scanners) (3). These injections result in concentrations of 5.92–8.88 kBq/mL (0.16–0.24  $\mu\text{Ci}/\text{mL}$ ) for a 6,283-mL phantom and 8.14 kBq/mL (0.22  $\mu\text{Ci}/\text{mL}$ ) for a 9,293-mL phantom. A standard patient (70 kg) injected with 555 MBq (15 mCi) and scanned 1 h after injection would have an average concentration of about 5.55 kBq/mL (0.15  $\mu\text{Ci}/\text{mL}$ ) at the time the scan is started. The phantom data must be acquired using the same protocol as is used for clinical  $^{18}\text{F}$ -FDG studies (i.e., 2-dimensional [2D] vs. 3-dimensional [3D] acquisition, scan time per bed position, reconstruction method, and filter). Most GE phantom acquisitions are done in 2D mode to match the clinical workflow, whereas Philips and Siemens images are acquired in 3D mode. The phantom imaging procedure is designed so that the resultant images will have noise characteristics similar to those of clinical  $^{18}\text{F}$ -FDG PET images.

The image datasets are sent digitally to the ACRIN PET Core Laboratory either on compact disk (CD) or, preferably, by transmission over the Internet, using FTP or ACRIN software platforms designed for image transmission (Preview 32 and its successor, Triad) (4). Preview is a proprietary software application for DICOM image acquisition and transmission that was used by ACRIN for its multicenter clinical trials until December 2007.

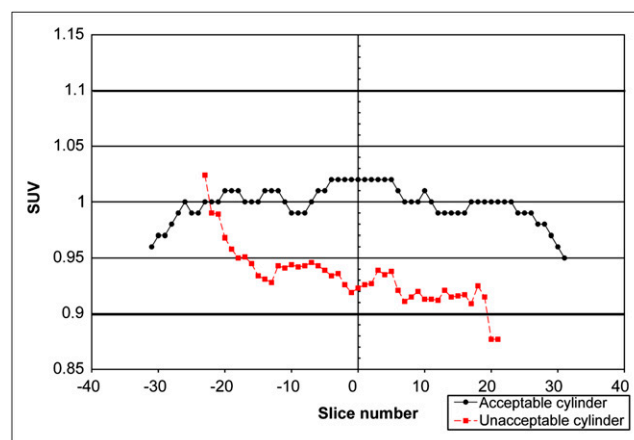
All digital data are imported onto manufacturer-specific workstations, namely a Siemens Leonardo, a GE Xeleris, or a Philips PETView workstation, depending on the type of PET scanner used for data acquisition.

The full qualification analysis has 2 parts. The first is to compare the information entered on the data sheets that accompany the application with the information contained in the DICOM headers of the phantom and test case image sets. The scanner qualification application consists of 5 pages (3). The first page contains basic information about the scanner, scanner quality control, and site and study personnel. The second and third pages require the site to record the phantom, dose, and scan acquisition information and the results of a basic region-of-interest (ROI) analysis. The fourth and fifth pages require the site to record the patient weight, height, blood glucose, dose information, and basic scan and reconstruction parameters for each test case. The patient weight, injected activity,

and assay time are all entered into the acquisition interface and are contained in the DICOM header. The information embedded in the header is compared with the information that was recorded in the application. Any discrepancies between the DICOM header and the application are investigated by contacting the site.

The second part of the analysis is the image review of both phantom and test cases. The standardized uptake values (SUVs) of the phantom are evaluated using 2D, circular ROIs that encompass approximately 90% of the interior diameter of the phantom. These ROIs are drawn on each axial slice throughout the entirety of the phantom. Because of variations in the size of phantoms used, a uniform ROI diameter was not used; however, the qualification procedure specifies that the phantom diameter must be between 18 and 22 cm and that the phantom length must be equal to or greater than the axial field of view (FOV). If the average SUV of the phantom is between 0.90 and 1.10 and does not show more than a 10% systematic variation from one end of the axial FOV to the other, the phantom results are considered to demonstrate acceptable scanner calibration. If the average SUV falls outside this range, the cause is investigated by contacting the site and troubleshooting the problem with the equipment operators. Figure 1 shows slice-by-slice results of the average SUV analysis for an acceptable uniform phantom and for one with potential normalization or calibration problems.

The review of the clinical test images is more qualitative. Images are checked to ensure that they are free of artifacts and not overly noisy and that patient positioning is reasonable. If a scan is performed on a PET/CT scanner, the PET and CT alignment is evaluated by viewing fused PET/CT images and verifying that the internal structures are overlaid properly. An SUV range is calculated in the liver by drawing 2D elliptical ROIs in several transverse slices and recording the average SUVs. If there are no major problems with the clinical test images, they will be judged as acceptable for scanner qualification. If the hepatic SUVs are outside the acceptable range (currently defined as 1.0–3.5), or the images appear to be suboptimal after technical assessment, then the underlying cause is investigated by contacting the site. In addition, if anything in the test case images or accompanying



**FIGURE 1.** Examples of uniform phantom with acceptable SUV measurements and another that shows potential problem with scanner normalization or calibration, indicated by significant axial variation of slice average SUV. Bolded horizontal lines indicate acceptable SUV limits for uniform cylinders (SUV = 0.90–1.10).

documentation (e.g., patient positioning,  $^{18}\text{F}$ -FDG dose administered, and blood glucose level) is at variance with the requirements or recommendations of the protocol in which the site is applying to participate, the site is notified.

### Review of Scanner Qualification Results

To date, the ACRIN PET Core Laboratory has reviewed cylinder and patient data for 169 PET scanners and has qualified 146 of these (86.3%). For the present study, we reviewed the records for PET qualification applications for 101 scanners (Fig. 2) submitted over 2 periods: PET qualification applications for 53 scanners between June 2005 and December 2006 and applications for 48 scanners between May 2007 and June 2008. Before June 2005, detailed records tracking the qualification process and outcome were not available. Also, for some sites that applied for qualification between January 2007 and April 2007, the details of the qualification procedure were not fully documented because of transitions in PET Core Laboratory personnel. Sites applying in these time frames were omitted from this analysis. During the study interval, all sites applying for ACRIN qualification submitted images that were reconstructed with the manufacturer's standard iterative reconstruction algorithm used for whole-body  $^{18}\text{F}$ -FDG PET for the particular scanner model. The available records allowed us to assess for the following: problems with data submission to ACRIN, problems with data importation at the ACRIN PET Core Laboratory, and reasons for initial or prolonged failure to achieve qualification. These data were cataloged into several broad categories to ascertain some of the most common problems in the application process and the most frequent reasons for qualification failure.

We also analyzed the phantoms and test images from 108 different PET scanners that passed qualification (Fig. 2). For analysis of the uniform phantom scans from these scanners, the average SUV and SD were recorded for each transverse slice. An overall average SUV for each phantom was then determined from the transverse slice data. The uniform phantoms were then grouped by scanner manufacturer and model, and the average SUV and SD were calculated for each scanner model.

For the patient image data from the 108 approved scanners, an average hepatic SUV was computed in an effort to improve the

quantitative technical assessment of the test case submissions. A range of typical liver SUVs was developed for each manufacturer across all sites from the average SUV results. This range will be used in the review of future PET qualification applications to flag studies that are likely to have quantitative inaccuracies. Only those test cases without signs of obvious hepatic lesions were analyzed. Two-dimensional, elliptic ROIs were drawn on 9 consecutive slices through the middle of the liver. The ROIs were drawn so that they encompassed the maximum area of the liver on each transverse slice while still remaining completely inside the boundaries of the liver (Fig. 3). The area of the ROI and average SUV were recorded for each transverse slice. An area-weighted average was then determined for each test case. Test cases were grouped by scanner manufacturer and model, and the mean hepatic SUV and SD were calculated for each manufacturer or model.

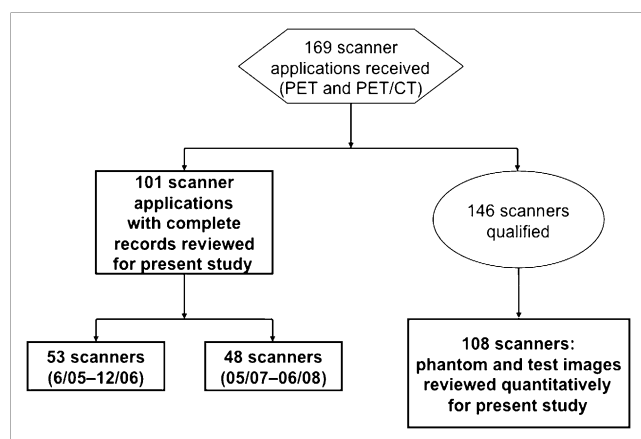
### RESULTS

Of the first 53 scanners to apply for PET qualification, 17 (32%) passed without any intervention from ACRIN, which means that the uniform phantom SUVs were in the acceptable range and all information in the headers matched the information entered on the application. Another 32 sites (60%) passed with some kind of intervention, and 4 sites (8%) failed to pass and opted not to continue the qualification process.

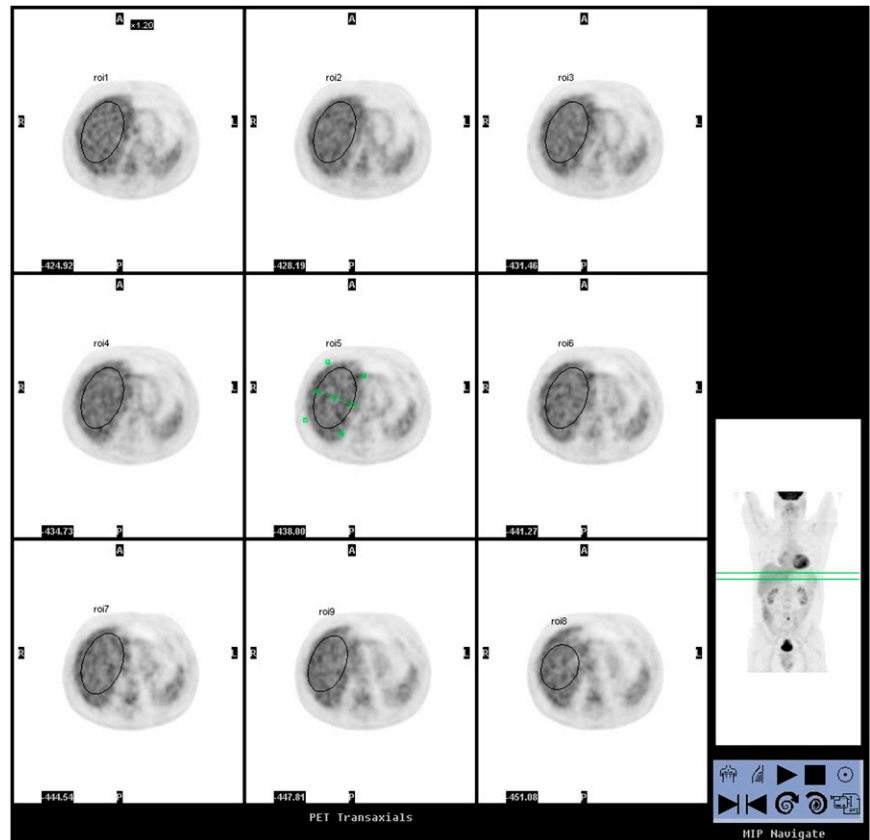
Of the more recent 48 scanners applying for PET qualification, 19 (40%) passed without any intervention from ACRIN. Another 25 sites (52%) passed with some kind of intervention, and 4 sites (8%) failed to pass and opted not to continue the qualification process. Table 1 provides a summary of the most common qualification problems encountered. To rectify the first 5 problems listed in the table, either verifying dose, patient information, or scan information with the site or requesting that the site resubmit the dataset in question was required. The final 2 problems listed, normalization calibration and SUV calibration, required the site to recalibrate the system and submit new phantom and patient datasets for qualification.

The results for the quantitative analyses of the uniform phantom images are summarized in Table 2. For the current study, the scanner models most frequently submitted for each manufacturer were reported, including Siemens Biograph PET/CT, GE Discovery LS PET/CT, GE Discovery ST PET/CT, Philips Allegro PET, and Philips Gemini PET/CT scanners. A comparison of the SUV results from Siemens Biograph scanners for  $^{18}\text{F}$ -filled phantoms and  $^{68}\text{Ge}$  solid phantoms showed that the  $^{18}\text{F}$ -filled phantom had an average SUV of  $0.99 \pm 0.04$  for 24 scanners and the  $^{68}\text{Ge}$  solid phantoms had an SUV of  $0.99 \pm 0.03$  for 19 scanners. The results show the same mean SUV for both phantom types with the fillable phantoms, suggesting that there is no bias in SUV measurement introduced using a long-lived  $^{68}\text{Ge}$  instead of an  $^{18}\text{F}$  fillable phantom.

The University of Pennsylvania provided multiple uniform phantom scans performed on the same PET/CT



**FIGURE 2.** Flow diagram analyzing relationship of 2 cohorts to each other and to total number of submitted applications.



**FIGURE 3.** Example of liver ROIs drawn on patient test images for SUV measurement.

scanner over a period of approximately 20 mo. The mean SUV of all cylinders was 0.96, with a range of 0.94–0.97. Figure 4 shows the stability of the average SUV of a uniform phantom over time.

The results for the quantitative analyses of hepatic SUVs in the test cases are shown in Table 3. For Biograph, Discovery LS, Discovery ST, Allegro, and Gemini scanners, 3, 2, 6, 2, and 3 cases, respectively, were excluded from analysis because hepatic lesions were apparent on review of the images. On average, the SUVs were highest with Siemens Biograph scanners and lowest with the Phillips scanners.

## DISCUSSION

All major models of PET scanners result in uniform cylinder acquisitions with SUVs of approximately 1.0, with Philips Gemini scanners having slightly lower values at 0.95 for the average SUV. The quantitative analysis of the liver SUVs showed that Philips systems have systematically lower SUVs than do Siemens or GE systems. Currently, there is no explanation for this disparity in average liver SUV between the Philips scanners and the scanners of the other manufacturers. Liver SUVs can vary because of many factors, such as patient weight, disease status, or uptake

**TABLE 1.** Problems Identified During Scanner Qualification in First and Second Intervals

Problem	Frequency of problem	
	Submission date, June 2005–December 2006 ( <i>n</i> = 53)	Submission date, May 2007–June 2008 ( <i>n</i> = 48)
Overanonymization of DICOM header	0 (0)	4 (0)
Incorrect information in DICOM header/on application	15 (1)	10 (0)
Clock synchronization problem	6 (0)	4 (0)
Incomplete application	2 (1)	3 (0)
Improper data format	7 (1)	1 (0)
Image display problem	0 (0)	1 (0)
Normalization calibration	5 (1)	1 (1)
SUV calibration	1 (0)	5 (3)
Total	36 (4)	29 (4)

Data in parentheses are number of scanners that failed to pass.

**TABLE 2. Uniform Phantom Mean SUVs in Relation to Scanner Model**

Scanner model	No. of scanners	Mean $\pm$ SD
Siemens Biograph	43	0.99 $\pm$ 0.03
GE Discovery LS	15	1.01 $\pm$ 0.05
GE Discovery ST	34	1.00 $\pm$ 0.04
Philips Allegro	5	0.98 $\pm$ 0.09
Philips Gemini	11	0.95 $\pm$ 0.09
Total	108	0.99 $\pm$ 0.04

Philips' software provides only 1 digit after decimal point for SUVs (i.e., 0.9, 1.0, 1.1, etc.); however, using information from DICOM header, a more precise value can be obtained.

**TABLE 3. Mean Hepatic SUVs in Relation to Scanner Model**

Scanner model	No. of scans	Mean $\pm$ SD
Siemens Biograph	83	2.34 $\pm$ 0.40
GE Discovery LS	28	2.13 $\pm$ 0.34
GE Discovery ST	62	2.27 $\pm$ 0.37
Philips Allegro	8	1.86 $\pm$ 0.62
Philips Gemini	19	1.92 $\pm$ 0.47
Total	200	2.23 $\pm$ 0.42

Philips' software provides only 1 digit after decimal point for SUVs (i.e., 0.9, 1.0, 1.1, etc.); however, using information from DICOM header, a more precise value can be obtained.

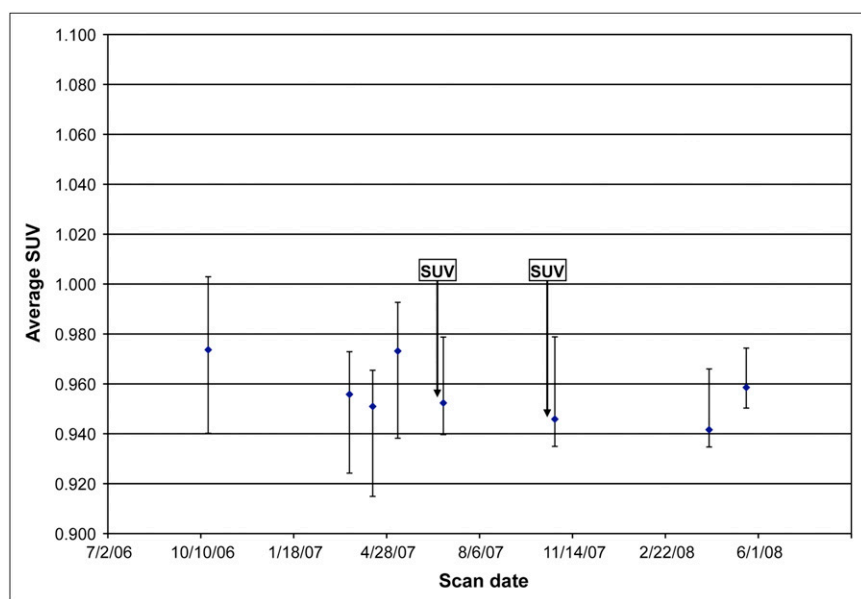
time. Average liver SUVs in different patient populations have been reported to range from 2.0 to 3.6 (5–8). On the basis of the results obtained from the quantitative analysis, a range of anticipated liver SUVs has been determined as follows: Siemens,  $2.3 \pm 1.2$ ; GE,  $2.2 \pm 1.2$ ; and Philips,  $1.9 \pm 1.3$ . Studies with values that fall outside these ranges are flagged and are more heavily scrutinized. Even if hepatic SUVs fall within the ranges, the information in the header is compared with that in the application and discrepancies are investigated. Results that fall outside the ranges always result in further investigation with the site to determine whether there is an explanation for the atypical results. The patients whose hepatic SUVs were analyzed in the present study potentially had a range of disease types at various stages and were not controlled in terms of injected dose, length of fasting before injection, uptake time, or activities during the uptake period, all of which may affect  $^{18}\text{F}$ -FDG uptake in the liver and other organs or tissues. The ranges determined in this study are for use in the qualification process to highlight submissions that may require

further inquiry, not to differentiate normal from diseased states.

SUVs are dependent on the quality of the information entered in the acquisition interface and also on the quality of the scanner calibrations. If either the input data or the calibrations are unreliable, then the SUV data obtained from the images will be suspect. Many sites are unable to supply a passing uniform cylinder (the most basic quality-control measurement for a PET scanner), raising concerns about the accuracy of SUVs in their patient studies.

There are a variety of underlying causes for a cylinder failing qualification, and the most common can be grouped into 4 broad categories: those rectified by recreating the CD or optical disk and resubmitting the data to ACRIN, those rectified by changing information in the DICOM header, those rectified by reacquiring the data, and those rectified only with a scanner recalibration.

At times, sites will submit data that are not in DICOM format (i.e., screen captures of the image sets or the data written to CD or optical disk will be in some way corrupted



**FIGURE 4.** Average SUVs of uniform phantom scanned repetitively on single PET/CT scanner over period of approximately 20 mo. Each data point represents average SUV over entire volume, and bars represent range of slice average SUVs. System was recalibrated at times noted (SUV). Over evaluation interval, system was stable.

or incomplete). Although the submitted data cannot be properly imported or analyzed at ACRIN, this kind of issue is usually rectified simply by requesting that the site resubmit the same datasets on new media or in a different format. These problems are usually specific to a particular dataset and do not indicate systematic problems that can affect the quality and validity of clinical patient scans.

Other problems that can be rectified without having to reacquire the data are those that involve incorrectly entering data into the DICOM header. The most common of these problems is when a site has entered the incorrect weight for the phantom. The proper value to enter is the weight of the water in the phantom, but some sites place the filled phantom on a scale and record that value, which results in systematically high SUVs.

Other common problems are simply mistyping the patient or phantom weight, dose, or dose assay time into the acquisition system; failing to compensate for a known offset between the clock used to record the dose assay time and the time on the scanning system; or failing to take the residual activity into account when entering the dose. These problems, although specific to a particular dataset, do affect the quality of clinical images.

Common problems that require sites to reacquire the data usually involve sites failing to record a piece of information that results in uncertainty either in the exact dose injected into the phantom or patient or in the dose assay time. The most common of these issues is when a site fails to measure the residual activity in the syringe after injecting the phantom or patient. This results in uniformly lower SUVs because the system expects more activity than is actually present. Another problem that can result in uncertainty about the dose is the presence of an offset between the clock used to record the dose assay or injection time and the acquisition system clock. If the offset is not recorded, or for some reason is not known, then the data cannot be corrected and must be reacquired. In these cases, there is nothing wrong with the actual data that were acquired and reconstructed by the system, but just with the dose information given to the system. Again, these issues can affect the quality of patient images. Failing to account for the residual activity in the syringe or an offset between the clock time used to record the dose assay or injection time and the clock time on the scanner will result in the wrong activity being used to calculate SUVs.

The final group of problems includes a bad normalization or an SUV calibration, which both require a system recalibration. These problems are suspected only after the site has submitted multiple cylinders that all failed in the same way, and no other cause can be found. In the case of a bad SUV calibration, the SUVs will be uniformly high or low; in the case of a normalization problem, the phantom will look heterogeneous in transverse or axial planes. If there is a bad scanner calibration, this will affect every patient study performed on the system. A bad SUV calibration will cause all measured SUVs to be uniformly

high or low, and a bad normalization calibration will cause variations across either or both the transverse and axial fields of view.

Figure 4 shows that scanner calibrations can be stable over time. There were 2 SUV calibrations performed in the 20-mo range displayed as a result of software upgrades and routine, preventive maintenance. However, the average cylinder SUV was consistently between 0.94 and 0.97 before and after the calibrations. Because the average SUV on a well-maintained PET system consistently falls in a small range, it is not unreasonable to require the phantom submissions of a site to be  $1.0\% \pm 10\%$  and to have less than a 10% variation across the axial FOV. Although the calibrations appear to be stable over time, it is important to check the calibrations routinely to ensure that they do not drift or abruptly change.

There are admittedly many ways to approach qualifying a PET or PET/CT scanner for participation in a research trial. ACRIN took an approach that aims at evaluating the basic scanner calibrations and ensuring that sites produce images that are of sufficient quality for nuclear medicine physicians to give a clinical interpretation. Uniform cylinder scans should be done periodically as recommended by the equipment manufacturer; the request by ACRIN should not be onerous because the phantom should be readily available and the scans should be occurring regularly. However, because most sites use a fillable phantom with diluted  $^{18}\text{F}$ , the measurements are susceptible to potential variations and errors in measuring and recording the activity in the phantom. The results show that there is not a systematic difference in the mean SUV results when using a fillable phantom versus a long-lived solid phantom. The fillable phantom provides an independent check of the system calibration and would help to avoid the propagation of errors from calibration to quality-control measurements that can occur when using the same phantom for both measurements. The performance of sites filling and analyzing a cylinder can be a useful metric for evaluating the level of experience of the operators of the PET cameras. Problems encountered acquiring data from fillable phantoms are likely to be indicators of problems acquiring clinical trial data.

The uniform cylinder scans are sensitive to basic scanner calibrations (such as SUV and normalization) and data corrections (such as scatter correction and attenuation correction) as well as the abilities of the system operators to measure and record the dose accurately and to analyze the cylinder data. They are not sensitive to more complex factors such as scanner performance factors (sensitivity, resolution, etc.) or reconstruction details (smoothing filter, resolution recovery, etc.). Phantoms being developed by the Society of Nuclear Medicine (SNM) and the American Association of Physicists in Medicine (AAPM) are aimed at reducing phantom-filling errors, using solid  $^{68}\text{Ge}/^{68}\text{Ga}$  phantoms or phantom inserts, and characterizing the lesion estimation performance of a given scanner while still testing its calibrations.

The SNM Validation Task Group developed a  $^{68}\text{Ge}$ -filled version of the NU-2 Image Quality Phantom for this purpose (9). This phantom was shipped to multiple institutions and imaged on a variety of PET scanners to quantify the repeatability and reproducibility of contrast-recovery measurements over a range of lesion sizes.

A more recent approach arising from AAPM Task group 145 is to modify the lid of the American College of Radiology PET accreditation phantom (10) by filling 4 of its cylinders with  $^{68}\text{Ge}$ . Phantoms were shipped to the 10 sites participating in the Pediatric Brain Tumor Consortium (PBTC). Each site filled the body of the phantom with enough  $^{18}\text{F}$  to establish a contrast ratio of 4:1 between the  $^{68}\text{Ge}$  cylinders and the  $^{18}\text{F}$  background and then imaged the phantom using the acquisition and processing protocols used for PBTC patient studies.

Other groups have tried to characterize the performance of PET cameras for use in specific research trials. One of the most stringent is the Alzheimer's Disease Neuroimaging Initiative, a National Institutes of Health-sponsored study in which applicants are required to submit PET scans of the 3D Hoffman brain phantom acquired on 2 separate days. There has also been an intense, standardized approach to system calibration verification and characterization of scanner performance for multicenter trials in The Netherlands. The protocol included standardization of patient preparation, dosing regimens, acquisition parameters, reconstruction settings, data analysis procedures, and quality-control procedures (11).

Although these other attempts at standard phantoms will provide additional information to better characterize the performance of a given PET scanner, unless the basic calibrations are well maintained the results of the more complex phantom scans will be meaningless in trials that use quantitative measurements. It is important to first verify that the system is well calibrated before trying to characterize scanner performance. In the setting of a multicenter research trial, scanner calibration verification via phantom data submission to a core laboratory is essential to ensuring that the same standard analysis procedures are applied across scanners. This will result in application of consistent quality-control standards across all sites participating in a multicenter research trial, yielding more consistent and reliable results.

## CONCLUSION

It has been observed that many sites are unable to produce a passing uniform cylinder dataset on the first attempt. This makes SUV validation before allowing sites to participate in multicenter research trials extremely important. The ACRIN PET Core Laboratory, with its validation phantom procedures, has ensured that scanners meet certain requirements before the scanners can be included in research trials, instilling more confidence that the data accrued over the course of the study is both quantitatively

accurate and consistent. These results highlight the importance of a central analysis of sample data before data accrued on a scanner are allowed to be used in a multicenter clinical trial. The central analysis ensures consistent handling of the data and allows for the application of standard approval criteria developed for the needs of the specific trial. Proper calibration of scanners is also essential to the use of quantitative measurements in routine clinical practice.

As trials become more complex, there is a desire to better characterize the capabilities of the equipment that is used to acquire the data for these trials. For PET scanners, this could mean requiring more complex phantoms to be scanned before permitting a site to participate in a trial. Given that some sites struggle to supply a passing uniform cylinder, research trial developers should be careful about making SUV validation too complex. Also, with the inability of some sites to reliably validate their SUV calibration, physicians reading outside studies, not acquired under their oversight, should be careful about putting too much emphasis on absolute SUVs.

## ACKNOWLEDGMENTS

We thank P. Duffy Cutler, Ramsey Badawi, and Richard Laforest for their early efforts to establish PET scanner qualification for ACRIN clinical trials. Dr. Siegel has no conflict of interest for this manuscript. Dr. Karp is principal investigator on a research agreement sponsored by Philips Healthcare. This work was supported in part by the National Cancer Institute, grants CA 080098 and CA 079778.

## REFERENCES

1. Young H, Baur R, Cremerius U, et al. Measurement of clinical and subclinical tumour response using [ $^{18}\text{F}$ ]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. *Eur J Cancer*. 1999;35:1773-1782.
2. Shankar LK, Hoffman JM, Bacharach S, et al. Consensus recommendations for the use of  $^{18}\text{F}$ -FDG PET as an indicator of therapeutic response in patients in National Cancer Institute trials. *J Nucl Med*. 2006;47:1059-1066.
3. ACRIN PET and PET/CT Scanner Qualification. Available at: <http://www.acrin.org/CORELABS/PETCORELABORATORY/PETQUALIFICATION/tabid/485/Default.aspx>. Accessed April 23, 2009.
4. Triad. Available at: <https://triad.acr.org/index.htm>. Accessed April 23, 2009.
5. Paquet N, Albert A, Foidart J, Hustinx R. Within-patient variability of  $^{18}\text{F}$ -FDG standardized uptake values in normal tissues. *J Nucl Med*. 2004;45:784-788.
6. Ramos CD, Erdi YE, Gonen M, et al. FDG-PET standardized uptake values in normal anatomical structures using iterative reconstruction segmented attenuation correction and filtered back-projection. *Eur J Nucl Med*. 2001;28:155-164.
7. Zasadny KR, Wahl RL. Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose: variations with body weight and a method for correction. *Radiology*. 1993;189:847-850.
8. Schomburg A, Bender H, Reichel C, et al. Standardized uptake values of fluorine-18 fluorodeoxyglucose: the value of different normalization procedures. *Eur J Nucl Med*. 1996;23:571-574.
9. A calibration phantom for multi-center PET/CT studies of assessing response to therapy [abstract]. *J Nucl Med*. 2007;48(suppl 2):194P.
10. MacFarlane CR. ACR accreditation of nuclear medicine and PET imaging departments. *J Nucl Med Technol*. 2006;34:18-24.
11. Boellaard R, Hoekstra CJ, et al. The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *Eur J Nucl Med Mol Imaging*. 2008;35:2320-2333.