Journal of Nuclear Medicine, published on March 2, 2017 as doi:10.2967/jnumed.116.186759

Qualification of NCI-Designated Cancer Centers for Quantitative PET/CT Imaging in Clinical Trials

Joshua S. Scheuermann[1], Janet S. Reddin[1], Adam Opanowski[2], Paul E. Kinahan[3], Barry A Siegel[4], Lalitha K Shankar[5], Joel S. Karp[1]

[1] Dept. of Radiology, University of Pennsylvania, Philadelphia, PA, USA

[2] American College of Radiology Imaging Network, Philadelphia, PA, USA

[3] Dept. of Radiology, University of Washington, Seattle, WA, USA

[4] Mallinckrodt Institute of Radiology and the Alvin J. Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO, USA

[5] Cancer Imaging Program, National Cancer Institute, Rockville, MD, USA

Corresponding Author:

Joshua S Scheuermann, MMP, DABR
John Morgan Building, Rm. 151A
3620 Hamilton Walk
Philadelphia, PA 19104
Phone: 215-573-7555

Funding Disclosure:

This work was supported in part by NCI-SAIC Subcontract #10XS070, NIH grants U01CA148131, U01CA190254, U10CA180820, and NIH-NCI contract HHSN261200800001E.

Word Count: 5572

Abstract

The National Cancer Institute (NCI) developed the Centers for Quantitative Imaging Excellence (CQIE) initiative in 2010 to pre-qualify imaging facilities at all of the NCI-designated Comprehensive and Clinical Cancer Centers for oncology trials using advanced imaging techniques, including positron emission tomography (PET). This paper reviews the CQIE PET/CT (Computed Tomography) scanner qualification process and results in detail.

Methods: Over a period of approximately 5 years, sites were requested to submit a variety of phantom, including uniform and ACR (American College of Radiology) phantoms, PET/CT images, as well as examples of clinical images. Submissions were divided into 3 distinct time points: initial submission (T0), followed by two requalification submissions (T1 and T2). Images were analyzed using standardized procedures and scanners received a pass or fail designation. Sites had the opportunity to submit new data for failed scanners. Quantitative results were compared: across scanners within a given time point and across time points for a given scanner.

Results: 65 unique PET/CT scanners across 42 sites were submitted for CQIE T0 qualification, with 64 passing qualification. 44 (68%) of the scanners from T0 had data submitted for T2. From T0 to T2 the percentage of scanners passing the CQIE qualification on the first attempt rose from 38% in T1 to 67% in T2. The most common reasons for failure were: standardized uptake value (SUV) out of specifications, incomplete data submission and uniformity issues. Uniform phantom and ACR phantom results between scanner manufacturers are similar.

Conclusions: The results of the CQIE process show that periodic requalification may decrease the frequency of deficient data submissions. The CQIE project also highlighted the concern within imaging facilities about the burden of maintaining different qualifications and accreditations. Finally, we note that for quantitative imaging-based trials the relationships between the level of the qualification (e.g., bias or precision) and the quality of the image data, accrual rates, and study power needs further evaluation.

Keywords: CQIE, PET Qualification, Quantitative Imaging

INTRODUCTION

Increasingly, PET/CT with ^{18}F-FDG and other radiopharmaceuticals is being used as a quantitative imaging biomarker in oncology to assess treatment efficacy (*1-8*). While there are many factors that influence quantitative accuracy, both physiological and instrumental (*9–12*), the ability to determine treatment efficacy based on PET is predicated on the ability of PET scanners to provide stable measurements of radiotracer concentrations, thus allowing treatment response to be tracked over months or years. Short-term scanner variability is expected to be low, and long-term scanner variability can likely be minimized with standardized quality control procedures (*13,14*), although recently it has recently been shown that long-term stability cannot be taken for granted, and should be checked (*15,16*).

An additional complication arises when accrual for research trials is accelerated by expanding imaging to multiple centers. Because of differences in procedures at imaging sites, some PET studies may not be quantitatively reliable or the data may not be useable in a pooled analysis. It has been reported that approximately one-third of PET studies acquired at community-based imaging facilities may lack the necessary information to obtain quantitative imaging data (*17*). In addition, the quantitative variability in multi-center trials can be expected to be larger than in single-center trials (*18,19*). In a study at a single institution with multiple PET scanners, which were clinically accredited and maintained according to manufacturer standards by qualified staff, it was shown that the variance of PET measurements is greater in clinical practice than under ideal study settings (*20*).

These reports of PET scanner bias and variability highlight the need for standardized qualification processes to minimize variability in multi-center research trials. However, the need for qualifying imaging systems prior to participation in research trials increases the time needed

to accrue the trial data, since it takes time to acquire the qualification data, send it to the study sponsor for analysis and get approval to participate in the trial. Methods for quantitative qualification of PET scanners range from testing accuracy and basic image quality (*21*) to prospectively assessing accuracy and contrast recovery and developing scanner-specific reconstruction parameters to unify contrast recovery coefficients across scanners (*22*,*23*). These efforts have resulted in guidelines for tumor imaging using ^{18}F-FDG PET in research trials (*24*–*26*).

Based on this understanding of the importance of quantitative accuracy of imaging biomarkers in clinical trials, the NCI developed the CQIE initiative in 2010 to pre-qualify imaging facilities at all of the NCI-designated Comprehensive and Clinical Cancer Centers for oncology trials using x-ray computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET). The intention of the CQIE project was to have a group of 'trial-ready' cancer imaging facilities to minimize the time between a multi-center research trial being developed and sites beginning accrual. An overview of the CQIE program for all three imaging modalities is provided elsewhere (*27*). This paper reviews the CQIE PET/CT qualification process and results in more detail.

MATERIALS AND METHODS

Study timeframe

The CQIE data submissions occurred over the period from June 2010 – March 2014. They were divided into three time periods: T0, T1 and T2. T0 was the baseline qualification and ran from June 2010 – December 2011. This was followed by (roughly) annual requalification

submissions for T1 and T2. T1 ran from January 2012 – December 2012. T2 ran from January 2013 – March 2014.

PET qualification procedure

Sites were provided with detailed instructions and training modules. If necessary, the test phantoms were provided. Sites were then required to submit a series of seven phantom and example patient images for review: uniform phantom acquired using a static body protocol; uniform phantom using a static brain protocol; uniform phantom using a dynamic body protocol; ACR phantom using a static body protocol; ACR phantom using a static brain protocol; two anonymized patient brain scan example cases; and two anonymized patient body scan example cases. In an effort to remain consistent with the existing ACRIN qualification and ACR accreditation programs, sites were requested to reconstruct all static body and brain images using their standard clinical reconstruction protocols.

Each site was requested to select one scanner for qualification. For the initial T0 period qualification tests, imaging centers were given the option of an on-site visit by a member of the CQIE qualification team to facilitate scanning and qualification. Phantoms (if needed) and CQIE standard operating procedure materials were then forwarded to the site within two weeks of the planned site visit. Methods for image transfer via secure file transfer protocol were also established at this time. Sites were encouraged to complete a review of learning modules describing the importance of the program prior to their site visit. For the T1 and T2 qualification tests, there were no visits by members of the CQIE qualification team.

Uniform phantom data acquisition and analysis

The uniform phantom data sets were based on a standard fillable cylinder without features, nominally 20 cm in diameter and length. The uniform phantom was filled with a dilute F-18 solution to a concentration of roughly 5.00 – 6.11 kBq/mL (135 to 165 nCi/mL). For the brain data set, the phantom was scanned for 1 bed position using the site's standard clinical brain acquisition and reconstruction protocols. For the body data set, the phantom was scanned for 2 bed positions using the site's standard clinical body acquisition and reconstruction protocols. The sites were asked to measure and report the SUV for a large central region of interest (ROI).

For the dynamic phantom study, since many sites do not routinely perform dynamic imaging, the site was required to acquire the phantom using one bed position and a specific timing protocol. The images were then reconstructed with a protocol that likely would be used for dynamic scanning. The results from this component are reported elsewhere and are not further discussed here (*27,28*).

The uniform phantom images were transferred to a central core laboratory and imported into an Osirix display platform (*29*) for analysis. For each phantom data set, several fields in the DICOM (Digital Imaging and Communications in Medicine) headers were compared to the site-reported data forms to verify the accuracy of the SUV calculations. These fields include: AcquisitionTime, Weight, RadiopharmaceuticalStartTime, and RadionuclideTotalDose. If these fields were found to be inconsistent, the site was contacted and asked to clarify the discrepancy. Once any DICOM header discrepancies were resolved, a 200 cm^2 (approximately 16 cm in diameter), circular ROI was placed on every axial slice in the phantom and the mean and standard deviation of the SUV for each voxel within the ROI were recorded. For the static brain and body phantoms, axial slices up to 1.5 cm from the axial edge of the field of view (FOV) or

the edge of the phantom were excluded from the analysis because of the typical fall off at the axial edge of the FOV and potential edge effects near the ends of the phantom.

For the static body and brain protocol phantom acquisitions, a Volume Average SUV was computed by taking the average of the Mean SUVs from each axial slice. The Maximum Axial Deviation also was calculated by finding the difference between the Maximum Mean Slice SUV and the Minimum Mean Slice SUV and dividing by the Volume Average SUV. For a static uniform phantom data set to pass the quantitative analysis the Volume Average SUV had to be between 0.90 and 1.10 and the Maximum Axial Deviation had to be < 10%. If the results were outside of these specifications the site was contacted to try to determine the reason for the failure and to resolve the problem.

ACR phantom data acquisition and analysis

The ACR PET phantom (*30*) contains a series of 4 ‘hot’ contrast cylinders with diameters of 8, 13, 17 and 25 mm, each 25 mm long, with a nominal cylinder:background ratio of 4:1. The phantom was filled according to the standard instructions (*30*), assuming a 444 MBq (12 mCi) patient dose, and scanned using the same body and brain protocols as were used for the uniform phantom. The ACR phantom was chosen for this study because it is a common phantom in use by many clinical centers, thus obviating the purchase of a more complicated phantom.

In addition to the phantom data sets, sites were required to submit 2 anonymized brain patient scans and 2 anonymized body patient scans. The patient test cases were acquired with the sites' standard clinical protocols. In addition the image acquisition and reconstruction parameters used by the site were recorded.

The ACR phantom data were transferred and imported into Osirix for analysis. The same comparison of DICOM headers done for the uniform phantom was done for each of the ACR phantom datasets The SUV analysis followed the ACR instructions. First, image planes were summed together to form images that were between 9 mm and 12 mm thick. The slice that best showed the four "hot" cylinders was selected, then circular background and small cylinder ROIs were drawn, with the background ROI about 6–7 cm in diameter in the center of the image and the small cylinder ROI drawn just inside the largest hot cylinder. Copies of this smaller (<25 mm diameter) ROI were drawn over the other hot cylinders and over the air, water, and bone cylinders. As part of the core laboratory analysis, we also recorded the SUVpeak measurement for each of the hot cylinders. The SUVpeak was defined as the average SUV in a 1.0 cm diameter circular ROI centered on the maximum pixel in each cylinder. The recovery coefficient (Recovery Coefficient = measured/true) for the hot cylinders was plotted as a function of the cylinder diameter.

Example patient test cases analysis

For the example patient brain and body test cases, the DICOM headers were reviewed for accuracy and compared to the data forms. Any discrepancies were investigated with the site to determine the source of the discrepancy. Once any discrepancies were resolved, a qualitative review of overall image quality was performed. The fusion between the PET and CT images was checked, the patient positioning in the FOV was evaluated and the appropriateness of the acquisition and reconstruction settings were evaluated based on the overall qualitative smoothness of the PET images. For the brain test cases SUVs were not recorded, but the ability to measure SUVs was verified. For the body test cases, SUV analysis of the liver was performed.

A large, 2D, elliptical ROI was drawn on 7 consecutive transverse slices through the middle of the liver as illustrated in Figure 1. The mean SUV for each 2D ROI was recorded and an area-weighted average of the means was computed to determine the average liver SUV for each test case.

T0 vs. T2 qualification summary comparison

For all of the T0 scanners reviewed, it was determined how many scanners passed the qualification review (1) without any failure, or intervention from the core lab, (2) with a single failure, or (3) with multiple failures. A failure was considered any issue that prevented the scanner from passing the qualification review without CQIE interacting with the site. The reasons for failure also were catalogued. They were divided into the following categories: uniformity phantom problem, SUV out of specification, phantom filling issue, reconstruction problem, improper acquisition, incomplete submission, and problem with data forms.

For the T2 scanners, the same analysis of qualification results was performed for those systems that also submitted at T0.

RESULTS

Accrual

T0 period: A total of 65 PET scanners underwent CQIE testing during the initial period. The majority of these sites opted for on-site visits by the PET CQIE team.

T1 period: Sites that participated in T0 period were sent reminders that requalification was needed. No specific follow-up was undertaken. In addition, no on-site visits were provided, and sites were not given access to funds to defray costs associated with scanner qualification.

Site participation decreased dramatically in year two relative to year one. The year two participation rate dropped to 39 scanners, for some of which data were not submitted for the T0 time point. The T2 participation rate increased to 52 scanners, with 44 also having submitted data for T0. Because of poor data accrual during the T1 period, only data from T0 and T2 are reported.

Data from 65 unique PET/CT scanners at 56 sites were analyzed for T0. Of the 65 scanners, one scanner could not be reviewed because the images could not be submitted in DICOM format. For the T2 period, data were analyzed for the 44 PET/CT scanners that were also submitted for the T0 period.

Uniform cylinder results

An example plot of the mean SUV for each image plane is shown in Figure 2, illustrating the calculation of the maximum axial deviation (MAD). MAD is a surrogate for evaluating the quality of the system normalization. We have found empirically that a re-normalization decreases the MAD and improves the flatness of the "axial" profile. Also shown is a typical roll-off of the mean SUV per image plane at the axial ends of the FOV, and the inclusion region for calculation of the MAD.

Average SUV and MAD results for the uniform cylinder acquired for time period T0 using Static Brain and Body protocols are summarized in Tables 1 and 2.

ACR phantom results

Figure 3 is a typical image of the ACR PET phantom showing the ROIs used for analysis. The contrast recovery coefficients plotted as a function of cylinder diameter for each of the three

PET scanner manufacturers for the Static Brain and Body acquisitions are shown in Figure 4. The error bars represent the standard deviation for each data point.

Example brain test case results

In some cases, a field in the DICOM header was not properly populated preventing SUV calculation. The most common reasons for this were as follows: the anonymization routine removed a required field; the operator did not enter a required data field in the acquisition interface; or a required DICOM field was changed or removed at some point in the processing.

Example body test case liver SUV results

Table 3 contains the results of the liver SUVs for the body test cases from the T0 period, showing what appears to be a divergence of average liver values between manufacturers. Despite careful evaluation of all aspects of the image acquisition protocols and processing chain, no systematic cause was found.

T0 vs. T2 comparison

For T0, 25 (38%) scanners passed without any core laboratory intervention, 30 scanners (46%) passed after the second submission, and 9 scanners (14%) required more than two submissions in order to pass (Table 4). The most common problem was that SUV results were out of specifications, followed by incomplete submissions (Table 5). For the 50 issues that were cataloged, 21 were likely linked to system calibration problems (uniformity problems and SUV out of specification). The remaining 29 issues were related to operator error. Note that the total

number of issues is not the same as the number of scanner disqualifications because some qualification attempts had multiple issues.

For the T2 period, data were submitted for 44 scanners that also were qualified for T0. All 44 scanners eventually passed the qualification review, with 31 passing without any CQIE core laboratory intervention. Table 6 shows a comparison of T0 and T2 results after the initial review in the core laboratory. 11 of 17 scanners that passed initially during T0 passed without any intervention during T2. 20 of 27 scanners that initially did not pass during T0 passed without any intervention during T2.

DISCUSSION

Our primary finding is that in the T2 period, there was a reduced frequency of qualification issues compared the T0 period (Tables 4-6). This indicates that a consistent scanner qualification process helps to ensure standardized scanner performance throughout the entirety of a trial. In the T0 period, there were a total of 50 issues identified with the data submissions for the 65 scanners. Quantification problems, which can be due to system calibration problems, accounted for 21 of the 50 issues (42%). The other 29 issues (58%) were attributable to user errors, which should be reduced with training. In the T2 period, there were only 14 issues identified with the data submission for 44 scanners; 3 of the 14 (21%) were quantification problems and the remaining 11 (79%) were user error. The lower overall rate of issues with submissions in the T2 period likely indicates that the sites better understood the submission process and were more comfortable with the requirements, leading to fewer mistakes and omissions. The lower rate of quantification problems may also indicate that sites were more

familiar with the analysis performed in the core laboratory and the passing criteria employed, so they were less likely to submit data that failed their internal analysis. Understanding the data analysis and passing criteria may also make them more sensitive to changes in performance and more likely to address potential problems with quantitative imaging sooner. However, approximately 7% (3 of 44) scanners had quantification issues that required a recalibration, which points to the need for periodic requalification.

The quantitative phantom results were mostly consistent between manufacturers. The uniform phantom results showed the average SUVs to be within one standard deviation of the expected value of 1. The MAD for the body FOV cylinders was consistent between manufacturers, but for the brain FOV, the MAD for Philips systems was higher than for the other 2 manufacturers. This may be related to the lack of post-processing smoothing, which is used by GE and Siemens, but not by Philips. The ACR phantom results for both brain and body were consistent between manufacturers. There were some differences between manufacturers, as seen in Figures 1 and 2, but the differences were within the error bars. In general, the quantitative differences between manufacturers are small.

For the current project, the inclusion of the ACR phantom did not appear to add value to the qualification process, because all scanners that passed the uniform cylinder analysis passed the ACR phantom analysis, unless there was a phantom filling problem. This could be due to the relatively wide acceptance criteria currently used for standard ACR submissions of clinical PET scanners, which were adopted for CQIE qualification. It may be appropriate to use tighter acceptance criteria for the ACR phantom to better assess differences in contrast recovery that may arise from different reconstruction parameters. This points to an unresolved issue: While it is clear some level of qualification and routine quality assurance/quality control (QA/QC) should

be included in clinical trials using quantitative imaging, the relation between the type and degree of the qualification (and QA/QC procedures) on the quality (e.g., bias or precision) of the image data has not been established. Thus, in cases where a higher degree of variance or bias in imaging data could be tolerated without affecting study power, imaging-based trials may use excess resources for qualification and QA/QC to improve data quality unnecessarily, or alternatively exclude sites unnecessarily, which in turn slows down accrual. However, the opposite scenario is also possible, where less rigorous qualification and QA/QC policies may allow for increased accrual rates, but at the cost of under-powering the study due to increased signal variation. Optimizing QA/QC procedures for imaging trials can, in theory, shift the power or accrual rate vs. cost curve substantially towards better optimized imaging trials. These trade-offs, which affect accrual and study power, need more evaluation.

The CQIE project suffered from a lack of accrual during the T1 period. This may have been due to some imaging facilities giving the CQIE qualification a low priority, which led to an increased number of incomplete or faulty submissions and difficulty resolving issues. During the T2 period the CQIE program intensified the requests for qualifications, which led to increased accrual. These issues also applied to the MRI and CT CQIE qualifications, and are discussed in more detail in the CQIE overview (*27*).

Some sites also expressed their hesitation to commit to another qualification regimen because of the time requirements to maintain other qualifications and accreditations, such as those from the ACR, ECOG-ACRIN, SNMMI, individual study sponsors, etc. Committing to another qualification program was seen as overly burdensome.

The ACR, ECOG-ACRIN, SNMMI and CQIE programs all require that sites submit phantom and patient image data to a core laboratory for analysis. The ACR and SNMMI

programs require that a fee be paid for accreditation/qualification, but the ECOG-ACRIN and CQIE do not require fees for qualification. ECOG-ACRIN uses a uniform cylinder phantom for qualification, which should be provided by the manufacturer for all scanners. ACR and SNMMI use more complicated resolution phantoms, with the ACR using a cylindrical phantom with fillable cylinders attached to the lid and the SNMMI using a chest simulator phantom that is more anthropomorphic and has spheres embedded throughout.

Unifying qualification and accreditation criteria between qualifying agencies and clinical research organizations would likely require the adoption of different levels of qualification. Depending on the specific aims of a given trial, more or less variability may be acceptable, which would require more or less rigorous scanner qualification. Unifying qualification programs would require the adoption of standard uniformity and contrast phantoms industry-wide, and agreement on filling and scanning procedures, analysis methodology and passing criteria. Given the time and resources already invested into various qualification and accreditation programs, it will be difficult to develop a single qualification methodology acceptable to all organizations. However, if research organizations were more transparent about their qualification programs, including the specifics of the analyses being performed and passing criteria, other organizations could more easily evaluate the qualification needs for their trials compared to the rigorousness of the various qualification programs and choose to accept specific organizations' qualifications in lieu of their specific qualification process.

CONCLUSION

The results of the CQIE process show that periodic requalification may decrease the frequency of deficient data submissions. This suggests that, as sites become more aware of the

qualification process and passing criteria, they will be more likely to address problems before submission of qualification data to a core laboratory or study sponsor.

The CQIE project also highlighted the concern within imaging facilities about the burden of maintaining different qualifications and accreditations. Discussions with personnel at various facilities emphasized the need to develop a common set of qualification criteria across various research organizations to reduce the burden on imaging facilities of participating in many different clinical trials. This may encourage facilities to participate in a greater number of multi-center clinical trials.

Finally, we note that for quantitative imaging-based trials the relationships between the level of the qualification (e.g., bias or precision) and the quality of the image data, accrual rates, and study power needs further evaluation.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of the following individuals for their time and effort in support of this publication: Mark Rosen, Deborah Harbison, Joseph Maffei, Mark Muzi, Finbarr O'Sullivan, Brian Elston, Darin Byrd, and Mattie Lloyd. This work was supported in part by NCI-SAIC Subcontract #10XS070, NIH grants U01CA148131, U01CA190254, U10CA180820, and NIH-NCI contract HHSN261200800001E.

REFERENCES

1. Pantel AR, Mankoff DA. Molecular imaging to guide systemic cancer therapy: illustrative examples of PET imaging cancer biomarkers. *Cancer Lett.* 2017;387:25-31.

2. Bengtsson T, Hicks RJ, Peterson A, Port RE. ^{18}F-FDG PET as a surrogate biomarker in non-small cell lung cancer treated with erlotinib: newly identified lesions are more informative than standardized uptake value. *J Nucl Med.* 2012;53:530-537.

3. Cornelis F, Storchios V, Violari E, et al. ^{18}F-FDG PET/CT is an immediate imaging biomarker of treatment success after liver metastasis ablation. *J Nucl Med*. 2016;57:1052-1057.

4. Humbert O, Riedinger JM, Charon-Barra C, et al. Identification of biomarkers including ^{18}FDG-PET/CT for early prediction of response to neoadjuvant chemotherapy in triple-megative breast cancer. *Clin Cancer Res*. 2015;21:5460-5468.

5. Kim JW, Oh JS, Roh JL, et al. Prognostic significance of standardized uptake value and metabolic tumour volume on ^{18}F-FDG PET/CT in oropharyngeal squamous cell carcinoma. *Eur J Nucl Med Mol Imaging*. 2015;42:1353-1361.

6. Kostakoglu L, Duan F, Idowu MO, et al. A phase II study of 3'-deoxy-3'-^{18}F-fluorothymidine PET in the assessment of early response of breast cancer to neoadjuvant chemotherapy: results from ACRIN 6688. *J Nucl Med*. 2015;56:1681-1689.

7. Pimiento JM, Davis-Yadley AH, Kim RD, et al. metabolic activity by ^{18}F-FDG-PET/CT is prognostic for stage I and II pancreatic cancer. *Clin Nucl Med*. 2016;41:177-181.

8. Siva S, Deb S, Young RJ, et al.. ^{18}F-FDG PET/CT following chemoradiation of uterine cervix cancer provides powerful prognostic stratification independent of HPV status: a

prospective cohort of 105 women with mature survival data. *Eur J Nucl Med Mol Imaging*. 2015;42:1825-1832.

9. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. *AJR Am J Roentgenol*. 2010;195:310-320.

10. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med.* 2009;50(Suppl 1):11S-20S.

11. Kinahan PE, Fletcher JW. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. *Semin Ultrasound CT MR*. 2010;31:496-505.

12. Shankar LK, Hoffman JM, Bacharach S, et al. Consensus recommendations for the use of ^{18}F-FDG PET as an indicator of therapeutic response in patients in National Cancer Institute Trials. *J Nucl Med*. 2006;47:1059-1066.

13. Doot RK, Scheuermann JS, Christian PE, Karp JS, Kinahan PE. Instrumentation factors affecting variance and bias of quantifying tracer uptake with PET/CT. *Med Phys*. 2010;37:6035-6046.

14. Doot RK, Thompson T, Greer BE, et al. Early experiences in establishing a regional quantitative imaging network for PET/CT clinical trials. *Magn Reson Imaging*. 2012;30:1291-1300.

15. Byrd D, Doot RK, Allberg KC, et al. Evaluation of cross-calibrated ^{68}Ge/^{68}Ga phantoms for assessing PET/CT measurement bias in oncology imaging for single center and multicenter trials. *Tomography*. 2016;2:353-360.

16. Doot RK, Pierce LA, Byrd D, Elston B, Allberg KC, Kinahan PE. Biases in multicenter

longitudinal PET standardized uptake value measurements. *Transl Oncol.* 2014;7:48-54.

17. Tahari AK, Wahl RL. Quantitative FDG PET/CT in the community: experience from interpretation of outside oncologic PET/CT exams in referred cancer patients. *J Med Imaging Radiat Oncol*. 2014;58:183-188.

18. Fahey FH, Kinahan PE, Doot RK, Kocak M, Thurston H, Poussaint TY. Variability in PET quantitation within a multicenter consortium. *Med Phys*. 2010;37:3660-3666.

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

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

21. Scheuermann JS, Saffer JR, Karp JS, Levering AM, Siegel BA. Qualification of PET scanners for use in multicenter cancer clinical trials: the American College of Radiology Imaging Network experience. *J Nucl Med*. 2009;50:1187-1193.

22. Barrington SF, Mackewn JE, Schleyer P, et al. Establishment of a UK-wide network to facilitate the acquisition of quality assured FDG-PET data for clinical trials in lymphoma. *Ann Oncol*. 2011;22:739-745.

23. Boellaard R, Oyen WJ, 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.

24. Boellaard R, Delgado-Bolton R, Oyen WJ, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42:328-

354.

25. FDG-PET/CT Technical Committee. FDG-PET/CT as an imaging biomarker measuring response to cancer therapy, Quantitative Imaging Biomarkers Alliance, Version 1.13. http://qibawiki.rsna.org/images/1/1f/QIBA_FDG-PET_Profile_v113.pdf. Published November 18, 2016. Accessed January 17, 2017.

26. Graham MM, Wahl RL, Hoffman JM, et al. Summary of the UPICT protocol for ^{18}F-FDG PET/CT imaging in oncology clinical trials. *J Nucl Med*. 2015;56:955-961.

27. Rosen M, Kinahan PE, Gimpel J, et al. Performance observations of scanner qualification of NCI-designated Cancer Centers: results from the Centers of Quantitative Imaging Excellence (CQIE) program. *Academic Radiology*. November 29, 2016. [Epub ahead of print].

28. Mou T, Huang J, Zhang Y, et al. Spatial covariance characteristics in a collection of 3-D PET scanners used in clinical imaging trials. 2014 IEEE Nuclear Science Symposium and Medical Imaging Conference 1-3, 2014.

29. Rosset A, Spadola L, Ratib O. OsiriX: an open-source software for navigating in multidimensional DICOM images. *J Digit Imaging*. 2004;17:205-216.

30. MacFarlane CR, Radiologists ACO. ACR accreditation of nuclear medicine and PET imaging departments. *J Nucl Med Technol*. 2006;34:18-24.

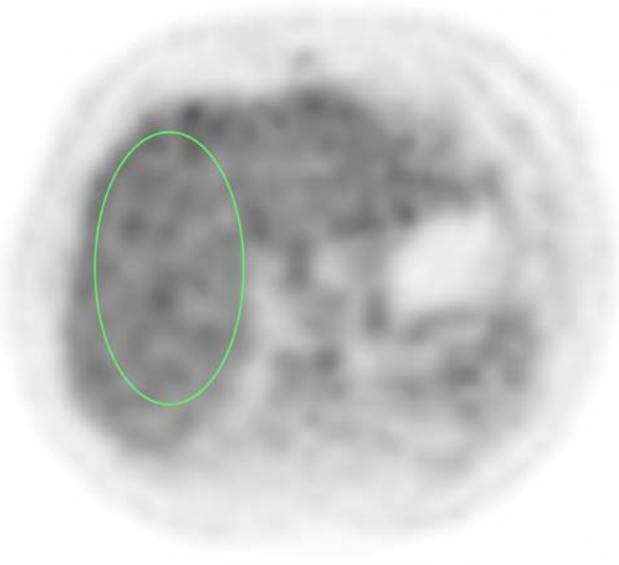

Figure 1. Illustration of 1 of the 7 adjacent ROIs in liver regions used in the body test cases.

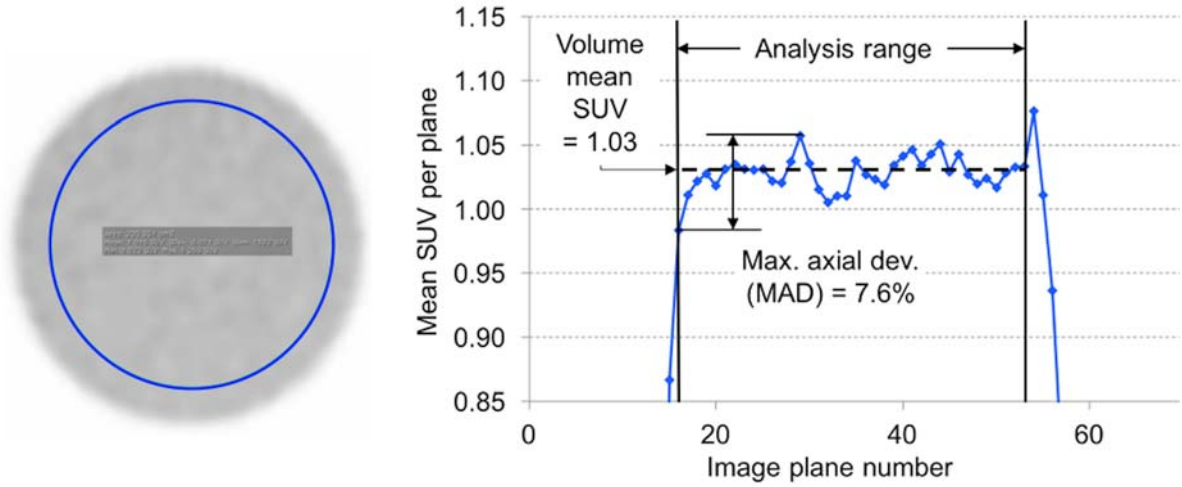

Figure 2. Left: Typical ROI used for uniformity analysis. Right: Example of a typical plot of mean SUVs per plane in the ROI and the calculation of the maximum axial deviation (MAD).

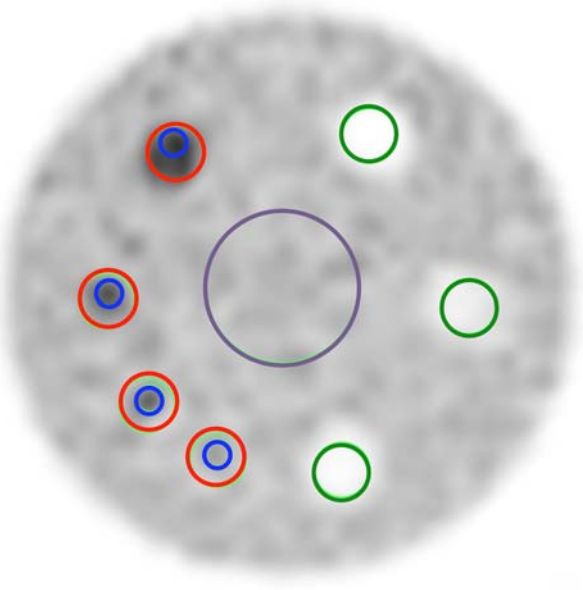

Figure 3. Image of the ACR PET phantom. Red ROIs are used for SUVmax and blue ROIs are used for SUVpeak measures. Also shown are a large background ROI and smaller green ROIs for the cold cylinders. The latter ROIs were not used in this analysis.

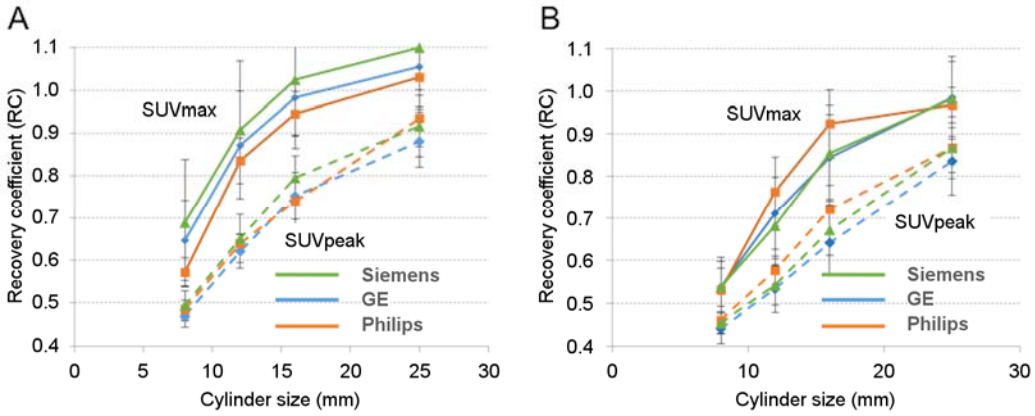

Figure 4: Recovery coefficients (defined as ratio of measured/true SUV) as a function of cylinder diameter for static brain (A) and body (B) acquisitions.

Table 1. Uniform cylinder results for static brain protocol for period T0.

Manufacturer	N	Vol. Avg. SUV	MAD
GE	36	1.00 ± 0.03	2.9% ± 1.2%
Philips	7	0.99 ± 0.03	8.5% ± 0.7%
Siemens	21	0.98 ± 0.05	3.4% ± 1.4%
Combined	64	0.99 ± 0.04	3.7% ± 2.1%

Table 2. Uniform cylinder results for static body protocol for period T0

Manufacturer	N	Vol. Avg. SUV	MAD
GE	36	1.00 ± 0.03	5.1% ± 1.8%
Philips	7	0.97 ± 0.04	5.4% ± 2.0%
Siemens	21	1.00 ± 0.03	5.5% ± 2.1%
Combined	64	1.00 ± 0.03	5.3% ± 1.9%

Table 3: Average liver SUVs by manufacturer.

Manufacturer	# of Cases	Liver SUV	Std. Dev.
GE	62	2.11	0.44
Philips	14	2.06	0.43
Siemens	42	2.42	0.51

Table 4. Differences in scanner qualification for the three time periods.

	T0		T1		T2	
# of Attempts to Pass	Number	Percentage	Number	Percentage	Number	Percentage
First Time	25	38%	34	87%	35	67%
Eventually	39	61%	5	13%	13	25%
Total Passing	64	98%	39	100%	48	92%
Total Scanners	65		39		52	

Table 5: Frequency of specific issues during scanner qualification. Issues are ranked by subjective order of relative importance. Uniformity problem and SUV out of specification are considered calibration issues, while the rest are attributable to operator error.

Issue	T0 (65 scanners)	T0 scanners in T2 (44 scanners)
Uniformity problem	7	0
SUV out of specification	14	3
Phantom filling issue	4	3
Reconstruction problem	6	5
Improper acquisition	3	0
Incomplete submission	11	2
Problem with forms	5	1
Total	50	17

Table 6: Cross-comparison by scanner of passing status on initial review for T0 and T2 periods for the 44 scanners that were in both qualification reviews.

		T2 period		
	Pass?	Yes	No	Total
T0 period	Yes	11	6	17
	No	20	7	27
	Total	31	13	44