Journal of Nuclear Medicine, published on December 18, 2014 as doi:10.2967/jnumed.114.148056

Quantitative PET/CT scanner performance characterization based upon the SNMMI Clinical Trials Network oncology clinical simulator phantom

John J. Sunderland[1], Paul E. Christian[2]

[1]Department of Radiology, Carver College of Medicine, University of Iowa, Iowa City, IA

[2]Center for Quantitative Cancer Imaging, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT

Corresponding Author:

John J. Sunderland

200 Hawkins Drive

University of Iowa

Iowa City, IA 52242

Phone: 319-356-1092

Fax: 319-353-6512

E-Mail: john-sunderland@uiowa.edu

Short running foot line: Quantitative scanner characterization

Financial Support: Provided by the Clinical Trials Network of the Society of Nuclear Medicine and Molecular Imaging

Word Count: 5926

Abstract

The Clinical Trials Network (CTN) of the Society of Nuclear Medicine and Molecular Imaging (SNMMI) operates a PET/CT phantom imaging program using the CTN's oncology clinical simulator phantom, designed to validate scanners at sites that wish to participate in oncology clinical trials. Since its inception in 2008 the CTN has collected 406 well-characterized phantom data sets from 237 scanners at 170 imaging sites covering the spectrum of commercially available PET/CT systems. The combined and collated phantom data describes a global profile of quantitative performance and variability of PET/CT data used in both clinical practice and clinical trials.

Methods

Individual sites filled and imaged the CTN oncology PET phantom according to detailed instructions. Standard clinical reconstructions were requested and submitted. The phantom itself contains uniform regions suitable for scanner calibration assessment, lung fields, and six hot spherical lesions with diameters ranging from 7-20mm at a 4:1 contrast ratio with primary background. The CTN Phantom Imaging Core assessed the quality of the phantom fill and imaging, measured background SUV for scanner calibration assessment, and SUV_{max} values of all six lesions to assess quantitative performance. Scanner make and model specific measurements were pooled and then subdivided by reconstruction to create scanner specific quantitative profiles.

Results

Different make and model scanners predictably demonstrated different quantitative performance profiles, including in some cases, small calibration bias. Differences in site-specific reconstruction parameters increased the quantitative variability among similar

scanners, with post-reconstruction smoothing filters being the most influential parameter. Quantitative assessment of this intra-scanner variability over this large collection of phantom data gives, for the first time, estimates of reconstruction variance introduced into trials from allowing trial sites to use their preferred reconstruction methodologies. Predictably, Time-of-Flight (TOF) enabled scanners exhibited less size-based partial volume bias than non-TOF scanners.

Conclusions

The CTN scanner validation experience over the past 5 years has generated a rich, well-curated phantom data set from which PET/CT make and model and reconstruction dependent quantitative behaviors were characterized for purposes of understanding and estimating scanner-based variances in clinical trials. These results should make it possible to identify and recommend make and model specific reconstruction strategies to minimize measurement variability in cancer clinical trials.

Keywords: phantom, PET quantitation, scanner calibration, multicenter clinical trials

INTRODUCTION

Multi-center oncologic clinical trials are increasingly using PET/CT imaging as primary and secondary endpoints to define success or failure of treatment regimens with considerable effort expended in understanding reproducibility and variability (*1-11*). PET, as an inherently quantitative imaging technique, is arguably the most powerful imaging modality available to researchers to assess response to therapy in the multi-center clinical trial setting. However, the accurate and reproducible quantitation methodology necessary to successfully complete a trial involving quantitative PET imaging has been complicated by vendors of commercial PET/CT scanner systems that understandably strive to generate higher quality diagnostic images to achieve market differentiation. While these efforts advance the field, they also paradoxically add variability to multi-center trials that include PET/CT equipment whose inherent hardware and software technologies can differ by more than a decade. The introduction of time-of-flight capable scanners and reconstruction advancements including iterative approaches that account for the position sensitive point response function, have further increased both quantitative and qualitative differences between older and newer generation scanners. The divergent image quality and varying quantitation make comparison of quantitative data associated with different makes and models of scanners of different vintages problematic within the context of multi-center clinical trials seeking to use metrics such as Standardized Uptake Values (SUV) and Total Lesion Glycolysis (*1, 12*).

Several professional societies have initiated programs and are devising and promoting standardization practices designed to reduce variability within the context of image quantitation in clinical trials. Organizations such as the American College of Radiology Imaging Network (ACRIN), The Radiological Society of North America's Quantitative Imaging Biomarker Alliance (QIBA), the American Association of Physicists in Medicine (AAPM), the European Association of Nuclear Medicine's Research 4Life (EARL), and the Society of Nuclear Medicine and Molecular Imaging (SNMMI), both alone and together have made significant strides in this area. Several of these organizations administer PET/CT phantom imaging programs to aid in standardization of quantitation in clinical trials and clinical practice (*13-16*). These programs are separate and distinct from clinical accreditations such as those administered by the American College of Radiology and the Intersocietal Accreditation Commission, and the Joint Commission.

In September 2008, the Clinical Trials Network (CTN) was created by SNMMI. The mission of the CTN is to advance the use of molecular imaging radiopharmaceuticals in clinical trials through standardization of chemistry and imaging methodology. This includes using imaging radiopharmaceuticals during the course of drug development, as well as bringing new radiopharmaceuticals to regulatory approval. The CTN operates a phantom-based validation program for PET/CT scanners that utilizes a unique anthropomorphic chest phantom specifically for validating the quantitative performance of PET/CT scanners for use in oncologic clinical trials.

From its inception through January of 2014 the CTN has collected and analyzed over 400 phantom data sets collected from 237 unique PET/CT scanners acquired from a

diverse group of 170 international imaging centers running the gamut from community-based imaging centers to academic sites. Virtually all manufacture make and model scanners from the last decade are represented in the data sets. Specifically excluded from the oncology phantom data are those collected from mobile PET/CT systems and PET-only systems. The image data from scanners that passed the validation criteria in these phantom studies form the basis of analysis presented here.

The study includes PET/CT scanners with technology advancements spanning more than a decade. Reconstruction methods have also evolved substantially during this period. GE and Siemens PET/CT systems have historically utilized similar iterative reconstructions, giving the user a broad level of flexibility determining their own level of convergence by specifying their preferred number of updates (iterations and subsets), and also allowing the ability to apply post-reconstruction Gaussian smoothing filters of user-defined width. Reconstructions of Philips scanners, although also iterative, allow the user less latitude in reconstruction and do not provide the ability to filter the images post-reconstruction.

The overall goal of this analysis is to assess quantitative variability of PET data in the context of single site and multi-center clinical trials that is introduced specifically by variability in scanner calibration and quantitative SUV_{max} measurement of spherical tumor-like lesions in the CTN oncology phantom. By better understanding the magnitude and sources of these variances, the field should be able to devise strategies to predictably enhance the quality of quantitative PET imaging data for clinical trials.

MATERIALS AND METHODS

Phantom Imaging and Data Collection

The CTN oncology clinical simulator phantom is an anthropomorphic chest phantom with lung fields and six spherical objects with inner-diameters ranging from 7-20 mm reproducibly secured at specific locations within the phantom (Fig. 1) (*16, 17*). The six spheres are serially connected via narrow–bore tubing allowing a single syringe to fill all six spheres. The phantom has a single 7 mm diameter sphere located in the mediastinum, two 10 mm spheres placed in the lung fields, and a third 10 mm sphere in an area corresponding to an axillary lymph node, a single 15 mm diameter sphere in the left shoulder, and a single 20 mm diameter sphere in the right lung field. Nominal concentration of the spheres and background at phantom imaging time are 24.0 kBq/mL and 6.0 kBq/mL respectively resulting in a 4:1 lesion: background concentration ratio with scanning commencing precisely 60 minutes after assay of the fill syringes. These concentrations were designed to simulate clinically relevant concentrations and contrasts found in FDG PET oncologic imaging. Phantom imaging is performed for four minutes per bed position for 3D imaging, and six minutes per bed position for 2D imaging. The sites are instructed to use their standard low-dose attenuation correction CT protocol and to reconstruct the images using their standard clinical reconstruction parameter set. However, the sites are instructed *not* to implement point-response-function assisted reconstructions because of variability of reconstructed quantitation using these techniques at this time. A predetermined “patient weight” (63 kg), and “injected dose” (555 MBq) is designed to produce a background SUV of 1.00 if the prescribed fill instructions are followed.

For validation purposes, each site submits to the CTN Phantom Imaging Core the attenuation corrected PET scans, non-attenuation corrected PET scans and the CT

used for attenuation correction. The phantom-fill data (activities and times), as well as PET and CT acquisition and reconstruction parameters and general information regarding the scanner are submitted on paper.

The Scanner Validation Core Lab performs a series of quality control steps prior to final quantitative analysis using Siemens syngo.via (va20), Siemens Inveon Research Workstation (IRW v4.2), and OsiriX (Pixmeo SARL, Switzerland v5.9). The PET/CT datasets were overlaid using the above software to assess the accuracy of the PET/CT registration for the scanner by comparing the 3D position of each of the six spheres on the CT with their location on the PET scan. Misregistrations on the order of 3mm in any dimension are visually detectable. The CT scan is carefully checked for the existence of air bubbles in the spherical lesions, because this will cause anomalously low SUV readings. An incomplete fill results in a request for the site to refill and rescan the phantom.

The sites are also asked to make both an SUV_{max} measurement of all identified lesions as well as a background measurement in the right shoulder region for assessment of scanner calibration accuracy. The CTN Scanner Validation Core lab subsequently makes its own measurements of the SUV_{max} for the spherical lesions, and SUV_{mean} for the background. Core lab measurements are those reported in this manuscript.

Acceptance criteria for the SUV_{mean} of the background region is set at 1.0 ± 0.1. This ±10% permissible variability is consistent with most other organizations that are currently addressing limits for acceptable quantitative PET scanner calibration performance for clinical trials (*2, 13-15, 18*). Because spheres of different sizes are placed within the phantom in different background settings, and scanner specific

performance in this complex environment was originally unknown, rigid sphere-specific acceptance criteria for SUV_{max} for the various sphere sizes are currently not strictly set. The current work presented here will act as the basis for these acceptance criteria moving forward.

Phantom Analysis Approach

For purposes of analysis and data reduction, scanner models from a particular vendor whose PET imaging properties are generally equivalent are bundled together. Fourteen distinct scanner groups were ultimately identified and are listed in Table 1. The proportion of GE, Siemens, and Philips scanners in this sample make up approximately 56%, 34%, and 10% of the scanners respectively.

For this analysis, the phantom data collected was analyzed in two general areas: overall scanner calibration, and scanner- and reconstruction-specific lesion quantitation.

Analysis of the reconstruction parameter sets (iterations, subsets, Gaussian filter width) of the over 240 PET/CT scanners revealed more than 100 different reconstruction parameter sets being used from the imaging sites in the database, demonstrating a substantial lack of standardization. Supplemental Table 1 details the reconstruction parameter sets and the frequency distribution per scanner. The database and data collection was not initially configured to collect Philips-specific parameters, and are therefore not reported in the supplemental table.

Scanner Validation Core lab analysis was performed using Siemens syngo.via workstations, Siemens Inveon Research Workstation, and OsiriX. All workstations were verified to generate the same SUVmax values generally to within 2% of one another,

however not all workstations were capable of generating SUV measurements from all scanner system image sets. OsiriX proved most universally capable of quantitation of concentration and SUV values and was used in those cases where the other workstations failed to generate quantitative information.

Scanner Calibration Analysis

For scanner calibration assessment, an approximately 30 mm diameter spherical volume of interest is created in the right shoulder, which is a uniform region devoid of complicating structures and concentrations. The region is placed far from the edges of the phantom to avoid partial volume effects. The mean and standard deviation of the VOI is recorded. The calibration data from similar models as described in Table 1 was pooled to assess scanner model specific trends. Two-sided t-test analysis was performed to determine whether the individual scanner specific background distributions were statistically significantly different from the parent background distribution of all scanners combined. An additional spherical VOI was placed in the uniform region located caudally in the phantom in the area near where the myocardium would be anatomically located (the “myocardial background region”). The difference between the right shoulder background SUV_{mean} and the background myocardial SUV_{mean} was calculated for all scanner studies. Results were compiled for each make and model scanner to determine whether scanner-specific quantitative anatomic biases exist.

Reconstruction Specific Quantitation

For the scanner and reconstruction specific lesion quantitation analysis, spherical VOIs with diameters at least two-times the diameter of the actual spheres were drawn over all

six spherical objects. CT information was used when the precise location of the lesion was not apparent on the PET scan. SUV_{max} measurements were made for each of the lesions. Both the imaging site and the Scanner Validation Core lab made this measurement. The Core measurements are those presented. For purposes of this analysis, only the SUV_{max} measurements from the five spheres 10 mm and larger are reported. They are first combined by scanner model, and then subsequently subcategorized by reconstruction. Measurements of the 7 mm sphere were specifically excluded from this analysis because so few scanners were able to detect it. Sub-categorization was performed by the width of the Gaussian reconstruction filter used, as this was determined to have the most significant quantitative impact. To achieve meaningful statistical numbers of phantom scans, Gaussian filter width ranges were typically used, rather than a specific filter width. Since Philips scanner reconstructions do not provide the ability to choose a post-reconstruction filter, Philips phantom data was analyzed per scanner, but not subsequently subcategorized.

RESULTS

Scanner Calibration

Assessment of accuracy of scanner calibration was performed on all submitted phantom studies by creating a spherical VOI in the uniform region of the left shoulder as described above. The SUV_{mean} was calculated for each attenuation corrected phantom study and the results were tabulated into frequency histograms for all 14 scanner models. Representative SUV_{mean} histogram distributions for background measurements (Nominally = 1.00) for two PET/CT scanner models are presented in Figures 2 A-B. Mean and standard deviations calculated for each of the 14 model scanners are also

shown in Figure 2C.

All pooled model-specific mean background values (Fig. 2C) are within ±4% of the actual concentration. However the GE Discovery 690-710 scanners and the Biograph 2/6 scanners both demonstrated a statistically significant positive bias when compared to the parent background SUV distribution. Four other scanner models (annotated in Fig. 2C) had p-values between 0.05 and 0.1, suggesting the possibility of slight bias.

Scanner specific differences between shoulder background SUV_{mean} and the background myocardial SUV_{mean} are listed (Table 2***).*** In nearly half of the 14 scanner models investigated there was a clear reconstruction-driven bias between the measurements in the shoulder region and the myocardial region. Investigating the GE line of PET/CT scanners gives insight into these phenomena. In 10 of 11 phantom scans with the GE 600 PET/CT scanner the myocardial background region concentration measurement was greater than that in the shoulder region. However, with the GE 690/710 scanners the opposite was found with 31 of 33 scans having the shoulder region greater than the myocardial region. GE's older models (the ST and STE) demonstrated no such bias.

Lesion Quantitation

Although updates (defined as iterations x subsets) impact quantitation, categorizing individual scanner data by the post-reconstruction Gaussian filter width demonstrated the most significant and systematic quantitative impact and is the basis of the data and analysis presented. The reconstructions for each of the PET/CT scanner models (Table 1) was sorted and pooled by Gaussian filter width. The complete set of data for the 14

scanner models is presented in Table 3. Representative results of the SUV_{max} values for each of the five spheres 10 mm and larger for the GE Discovery STE, GE Discovery 690-710, Siemens Biograph TruePoint, and Philips TF are graphically presented in Figure 3. All results for all individual scanner models are presented in histogram plots in Supplemental Figures 1-3. In each of these histogram plots the leftmost bar is the mean SUV_{max} for that sphere for the entire 406 phantom datasets. Subsequent bars represent mean SUV_{max} for increasing Gaussian filter width ranges used in reconstructions for that model scanner. Three filter bin widths were typically selected for each of the scanner models primarily to balance, to the extent possible, the number of phantom scans in each bin. However, balanced distribution was often not possible. Philips, as previously mentioned, does not allow the user the capability to filter the image post-reconstruction. Given the limited number of scanners per model in our sample, refining filter bin widths beyond three bins would have resulted in too little data per bin for conclusions to be drawn.

Differences in general quantitative performance between vendors was not observed, however the vintage of scanner models did appear to impact the range and distribution of measured SUV_{max} values for the spheres. For purposes of this analysis, early generation PET/CT scanners (Discovery LS, Biograph Duo and Biograph 6, and Philips Gemini and Gemini GS) were bundled into a one category, recent higher-performance time-of-flight scanners (GE 690/710, Siemens mCT, and Philips Ingenuity) were put into a second category, while the remaining PET/CT scanners were segregated into a third mid-range performance category. Examples of the different SUV_{max} distributions for these three categories for the 15 mm left shoulder sphere, and the 10 mm right lung

sphere are shown in Figure 4 A and B. It should be noted that virtually all of the anomalously high-SUV_{max} values in the plots in the high-performance time-of-flight scanner distribution are associated with point-response-function reconstructions that were inadvertently submitted to CTN, (CTN specifically excludes point-response-function reconstructions from their official analyses). The inclusion of these data in these plots is to demonstrate the broad and largely unpredictable quantitative behavior of these reconstructions with current implementations.

DISCUSSION

Multi-center clinical trials typically, and sometimes necessarily, recruit a cross-section of medical centers that range from community-based clinics to world-class academic centers. Imaging sites at these institutions employ a range of scanners of different make and model, and the trial protocol generally asks the sites to image their study subjects using their standard clinical acquisition and reconstruction. The impact of this uncontrolled approach to imaging on any quantitative endpoint within the context of a multi-center clinical trial is largely unknown. However, it is clear that any additional variance that results from quantitative variability across imaging equipment and technique will detrimentally impact the statistical power of the study, and require more subjects at significantly greater expense.

The collection of over 400 CTN oncology phantom data sets is a rich and diverse set of qualitative and quantitative information on scanner performance across site-type, scanner make and model, and vintage. The data presented provides the first large-scale controlled systematic analysis of the impact of scanner and reconstruction specific quantitative performance.

Perhaps the most surprising result of the phantom dataset is the diversity of reconstruction parameter sets even when limited to a single scanner model. Each scanner site typically begins with a default reconstruction parameter set, but then experiments with different parameter sets to achieve a clinical image quality with which the particular site physician(s) are comfortable. Vendors understandably are providing both the means and the opportunity for each site to optimize reconstructions to their own preferences. However, this creates an environment where quantitative variability will be inevitable in any multi-center trial.

Scanner Calibration

By convention, all PET scanners are calibrated with a 20 cm diameter cylindrical phantom with known concentration. The accuracy of this calibration is tied to the accuracy of the dose calibrator, timing, and volume measurements associated with the calibration procedure. A properly calibrated scanner will demonstrate accurate concentration measurements in the cylindrical phantom across the entire axial field of view, which is precisely what the ACRIN phantom procedure measures and verifies.

The CTN oncology phantom is neither designed to nor capable of confirming full axial FOV calibration. Since the VOI for background measurement in the anthropomorphic chest phantom is in the right shoulder, far from the center of the scanner field of view, and because of phantom asymmetry, there is possibility for calibration measurement bias as compared with that obtained from a standard 20 cm diameter cylindrical phantom. The background SUV distributions for each of the three time-of-flight systems from the three vendors each demonstrated a non-statistically significant, but suggestive, calibration bias as measured in the shoulder area of the phantom. These biases, if real,

may result from scatter corrections tuned to standard simple geometries that may be rendered inaccurate under more complex situations.

The hypothesis that the complexity of the phantom presents a more significant quantitative challenge is supported by additional background measurements that were made in the uniform myocardial region of the phantom. Specific scanner models frequently showed significant differences between the shoulder background and myocardial background measurements. These differences are not evident in the more common ACRIN-style cylindrical phantom test of scanner uniformity. ACRIN's own observation of differences in mean liver SUV between vendors supports the existence of this problem.[13]

Current scatter correction assessments, like in NEMA measurements, or with the NEMA image quality phantom, are made closer to the center of the scanner field of view and have a uniform concentration and density. The CTN Oncology Phantom is complex in design and geometry with multiple density internal objects and therefore presents a different and more challenging imaging scenario.

Benchmarking

One of the primary uses of the current CTN oncology phantom image and reconstruction database is benchmarking. An individual scanner can be quantitatively benchmarked against itself, based on prescribed periodic phantom imaging during the course of a clinical trial to determine long-term quantitative stability and variance. Additionally, a particular scanner's performance can be benchmarked both against identical scanners that use different reconstructions and also identical scanners with

virtually identical reconstructions. In either case, an individual phantom scan result, when compared to the compiled and categorized data, can inform the site and trial sponsor of a scanner's performance relative to relevant statistical parent distributions.

With this data, it is also possible for a trial sponsor to estimate an anticipated variance of quantitative data based upon the mix of make and model scanners used in a multi-center trial (with associated reconstructions) using the compiled SUV_{max} database for the phantom.

For trial sponsors interested in more prospectively harmonized quantitative data, the database can help sponsors identify make and model specific candidate reconstructions that might help reduce variances prospectively. Because current TOF enabled scanners demonstrated significantly higher quantitative performance (higher SUV_{max} values) than those without TOF capabilities (Figure 4A-B), a sponsor might consider requiring TOF scanners to reconstruct without the TOF information in order to reduce differences between scanners. Alternatively, excluding earlier vintage scanners from multi-center clinical trials may be a reasonable strategy for trials where absolute quantitative measurements are critical.

Quantitative scanner performance as defined by SUV_{max} of the spheres in the CTN phantom demonstrated significant variability. This was not unexpected given the broad range of scanner vintages and the diversity of reconstructions. Categorizing SUV_{max} results by scanner and subcategorizing by post-reconstruction Gaussian filter width demonstrated expected reduction of SUV_{max} with increasing filter width for all spheres and all scanner make and models. Within a given model, this decrease in SUV_{max} occurred at a rate of approximately 0.2-0.3 SUV units per additional mm of filter width.

CONCLUSION

The current assembly of over 400 CTN oncology phantom scans includes multiple image sets from virtually all make and model PET/CT scanners. The CTN oncology phantom demonstrated utility in both validating scanner calibration and characterizing the reconstruction-specific quantitative imaging characteristics of 14 different make and model PET/CT scanners through the measurement of SUV_{max} values for the phantom's 5 spherical objects (10-20 mm). Analysis of the variability in the reported phantom lesion measurements should enable sponsors and designers of clinical trials to better estimate quantitative variance within a multi-center clinical trial setting. The reconstruction specific data should also be useful to help trial designers minimize variance by selecting scanner specific reconstructions towards quantitative harmonization.

Acknowledgements: We kindly thank Tina Kiss, Jina Kim, and Bonnie Clarke of the SNMMI for their hard work in administering the CTN scanner validation program and its associated database. Further thanks to Keith Bingham, Medical Designs, Inc, Newtown, CT for his work in designing, manufacturing and maintaining the phantoms. Lastly thanks to the hard-working members of the CTN scanner validation committee for shepherding the program.

References

1. Boellaard R. Need for standardization of 18F-FDG PET/CT for treatment response assessments. *J Nucl Med.* 2011;52 Suppl 2:93S-100S.
2. Boellaard R, O'Doherty MJ, Weber WA, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging.* 2010;37:181-200.
3. Doot RK, Pierce LA, 2nd, Byrd D, Elston B, Allberg KC, Kinahan PE. Biases in multicenter longitudinal PET standardized uptake value measurements. *Transl Oncol.* 2014;7:48-54.
4. Feuardent J, Soret M, de Dreuille O, Foehrenbach H, Buvat L. Reliability of uptake estimates in FDG PET as a function of acquisition and processing protocols using the CPET. *IEEE T Nucl Sci.* 2005;52:1447-1452.
5. Kurland BF, Gerstner ER, Mountz JM, et al. Promise and pitfalls of quantitative imaging in oncology clinical trials. *J Magn Reson Imaging.* 2012;30:1301-1312.
6. Lammertsma AA. Measurement of tumor response using [18F]-2-fluoro-2-deoxy-D-glucose and positron-emission tomography. *J Clin Pharmacol.* 2001;Suppl:104S-106S.
7. Lammertsma AA, Hoekstra CJ, Giaccone G, Hoekstra OS. How should we analyse FDG PET studies for monitoring tumour response? *Eur J Nucl Med Mol Imaging.* 2006;33 Suppl 1:16-21.
8. Quak E, Hovhannisyan N, Lasnon C, et al. The importance of harmonizing interim positron emission tomography in non-Hodgkin lymphoma: focus on the Deauville criteria. *Haematologica.* 2014;99:e84-85.

9. Shankar LK, Hoffman JM, Bacharach S, et al. Consensus recommendations for the use of F-18-FDG PET as an indicator of therapeutic response in patients in national cancer institute trials. *J Nucl Med.* 2006;47:1059-1066.

10. Vriens D, Visser EP, de Geus-Oei LF, Oyen WJ. Methodological considerations in quantification of oncological FDG PET studies. *Eur J Nucl Med Mol Imaging.* 2010;37:1408-1425.

11. Westerterp M, Pruim J, Oyen W, et al. Quantification of FDG PET studies using standardised uptake values in multi-centre trials: effects of image reconstruction, resolution and ROI definition parameters. *Eur J Nucl Med Mol Imaging.* 2007;34:392-404.

12. Lasnon C, Desmonts C, Quak E, et al. Harmonizing SUVs in multicentre trials when using different generation PET systems: prospective validation in non-small cell lung cancer patients. *Eur J Nucl Med Mol Imaging.* 2013;40:985-996.

13. 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.

14. QIBA-UPICT Protocol, Version for Public Comment. FDG-PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy, v1.0: Radiological Society of North America; 2013.

15. Boellaard R, Willemesen AT, Arends B, Visser E. EARL procedure for assessing PET/CT system specific patient FDG activity preparations for quantitative FDG PET/CT studies. Guidelines. 2011. Available from EARL.EANM.ORG.

16. Christian PE. Use of a precision fillable clinical simulator phantom for PET/CT scanner validation in multi-center clinical trials: The SNM Clinical Trials Network (CTN) Program [abstract]. *J Nucl Med.* 2012;53:437.

17. Christian PE. Longitudinal PET scanner stability: SNMMI Clinical Trials Network experience. *J Nucl Med. 2014;* 55:2156.

18. FDG-PET/CT Technical Committee. FDG-PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy, Quantitative Imaging Biomarkers Alliance, Version 1.05 Publically Reviewed Version. QIBA, December 11, 2013. Available from:RSNA.ORG/QIBA.

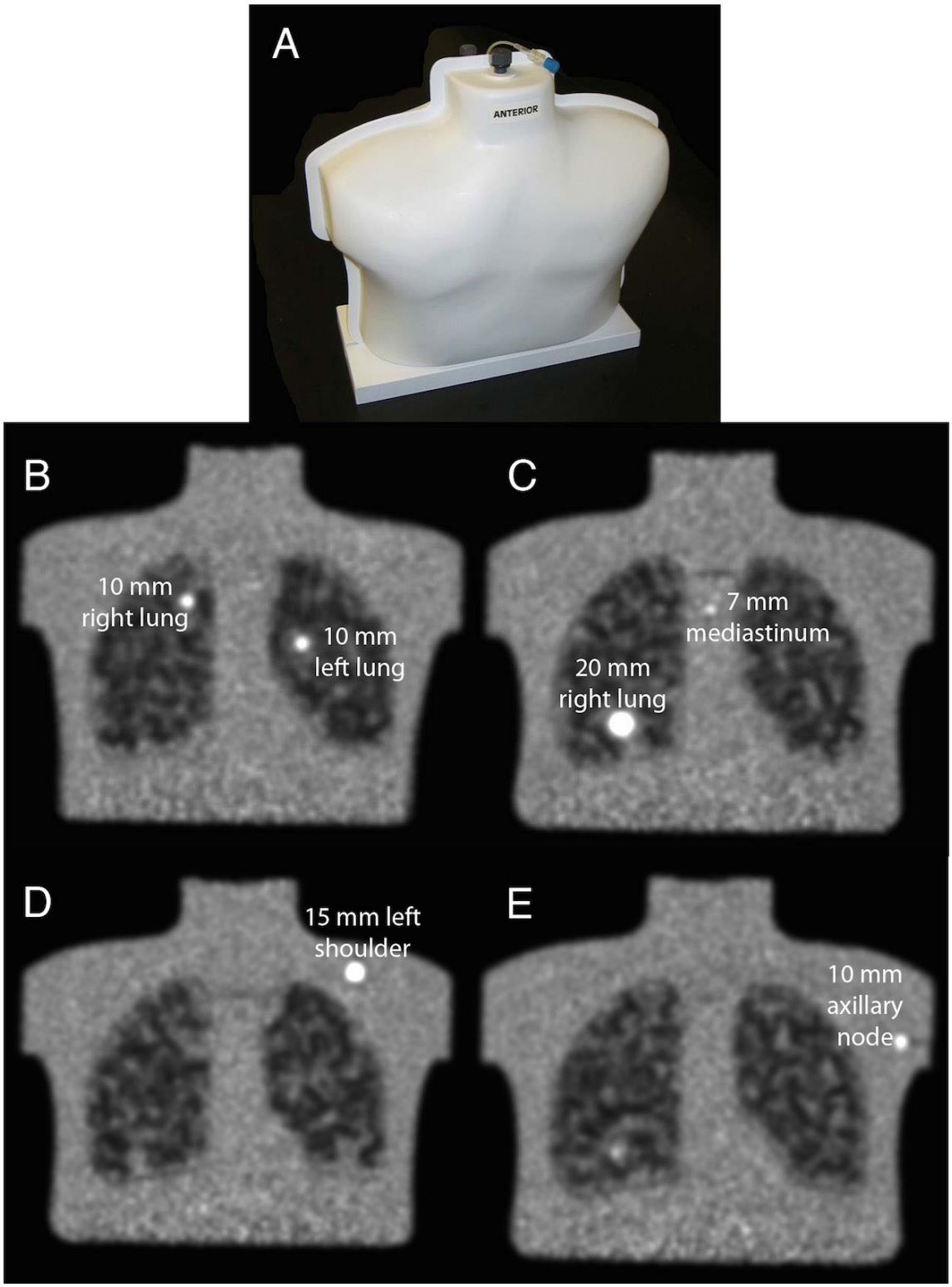
A
ANTERIOR
B
10 mm
right lung
10 mm
left lung
C
7 mm
mediastinum
20 mm
right lung
D
15 mm left
shoulder
E
10 mm
axillary
node

FIGURE 1. Representative phantom images from a later model time-of-flight enabled PET/CT scanner capable of visualizing of all six spheres. A) The CTN Oncology Phantom. B) Coronal slice visualizing both the left and right 10 mm lung lesions. C) Coronal slice visualizing the 7 mm mediastinal lesion and 20 mm right lung sphere D) Coronal slice visualizing the 15 mm sphere in the left shoulder. E) Coronal slice visualizing the 10 mm axillary lymph node.

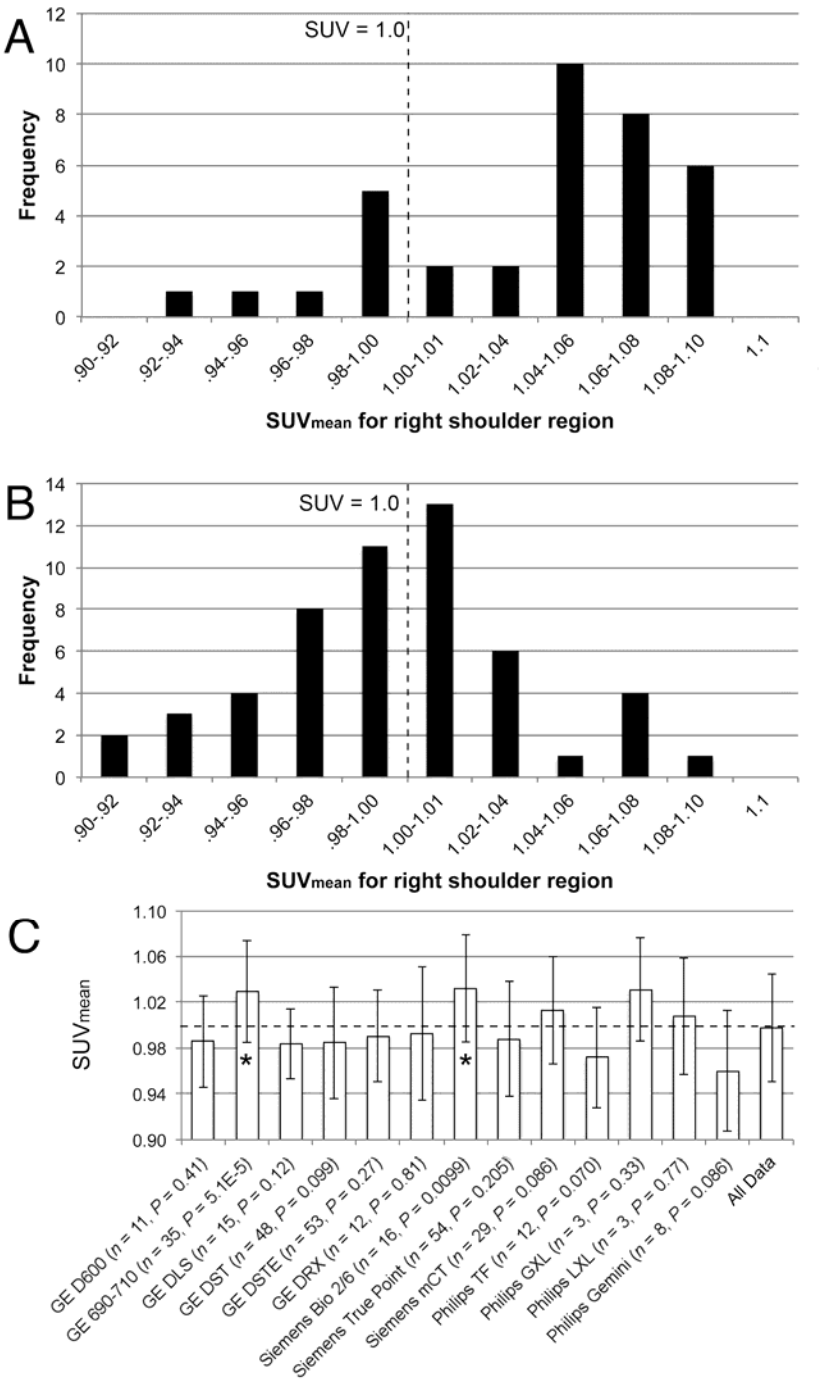

FIGURE 2. Representative background SUV_{mean} measurements in the right shoulder region. A) Asymmetrically distributed histogram distribution of background measurements for the GE 690-710 PET/CT scanner models. B) Generally symmetric

histogram distribution for the GE Discovery STE PET/CT scanner platform centered around SUV=1.0. C) Mean of all the background SUV_{mean} compiled for each scanner make and model. The GE 690/710 models, and the Siemens Biograph 2/6 both had means statistically significantly higher than 1.0 as designated by the *.

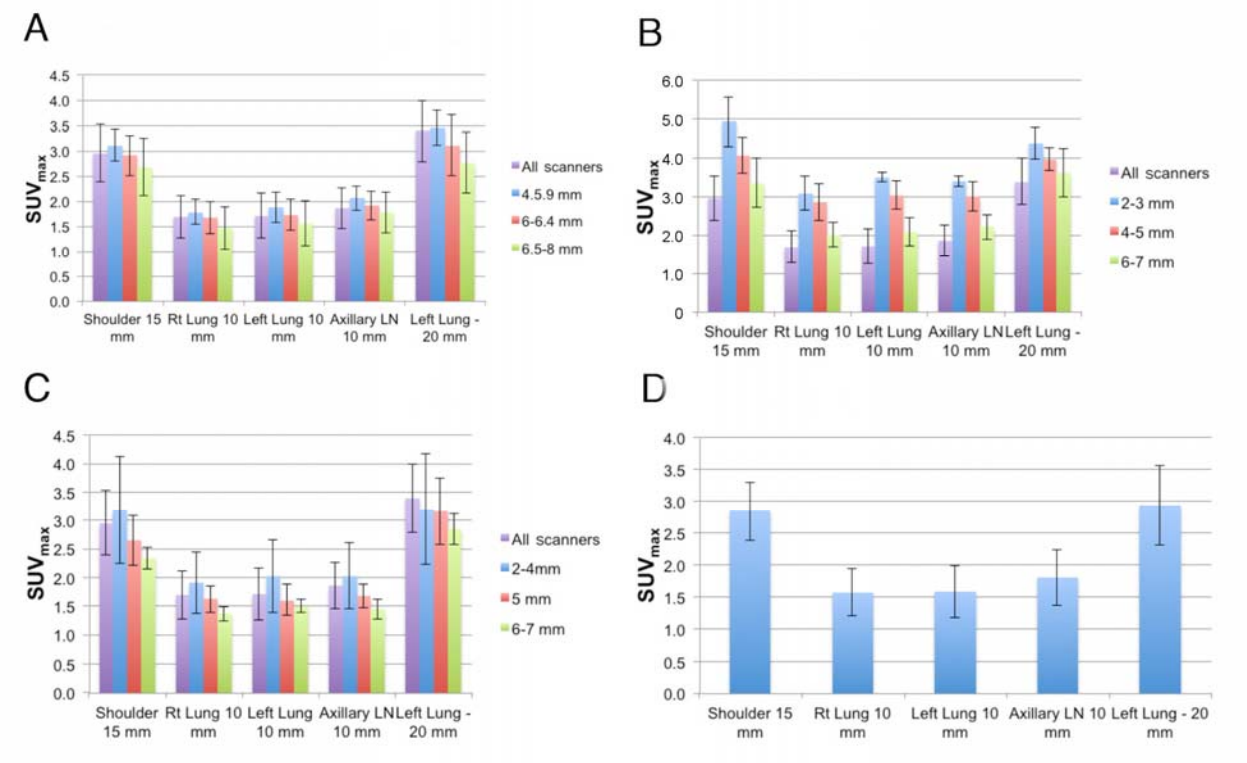

FIGURE 3. Representative SUV_{max} histograms of the five spherical lesions in the CTN oncology phantom ≥ 10mm for four different PET/CT scanner make and models. In A-C the first bar in each histogram grouping is the mean value for that lesion in all phantom studies from all scanners. Subsequent histogram bars are the averages for the specified reconstruction filter width bins. A) GE Discovery STE. B) GE Discovery 690-710. C) Siemens Biograph TruePoint. D) Philips Gemini TF. The Gemini TF shows only a single bar because reconstructions were not broken down for Philips scanners because they do not allow the user to apply a reconstruction filter

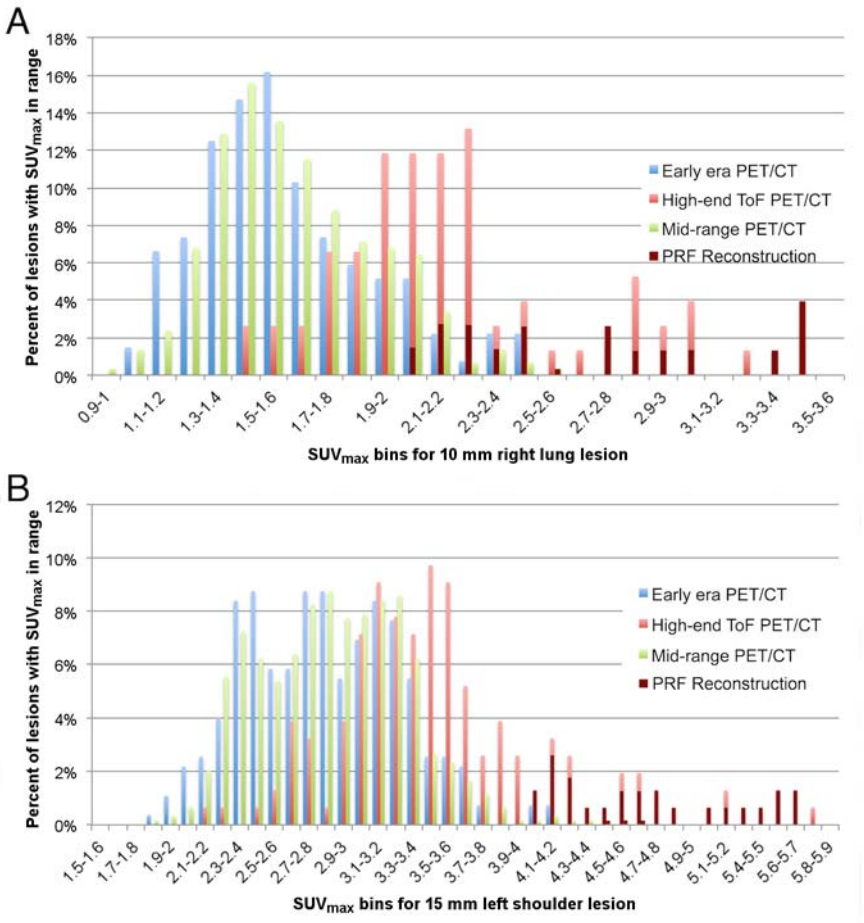

FIGURE 4. A) Histogram distribution of SUV_{max} values for the 10 mm right lung lesion of the CTN oncology phantom for three different vintage/performance PET/CT scanner models. B) Similar SUV_{max} histogram distribution for the 15 mm left shoulder spherical lesion. More recent model time of flight enabled scanners demonstrated higher SUV_{max} values, in general, than non-time of flight machines. Point response function reconstructions (PRF) primarily but not exclusively from some

time-of-flight (ToF) enabled machines are designated by the maroon bars in both A and B.

TABLE 1. Categorization of Scanners into groups of like quantitative performance.

	Scanner Model	Scanner Grouping	Number of unique scanners	Number of phantom scans
GE PET/CT Scanner Models	STE	STE	25	47
	VCT		17	29
	LS	LS	16	23
	ST	ST	34	59
	RX	RX	7	16
	600	600-610	6	14
	610		0	0
	690	690-710	18	31
	710		4	6
	TOTAL GE		127	225
Siemens PET/CT Scanner Models	Biograph TruePoint	Biograph TruePoint	43	83
	Biograph Duo	Biograph 2-6	7	12
	Biograph 6		6	8
	Biograph mCT	mCT	23	36
	TOTAL Siemens		79	139
Philips PET/CT Scanner Models	Gemini TF	Gemini TF	16	18
	Ingenuity	Ingenuity	1	1
	Gemini LXL	Gemini LXL	1	3
	Gemini GS2	Gemini GS2	6	10
	Gemini GXL	Gemini GXL	7	10
	TOTAL Philips		31	42
TOTAL	TOTAL ALL VENDORS.		237	406

TABLE 2. Differences in background SUV_{mean} measurements for uniform areas in phantom

	GE Discovery 600	GE Discovery 690-710	GE Discovery LS	GE Discovery RX	GE Discovery ST	GE Discovery STE
Number of Phantom Scans with Shoulder SUV_{mean} > Myocardial SUV_{mean}	1	31	11	1	27	33
Number of Phantom Scans with Shoulder SUV_{mean} < Myocardial SUV_{mean}	10	2	7	12	27	29
Average SUV Difference	-0.03	0.05	0.00	-0.03	0.00	-0.01

	Siemens Biograph 2/6	Siemens Biograph TruePoint	Siemens Biograph mCT	Philips Gemini TF	Philips Gemini GXL	Philips Gemini LXL	Philips Gemini GS
Number of Phantom Scans with Shoulder SUV_{mean} > Myocardial SUV_{mean}	9	51	25	10	2	3	7
Number of Phantom Scans with Shoulder SUV_{mean} < Myocardial SUV_{mean}	10	15	6	2	0	0	0
Average SUV Difference	0.01	0.02	0.02	0.05	0.09	0.08	0.02

TABLE 3. SUV_{max} measurements for the five spherical lesions ≥10 mm in the CTN Oncology phantom

	Filter Width (mm)	n	Left Shoulder (15 mm)	Rt Lung (10 mm)	Left Lung (10 mm)	Axillary LN (10 mm)	Left Lung (20 mm)
GE Discovery 600	6.0	5	3.09 ± 0.51	1.76 ± 0.27	2.09 ± 0.52	2.18 ± 0.36	3.30 ± 0.42
	6.1-7.0	7	2.91 ± 0.33	1.73 ± 0.13	1.79 ± 0.17	1.87 ± 0.12	3.35 ± 0.37
	7.1-9.0	2	2.41 ± 0.04	1.41 ± 0.10	1.35 ± 0.03	1.64 ± 0.06	2.89*
GE Discovery 690-710	2.0-3.9	2	4.94 ± 0.62	3.09 ± 0.43	3.51 ± 0.12	3.40 ± 0.13	4.38 ± 0.40
	4.0-5.9	8	4.07 ± 0.45	2.86 ± 0.47	3.04 ± 0.35	3.01 ± 0.37	3.96 ± 0.28
	6.0-7.0	27	3.35 ± 0.62	2.02 ± 0.31	2.08 ± 0.36	2.22 ± 0.31	3.61 ± 0.62
GE Discovery LS	5.0-5.4	3	3.06 ± 0.51	1.66 ± 0.28	1.75 ± 0.44	1.90 ± 0.25	3.21 ± 0.66
	6.0	12	2.86 ± 0.22	1.55 ± 0.14	1.57 ± 0.44	1.74 ± 0.27	3.54 ± 0.39
	7.0-10.0	6	2.14 ± 0.15	1.23 ± 0.17	1.20 ± 0.10	1.30 ± 0.10	2.68 ± 0.02
GE Discovery RX	3.0	3	3.39 ± 0.25	2.00 ± 0.13	2.18 ± 0.18	2.45 ± 0.09	3.02 ± 0.37
	4.0-5.9	11	2.89 ± 0.38	1.73 ± 0.28	1.79 ± 0.20	1.98 ± 0.22	3.19 ± 0.59
	6.0-7.0	3	2.74 ± 0.48	1.69 ± 0.37	1.60 ± 0.25	1.58 ± 0.48	3.25 ± 0.35
GE Discovery ST	4.0-5.9	16	2.98 ± 0.27	1.83 ± 0.32	1.81 ± 0.23	1.92 ± 0.23	3.43 ± 0.46
	6.0-6.4	32	2.83 ± 0.43	1.60 ± 0.26	1.69 ± 0.30	1.81 ± 0.25	3.13 ± 0.61
	6.5-8.0	8	2.58 ± 0.31	1.46 ± 0.36	1.50 ± 0.36	1.59 ± 0.25	2.98 ± 0.53
GE Discovery STE	4.0-5.9	21	3.11 ± 0.30	1.78 ± 0.24	1.87 ± 0.29	2.06 ± 0.23	3.45 ± 0.33
	6.0-6.4	33	2.90 ± 0.38	1.67 ± 0.31	1.72 ± 0.30	1.91 ± 0.28	3.11 ± 0.60
	6.5-8.0	18	2.66 ± 0.31	1.46 ± 0.18	1.55 ± 0.20	1.78 ± 0.17	2.76 ± 0.50
Siemens Biograph 2-6	5.0	17	2.47 ± 0.38	1.34 ± 0.20	1.37 ± 0.24	1.58 ± 0.18	2.93 ± 0.48
	6.0	3	2.34 ± 0.37	1.56 ± 0.16	1.55 ± 0.22	1.62 ± 0.16	2.64 ± 0.61
Siemens Biograph TruePoint	2.0-4.0	18	3.17 ± 0.93	1.90 ± 0.52	2.02 ± 0.63	2.02 ± 0.57	3.19 ± 0.97
	5.0	52	2.65 ± 0.43	1.62 ± 0.22	1.60 ± 0.26	1.67 ± 0.20	3.16 ± 0.57
	6.0-7.0	11	2.33 ± 0.18	1.36 ± 0.12	1.50 ± 0.11	1.44 ± 0.16	2.84 ± 0.26
Siemens Biograph mCT	1.0-3.0	11	3.82 ± 0.82	2.48 ± 0.43	2.38 ± 0.41	2.51 ± 0.45	3.85 ± 0.42
	4.0	9	3.23 ± 0.37	2.21 ± 0.38	2.14 ± 0.35	2.16 ± 0.25	3.04 ± 0.73

	5.0	15	3.18 ± 0.39	2.02 ± 0.24	2.01 ± 0.24	2.03 ± 0.26	3.15 ± 0.97
Philips Gemini TF	N/A	18	2.84 ± 0.45	1.56 ± 0.36	1.58 ± 0.40	1.80 ± 0.43	2.94 ± 0.62
Philips Gemini GXL	N/A	10	2.89 ± 0.36	1.38 ± 0.18	1.44 ± 0.19	1.83 ± 0.18	3.06 ± 0.55
Philips Gemini LXL	N/A	3	3.47 ± 0.24	1.48 ± 0.07	1.50 ± 0.11	1.97 ± 0.25	3.61 ± 0.07
Philips Gemini GS	N/A	10	2.58 ± 0.23	1.35 ± 0.14	1.36 ± 0.20	1.57 ± 0.14	3.29 ± 0.24

* Only a single scanner used a post-reconstruction filter width in this range, making calculation of standard deviation impossible.