Nuclear Medicine Communications

Harmonization based on quantitative analysis of standardized uptake value variations across PET/CT scanners: a multicenter phantom study

Abbas Monsef[a,b], Mohammad Reza Ay[a,b], Peyman Sheikhzadeh[a,c], Parham Geramifar[d], Arman Rahmim[e,f] and Pardis Ghafarian[g,h]

Objectives This study aimed to measure standardized uptake value (SUV) variations across different PET/computed tomography (CT) scanners to harmonize quantification across systems.

Methods We acquired images using the National Electrical Manufacturers Association International Electrotechnical Commission phantom from three PET/CT scanners operated using routine imaging protocols at each site. The SUVs of lesions were assessed in the presence of reference values by a digital reference object (DRO) and recommendations by the European Association of Nuclear Medicine (EANM/EARL) to measure inter-site variations. For harmonization, Gaussian filters with tuned full width at half maximum (FWHM) values were applied to images to minimize differences in SUVs between reference and images. Inter-site variation of SUVs was evaluated in both pre- and postharmonization situations. Test-retest analysis was also carried out to evaluate repeatability.

Results SUVs from different scanners became significantly more consistent, and inter-site differences decreased for SUV_{mean}, SUV_{max} and SUV_{peak} from 17.3, 20.7, and 15.5% to 4.8, 4.7, and 2.7%, respectively, by harmonization (*P* values <0.05 for all). The values for contrast-to-noise ratio in the smallest lesion of the phantom verified preservation of image quality following harmonization (>2.8%).

Conclusions Harmonization significantly lowered variations in SUV measurements across different PET/CT scanners, improving reproducibility while preserving image quality. *Nucl Med Commun* XXX: 000–000 Copyright © 2022 Wolters Kluwer Health, Inc. All rights reserved.

Nuclear Medicine Communications XXX, **XXX**:000–000

Keywords: PET/CT, SUV, harmonization, multi-center study, repeatability

[a]Department of Medical Physics and Biomedical Engineering, [b]Research Center for Molecular and Cellular Imaging, [c]Department of Nuclear Medicine, Imam Khomeini Hospital, [d]Research Center for Nuclear Medicine, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran, [e]Departments of Radiology and Physics, University of British Columbia, [f]Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC, Canada, [g]Chronic Respiratory Diseases Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran and [h]PET/CT and Cyclotron Center, Masih Daneshvari Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Correspondence to Pardis Ghafarian, PhD, Chronic Respiratory Diseases Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran
Tel: +9821 2712 2705; fax: +9821 2610 9504;
e-mail: pardis.ghafarian@sbmu.ac.ir

Received 4 February 2022 **Accepted** 31 May 2022

key point

- This study measured net variations in standardized uptake value (SUV) across different PET/CT scanners.
- Harmonization significantly improves reproducibility of quantitative metrics while preserves image quality acceptably.
- Current reference ranges for the SUVs need to be updated compatible with modern PET/CT scanners.

Introduction

PET combined with X-ray computed tomography (PET/CT) imaging is nowadays a standard component of cancer care management [1,2]. Specifically, PET/CT using the radiotracer ^{18}F-2-fluoro-2-deoxy-D-glucose (FDG) plays an essential role in diagnosis, staging and responses to therapy assessment for various malignancies in oncology [3,4].

Although most clinical PET/CT interpretations are based on visual assessment of FDG accumulation, quantification of metabolic information has significant potential [5]. The most common indicator used to quantify radiotracer accumulation in images is the standardized uptake value (SUV), whose change can be a more effective biomarker relative to anatomic size information in monitoring disease progression [6].

At the same time, the SUV is associated with multiple sources of variation, including data acquisition, reconstruction parameters, scanner calibration, patient physiological factors, and so on [7,8]. This can cause measurement variability among different systems, even within the same scanner using different protocols. Although guidelines recommend pre- and post-treatment PET scans to be ideally performed with an identical scanner, this is not always practical. Hence, the resulting uncertainty can be misleading to discern real percentage changes in tumor SUV and measurement fluctuations [9–11]. This becomes even more essential when low uptake malignancies (low SUVs) are considered.

DOI: 10.1097/MNM.0000000000001598

A systematic solution may thus be needed to reduce inter-site and intra-site SUV variabilities. The European Association of Nuclear Medicine (EANM) has identified this problem since 2010. Harmonizing different PET/CT sites increases consistency in multicenter clinical trials, for example, quantitative assessment of clinical images for monitoring treatment response and disease progression in oncology. This is a very important consideration, to significantly improve the power of clinical trials [12]. Overall, harmonization enables improved consistency, while aiming to preserve high PET image qualities for diagnostic and follow-up purposes [13].

Several studies have aimed to address factors affecting SUV measurement errors and assess SUV reproducibility and repeatability, which are related to inter- and intra-scanner variabilities, respectively [14]. The quantitative imaging biomarker alliance (QIBA) has provided definitions for reproducibility and repeatability [15]. Many oncology follow-up studies evaluate SUV change based on criteria characterized by the European Organization for Research and Treatment of Cancer (EORTC) [16] and PET Response Criteria in Solid Tumors [5]. A number of surveys have adopted SUV harmonization strategies by following the EANM/EARL guidelines [10]. Tsutsui *et al*. [17] used the Japanese Harmonization technology (J-hart) and performed a multicenter analysis of PET SUV according to the Japanese Society of Nuclear Medicine (JSNM) standards. These studies often work with a specified range of recovery coefficients as a function of sphere sizes so as to align the image SUVs within the standard allowed range.

In the present study, we propose using a digital reference object (DRO) for the purpose of harmonizing SUVs among our PET/CT centers. The objective was to measure differences in SUVs acquired across scanners pre- and post-harmonization and to decrease inter-site variations to improve the reproducibility of the SUV. Our study is distinct from previous works in its special focus to evaluate SUV variability in low uptake lesions. We aim to enable reliably comparable inter-site quantification and to facilitate multicenter PET/CT studies.

Materials and methods

Figure 1 illustrates the overall schematic of our study.

Phantom preparation and data acquisition

We used a National Electrical Manufacturers Association (NEMA) International Electrotechnical Commission body phantom for our experiments based on the NEMA NU-2 2012 standard [18]. The phantom consists of six spheres 10, 13, 17, 22, 28 and 37 mm in diameter with a wall thickness of 1 mm, as a benchmark to evaluate various sizes of clinical lesions. To study low uptake lesions, we focused on low SUV values. Hence, we designed our phantom study having two values for lesion-to-background ratio (LBR), namely 2.0 and 4.0. The background radioactivity concentration was prepared at 3.7 kBq/ml of ^{18}F-FDG solution (assuming standard 185 MBq per 50 kg). For LBR values of 2.0 and 4.0, we filled the spheres with 7.4 and 14.8 kBq/ml of ^{18}F-FDG solution, respectively, on separate days at each center. We repeated the entire process at all centers identically. Every prepared phantom was scanned by each system using routine protocols for data acquisition and image reconstruction (Table 1). To evaluate intra-scanner variability, we repeated each scan under similar conditions, 110 min apart, for test-retest study and analysis.

PET/CT scanners

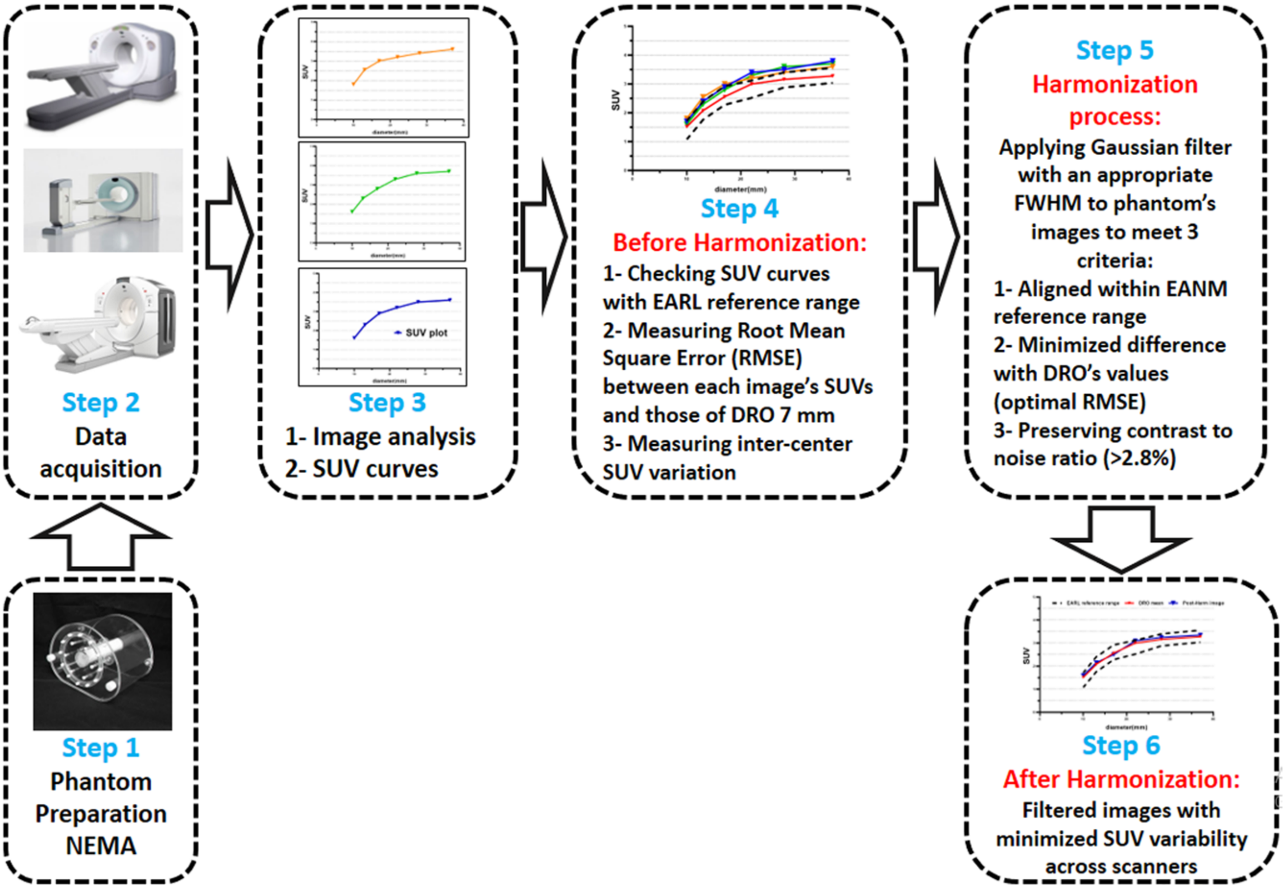
We acquired data from three PET/CT systems (Table 1), namely GE Discovery 690 (GE Healthcare, Wisconsin, USA), Siemens Biograph (Siemens Healthcare), and GE Discovery IQ (GE Healthcare), at three collaborating PET/CT centers within three university hospitals contributing to cancer care and research. Dose calibrator cross-calibration before phantom scanning was carried out by each center's medical physicist.

Digital reference object

In addition to EANM/EARL reference bandwidths (acceptable ranges), we used an FDG PET/CT DRO that has been mathematically developed by QIBA [19], to generate reference SUVs to assess quantitative performances for the different sites (the objective behind the use of the DRO will become more clear in subsection *Harmonization process*). This is a digital image created based on NEMA body phantom characteristics in the Digital Imaging and Communications in Medicine (DICOM) format. The original DRO has a uniform background region that is 20 cm in the transaxial simulated human abdominal cross-section with an SUV of 1.00 (3.61 kBq/ml), while the six spheres contain FDG solutions with an SUV of 4.00 (14.47 kBq/ml). The central 5 cm-diameter cylinder is cold (SUV is 0.0) [20]. As two LBR values were included in this study, we created a modified version of the DRO using MATLAB (Mathwork 2016, which had an identical background while sphere SUVs were 2.00 for evaluating LBR = 2.

A three-dimensional (3-D) Gaussian filter was applied to the DROs to simulate the PET image with relatively decreased spatial resolution. The recovery coefficient of DROs for three types of SUVs (mean, max and peak) was situated within the EARL range under Gaussian filter when applying full width at half maximum (FWHM) values of 7–10 mm. The FWHM = 7 mm provided the highest recovery coefficients as well as being relatively the closest to the FWHM of collaborating PET systems. We selected $\mathrm{DRO}_{7\mathrm{mm}}$ followed by measurement of the three types of SUVs separately for both LBRs. These SUVs were used as references for reproducibility analysis and harmonization purposes.

Fig. 1


Step 2
Data acquisition
SUV
diameter(mm)
SUV plot
Step 3
1- Image analysis
2- SUV curves
SUV
diameter(mm)
Step 4
Before Harmonization:
1- Checking SUV curves with EARL reference range
2- Measuring Root Mean Square Error (RMSE) between each image's SUVs and those of DRO 7 mm
3- Measuring inter-center SUV variation
Step 5
Harmonization process:
Applying Gaussian filter with an appropriate FWHM to phantom's images to meet 3 criteria:
1- Aligned within EANM reference range
2- Minimized difference with DRO's values (optimal RMSE)
3- Preserving contrast to noise ratio (>2.8%)
Step 1
Phantom Preparation NEMA
SUV
diameter(mm)
Step 6
After Harmonization:
Filtered images with minimized SUV variability across scanners

Overall schematic of the study.

Table 1 **Specifications of collaborating PET/CT scanners**

PET/CT system		GE Discovery IQ-5 ring	Siemens Biograph	GE Discovery 690
Scanner characteristics	Crystal	BGO	LSO	LYSO
	Spatial resolution	5.1 mm	4.6 mm	5.6 mm
	Sensitivity	22 cps/kBq	4.2 cps/kBq	7 cps/kBq
	Axial field of view	26 cm	15.2 cm	15.7 cm
Dara acquisition	Scan duration	120 s	180 s	180 s
Image reconstruction	Reconstruction method	Iterative (OSEM) Non-Q.clear	Iterative (OSEM)	Iterative (OSEM) Non-TOF
	Iteration-Subset	4 It–12 Sub	2 It–21 Sub	3 It–18 Sub
	Matrix size	192 × 192	128 × 128	256 × 256
	Pixel spacing	3.64 × 3.64 × 3.26	4.07 × 4.07 × 3	2.73 × 2.73 × 3.27
	Filter – FWHM	Gaussian-6.4 mm	Gaussian-5 mm	Gaussian-6.4 mm
CT protocol	kV	120 kV	110 kV	120 kV
	mA	80 kV	70 mA	80 mA

FWHM, Full-width at half maximum; OSEM, ordered subset expectation maximization.

Image and data analysis

SUV is computed as follows:

$$\mathrm{SUV} = \frac{\text{Activity concentration in region of interest}(\mathrm{kBq/ml})}{\text{Injected activity }(\mathrm{kBq})/\text{Body weight}(\mathrm{kg})} \quad (1)$$

where maximum SUV in the lesion yields $\mathrm{SUV}_{\mathrm{max}}$, the average over a volume of interest gives $\mathrm{SUV}_{\mathrm{mean}}$ [21], and the $\mathrm{SUV}_{\mathrm{peak}}$ is defined as the average of all voxels within a 10 mL spherical region positioned within the lesion so as to maximize its mean value [11]. We placed six size- and location-matched spherical volume of interest over lesions for each acquired image, and measured three SUV categories in body weight scale.

The image quality was analyzed using $Q_{H,10\ mm}$ for percent contrast, $N_{10\,mm}$ for the background variability. These parameters were calculated using the following equations:

$$Q_{H,10\ mm10\ mm} = \frac{\frac{CH.10mm}{CB.10mm} - 1}{\frac{aH}{aB} - 1} \times 100\ (\%) \qquad (2)$$

$$N_{10mm} = \frac{SD10mm}{CB.10mm} \times 100\ (\%) \qquad (3)$$

where $C_{H,10\ mm}$ is the average count in the region of interest (ROI) for a 10 mm sphere, $C_{B,10\,mm}$ is the average count of the 60 circular ROIs on background located through five central slices, a_H and a_B are the true activity concentrations in the hot sphere and background, respectively and SD_{10mm} is the SD of the 60 background ROIs [6]. The image quality was assessed and compared with the reference value in the 'Japanese Guideline for Oncology FDG PET/CT Data Acquisition Protocol: Synopsis of Version 2.0' published by the JSNM and the Japanese Society of Nuclear Medicine Technology. The guideline recommended $Q_{H,10\ mm}/N_{.10mm}$ >2.8 (%) as the threshold for contrast noise ratio in image quality [22].

To assess SUV variability, we analyzed SUVs for different LBR values, SUV type, and multicenter considerations. The EANM/EARL specifications in recovery coefficient reference bandwidths as a function of lesion diameter for SUV_{mean} and SUV_{max} were our main standard ranges to check image SUV plots with. We also used the SUV_{peak} reference range proposed by Boellaard *et al.* 2017 that is not mentioned by the current EARL ranges [10]. The ranges were calculated for each collaborating LBR value by the following formula:

$$\text{Recovery Coef f icient } = \frac{\text{Measured}\quad \text{SUV}}{\text{Target}\quad \text{SUV}} \qquad (4)$$

First, in the absence of harmonization, we evaluated SUV inter-site variation for both LBR values and each SUV type based on the EARL range and DRO_{7mm} SUV plots. The root mean square error (RMSE) was used to measure the difference in SUVs between the DRO_{7mm} ($SUV_{ref,i}$) and actual phantom images for all the spheres (SUV_i; i=1…6).

$$\text{RMSE } = \sqrt{\frac{1}{6} \sum_{i=1\ldots6} (SUVi - SUVref.i)^2} \qquad (5)$$

Harmonization process

Harmonization was implemented as a postprocessing approach, which is briefly described in the following steps: First, each reconstructed image was loaded into the PMOD image quantification toolkit. We measured the SUVs of the hot spheres and checked against their corresponding EARL ranges and DRO_{7mm} plots. Then an additional 3D Gaussian filter with various FWHM ranging from 1 to 12 mm was incrementally applied to images. At each of applied FWHM, three criteria/metrics were regularly evaluated: (1) SUV plots in reference to EARL bandwidths, (2) RMSEs between image and DRO_{7mm} SUVs and (3) contrast to noise ratio for the smallest lesion had to satisfy >2.8% regarding image quality preservation. The FWHMs that enabled SUV plots to meet the bandwidths while preserving image quality, were selected to be candidates as FWHM values. Eventually, the FWHM that was able to minimize RMSE was selected as the appropriate harmonization parameter, which had to be applied as a postreconstruction filter. In other words, while a number of filters enabled satisfaction of EARL bandwidth and minimum quality criteria, we used RMSE with respect to blurred DROs to select optimum filters. It should be noted that we implemented this entire procedure for the three types of SUVs (mean, max and peak) separately for each loaded image.

Statistical evaluation and test-retest analysis

The coefficient of variation (CV) was computed in the SUVs of different lesions in images that were acquired across different centers to measure inter-site variability. CV measurement was performed for pre- and post-harmonization for all SUV categories. For the test-retest images acquired, we used several statistical metrics such as the Intra-class correlation coefficient (ICC) and R-squared so as to assess SUV repeatability. We also used the mean absolute percentage difference (MAPD) to measure test-retest SUV fluctuations, as determined by the following equation [11]:

$$\text{MAPD } = \frac{1}{N}\sum_{i=1}^{N} \frac{|\text{test (i)} - \text{retest(i)}|}{(\text{test (i)} + \text{retest(i)})/2} \times 100 \qquad (6)$$

where N is the number of cases. Pre- and post-harmonization CV differences were analyzed by Wilcoxon signed-ranked tests using SPSS packages (SPSS, version 22.0, Armonk, New York, USA). Statistical significance was set at $P < 0.05$.

Results

Standardized uptake values

Figure 2 shows SUV plots as a function of lesion diameter obtained from the scanners, as evaluated in the presence of EARL reference bandwidth and corresponding DRO reference plots. Graphs A–F indicates identical analysis for SUV_{mean}, SUV_{max} and SUV_{peak} in the case of two LBR values of 2 and 4. In the absence of harmonization, noticeable differences were observed in SUV plots relevant to different scanners. The largest inter-site relative differences in SUVs were 17.3, 20.7 and 15.5% for $\mathrm{SUV}_{\mathrm{mean}}$, $\mathrm{SUV}_{\mathrm{max}}$ and $\mathrm{SUV}_{\mathrm{peak}}$, respectively.

Harmonization

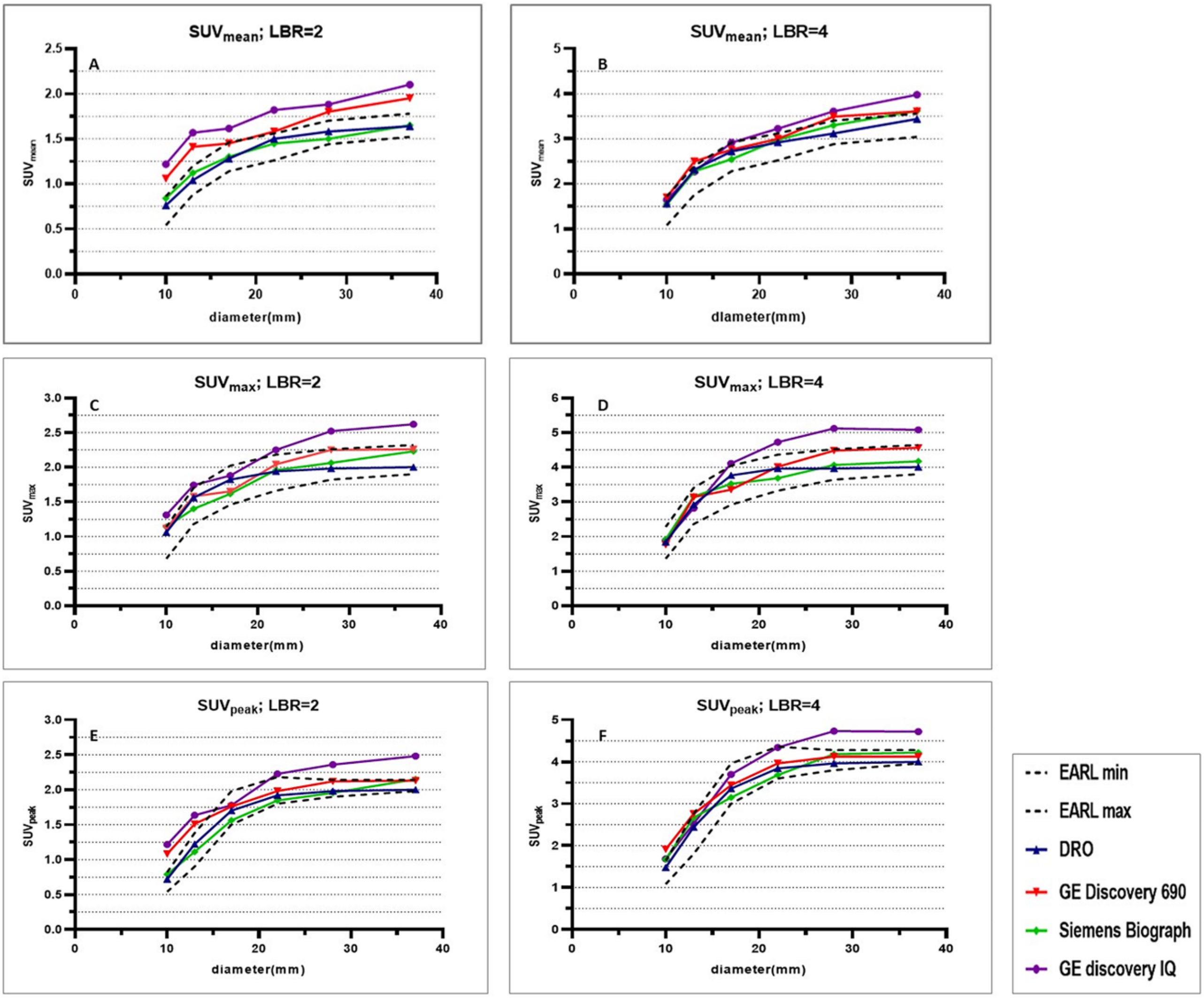
Table 2 shows the FWHM values used for the postprocessing Gaussian filter, which brings about the alignment of the SUVs within the EARL bandwidth. The optimum value introduces the FWHM that minimizes RMSE between DRO and image SUVs. As different ranges and DRO values were used for different SUV types, the FWHM range and optimum values were determined separately.

Figure 3 shows postharmonization SUV plots of three scanners. Similar to before, graphs A–F depict analysis for $\mathrm{SUV}_{\mathrm{mean}}$, $\mathrm{SUV}_{\mathrm{max}}$ and $\mathrm{SUV}_{\mathrm{peak}}$ in LBR = 2 and 4 separately. SUVs fell within the EARL reference bandwidths and became closer together as a result of harmonization. The largest inter-site differences in SUVs decreased to 4.8, 4.7 and 2.7% for $\mathrm{SUV}_{\mathrm{mean}}$, $\mathrm{SUV}_{\mathrm{max}}$ and $\mathrm{SUV}_{\mathrm{peak}}$, respectively.

Reproducibility

Figure 4 shows the inter-site coefficient of variation ($\mathrm{CV}_{\mathrm{reproducibility}}$) of PET SUVs across three scanners pre- and postharmonization, where the six graphs relate to the three SUV types and two LBR values. Harmonization lowered $\mathrm{CV}_{\mathrm{reproducibility}}$ range in $\mathrm{SUV}_{\mathrm{mean}}$, $\mathrm{SUV}_{\mathrm{max}}$ and $\mathrm{SUV}_{\mathrm{peak}}$ from 7–11%, 6–11% and 5–10% to 2–7%, 3–7% and 2–5%, respectively ($P < 0.05$).

Fig. 2



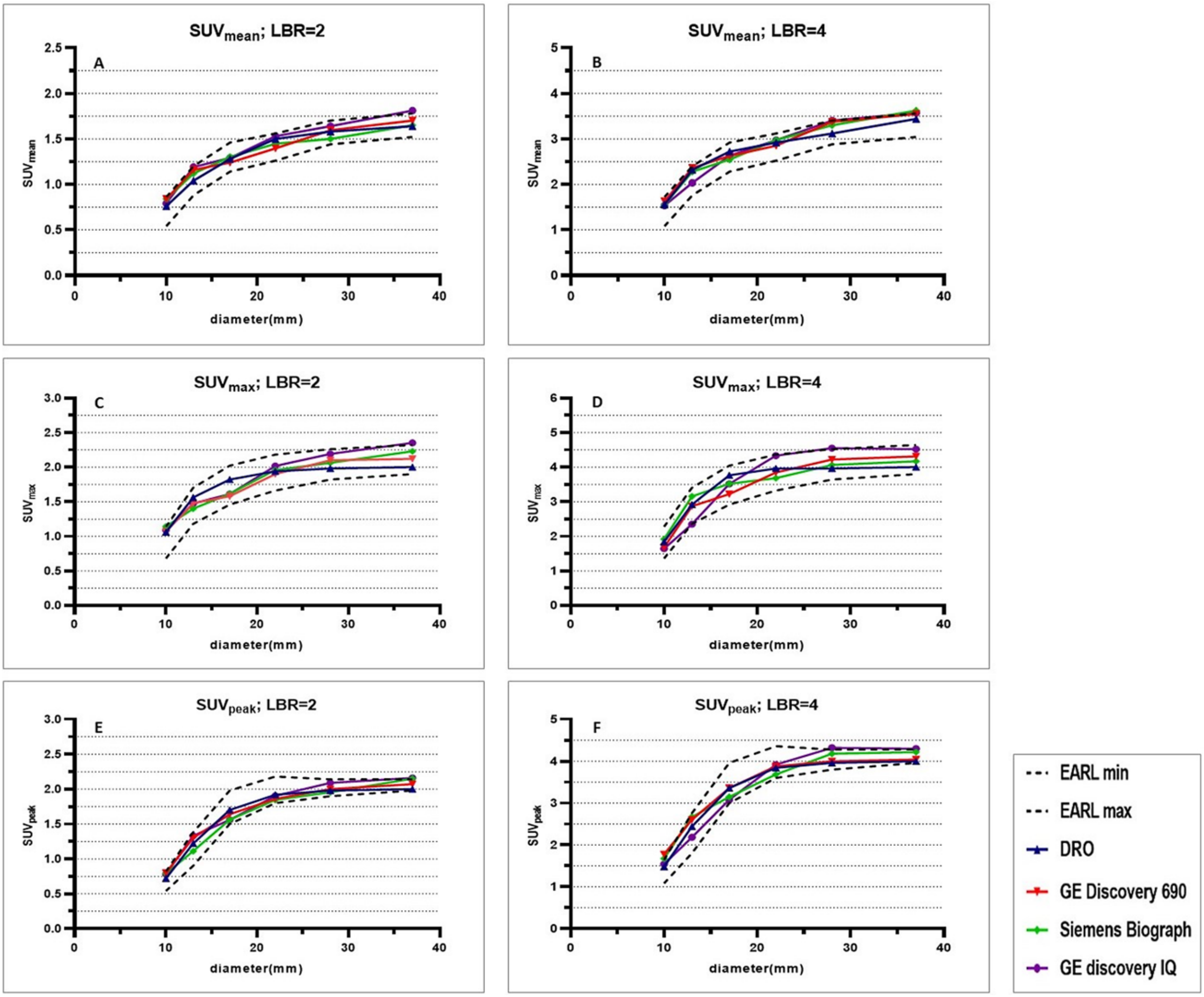
Preharmonization SUV plots of the three PET/CT centers, in the context of EARL bandwidths and DRO plot. Plots of $\mathrm{SUV}_{\mathrm{mean}}$, $\mathrm{SUV}_{\mathrm{max}}$ and $\mathrm{SUV}_{\mathrm{peak}}$ for two LBR values are shown. DRO, digital reference object; LBR, lesion-to-background ratio; SUV, standardized uptake value.

Table 2 **Full-width at half maximum of post-processing filters in standardized uptake value harmonization of the scanners**

SUV type	SUV mean	SUV max		SUV peak		
Scanner	FWHM range	Optimum FWHM	FWHM range	Optimum FWHM	FWHM range	Optimum FWHM
GE Discovery IQ	8–11 mm	8 mm	7–9 mm	7 mm	8–9 mm	8 mm
GE Discovery 690	3–7 mm	3 mm	0–6 mm	4 mm	4–7 mm	4 mm
Siemens Biograph	No filter	–	No filter	–	No filter	–

FWHM, Full-width at half maximum; SUV, standardized uptake value.

Fig. 3



Postharmonization SUV plots of the three PET/CT centers, in the context of EARL bandwidths and DRO plot. Plots of SUV_{mean}, SUV_{max} and SUV_{peak} for two LBR values are shown. DRO, digital reference object; LBR, lesion-to-background ratio; SUV, standardized uptake value.

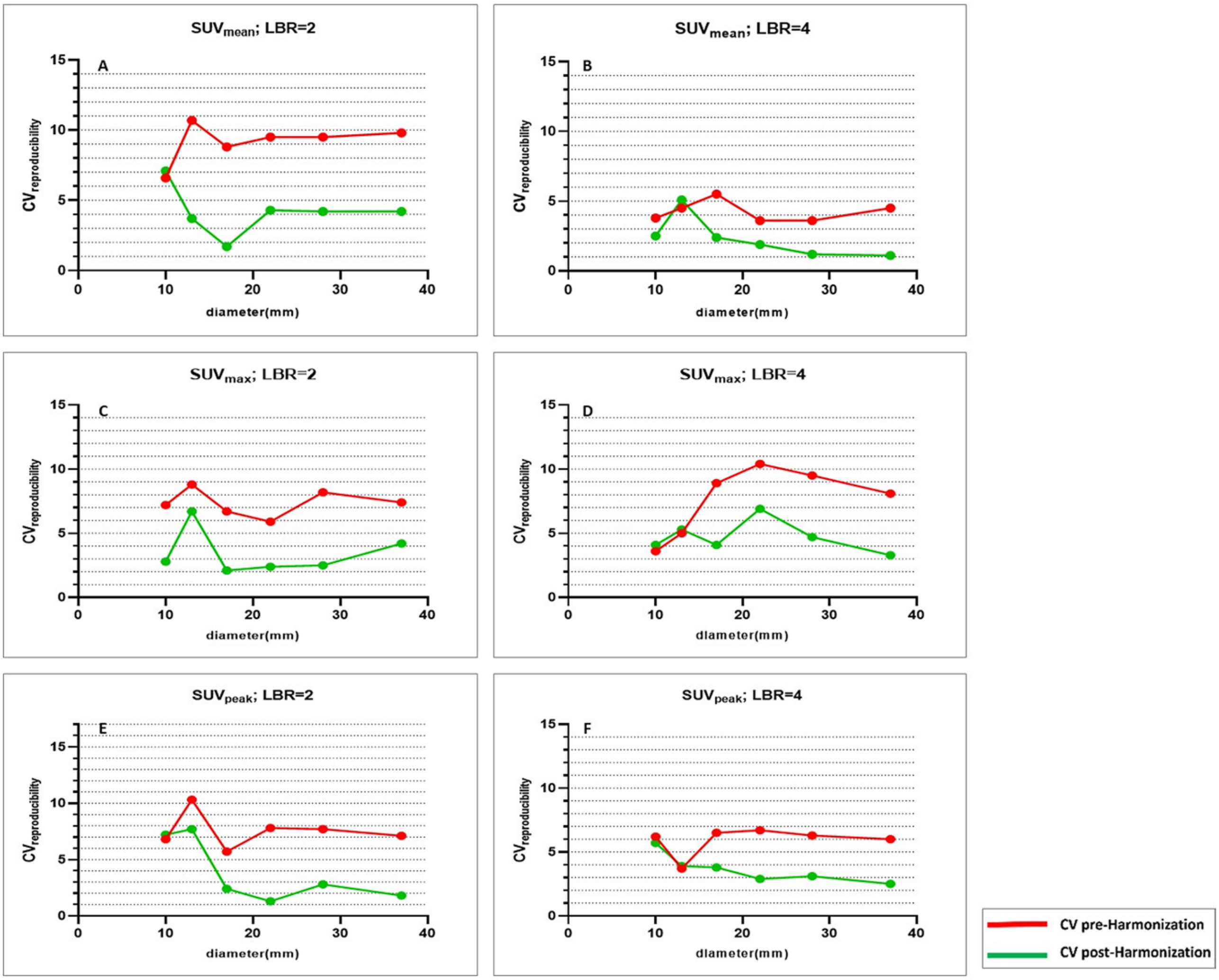
Image quality

Figure 5 shows the contrast to noise ratio for the smallest lesion (10 mm in diameter) for each scanner image in relation to image quality preservation in harmonization. This parameter clearly declined following harmonization. Nevertheless, they satisfied the above-mentioned condition (>2.8%) in all cases.

Test-retest analysis

Table 3 is associated with SUV repeatability assessment using test-retest analysis, where each row relates to a specific statistical parameter calculated for each SUV type in three scanners. Figure 6 shows a set of nine test-retest graphs, whose horizontal and vertical axes represented test and retest values respectively. The linear regression

Fig. 4



Influence of harmonization on SUV reproducibility. Pre- and postharmonization plots of the coefficient of variation across the scanners for SUV_{mean}, SUV_{max} and SUV_{peak}. SUV, standardized uptake value.

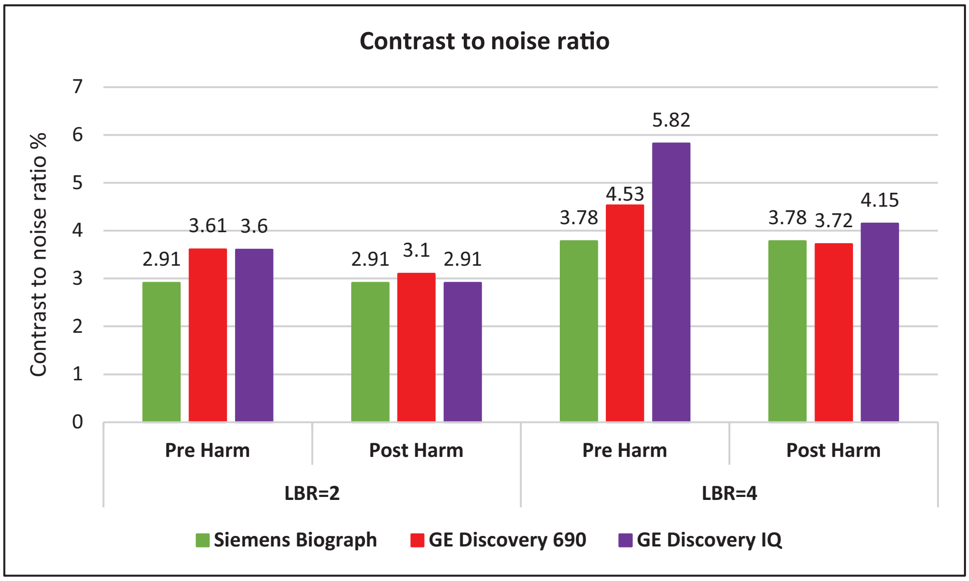
equation, as well as R-square, are shown on each graph. R^2 values were above 0.9 and slopes were within a range of 0.89–1.11, indicating that test and retest values are approximately located on the bisector line in the first sector of the Cartesian coordinates. In addition, ICC and MAPD values were in the range of 0.986–0.998 and 3.35–9.08, respectively.

Discussion

The phantom scanned by the different systems yielded distinct SUV curve outputs (Fig. 2), with images from the Siemens Biograph, GE 690 and GE IQ depicting the least to the highest values. Considering Table 1 and Fig. 2, the reason for the discrepancy and also ordering in SUV values can be attributed to different characteristics including scanner design and image reconstruction. In spite of having the same reconstruction algorithm ordered subset expectation maximization, the systems had differences in parameters such as the number of subsets, iteration and matrix size, which led to significant differences in SUV values. Shiri *et al.* [23] showed that the reconstruction parameters have a profound impact on the variation of PET/CT quantitative metrics and their reproducibility.

Figure 2 indicated that inter-site variations had different behaviors depending on the SUV type. The highest variability could be noticed in SUV_{max} while the least was seen in SUV_{peak}. Because the SUV_{max} is computed by using one voxel within an ROI, it is more sensitive to noise and the factors that contribute to counts in voxels. Burger *et al.* [24], by taking into account several types of SUV_{max} and comparing it with SUV_{mean}, showed that SUV_{max} undergoes more fluctuations. De Langen *et al.* [25] in their meta-analysis showed that SUV_{mean} had

Fig. 5



Contrast to noise ratio of the smallest lesion (10 mm in diameter) for two LBR values. Comparison of pre- and post-harmonization for images from the centers. LBR, lesion-to-background ratio

Table 3 **Statistical parameters for standardized uptake value repeatability assessment in test-retest analysis.**

SUV type	SUV mean			SUV max			SUV peak		
Scanner Statistical parameter	Siemens	GE 690	GE IQ	Siemens	GE 690	GE IQ	Siemens	GE 690	GE IQ
R-square	0.99	0.98	0.99	0.98	0.99	0.99	0.98	0.99	0.99
Linear regression slope	1.005	0.89	1.009	1.11	1.019	1.12	1.05	1.05	1.11
ICC	0.996	0.997	0.998	0.991	0.994	0.986	0.992	0.997	0.992
MAPD (%)	3.95	3.35	3.47	5.75	4.45	9.08	6.68	4.11	5.4

ICC, Intra-class correlation coefficient; MAPD, mean absolute percentage difference;SUV, standardized uptake value.

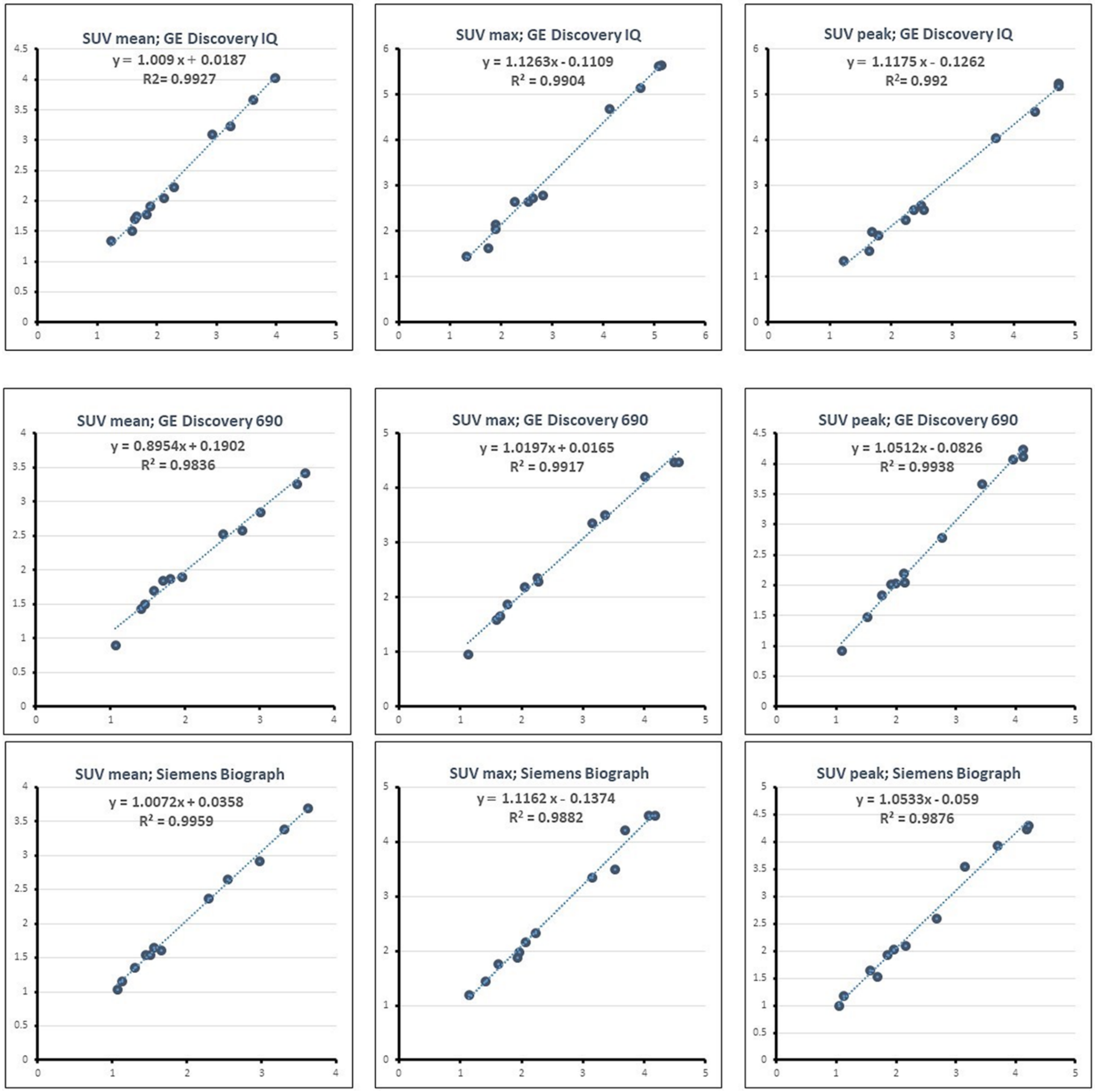
better repeatability performance than SUV_{max}; however, both measures showed poor repeatability for lesions with low FDG uptakes. SUV_{peak}, having no dependency on ROI placement/definition compared to SUV_{mean} and lower susceptibility to noise compared to SUV_{max}, resulted in less variability across the scanners. Sher *et al.* [26] presented SUV_{peak} as the most reliable parameter for FDG-PET/CT quantification. A similar pattern is also seen in the thorough review by Lodge *et al.* [11], where variability appears least (i.e. repeatability best) for SUV_{peak}, followed by SUV_{mean}, followed by SUV_{max}, though the differences are small.

According to Table 2, ‘No filter’ for the images of Biograph indicated that the images did not need additional Gaussian filters for harmonization. This resulted from the fact that this relatively older scanner had good agreement with EARL bandwidths, while the two later-generation scanners provided higher SUVs that fell outside the bandwidths; as such the latter two scanners needed additional postsmoothing, as stated in Table 2, to be harmonized.

According to Fig. 3, the SUVs generally declined by applying a smoothing filter in harmonization. All SUV curves fell within the EARL range and the discrepancy of the SUV curves decreased to a minimum level compared with before harmonization. The recovery coefficient ranges of the EARL bandwidths are more compatible with relatively older scanners. In newer-generation PET/CT scanners, with improved resolution performances, new SUV curves will need filters with higher FWHM to be harmonized (e.g. GE IQ in this study). This poses a challenge as more such scanners are released to the market: to tackle this, Kaalep *et al.* [10] argued that current EARL ranges need to be updated by considering higher recovery coefficient, which is a work in progress.

Comparison of $CV_{reproducibility}$ before and after harmonization in Fig. 4 indicated that harmonization improved the reproducibility in all SUV types by lowering $CV_{reproducibility}$ (*P* value <0.05 for all three SUV types). By applying proper filters, SUVs achieved lowered values whereby

Fig. 6



Test-retest plots of SUV where the horizontal and vertical axes represent test and retest SUVs, respectively. Plots of SUV_{mean}, SUV_{max} and SUV_{peak} for three PET/CT centers are shown. The linear regression equation and R^2 have been calculated for each. SUV, standardized uptake value.

the resulting curves fell within the standard range and closely overlaid one another.

With the increase in the FWHM of filters, the percent contrast decreased. This caused degradation in image detectability, particularly for small lesions. The condition for acceptability of FWHM while preserving image quality was that it should not reduce the contrast-to-noise ratio in the smallest lesion to less than 2.8%. Figure 5 verified that images of all three scanners, for both values of LBR, in the two pre- and post-harmonization situations, subscribed to the aforementioned rule. Images after harmonization at LBR = 2 had smaller percent contrast compared to LBR = 4, which indicates that after harmonization low-uptake lesions (particularly small sizes because of the partial volume effect) are more compromised to being invisible in images. As such, a recommended practice is to apply postsmoothing to render

SUVs consistent across centers, while for diagnostic purposes, original images are utilized. In other words, two sets of images are applied, one for diagnosis and one for quantification, where original images are used for the former, and harmonization for the latter [9,27].

The data in Fig. 6 and Table 3 in which all values were ICC >0.9, 0.9 <linear regression slope <1.1, MAPD <10% and R^2 >0.9, showed that all SUV types in all three scanners benefited from excellent repeatability. In investigations involving patient data, variability in test-retest measurements are more significant: this is because biologic factors contribute significantly to higher test-retest variability in SUV [28], estimated in one study to be nearly half of the overall repeatability [29]. This means that differences in test-retest repeatability between scanners, in the present phantom study, would be amplified compared to patient studies, as biologic variability should not be dependent on the specific scanner.

Similar to the research conducted by Nakahara *et al.* [30], there are several limitations in our study. First, while a large number of harmonization surveys use many PET/CT scanners, only three scanners were included in this study. In addition, the DRO_{7mm} was determined as a reference image, however, it is not a universal rule. As mentioned earlier, according to situating recovery coefficients plots of the smoothed DRO within the EARL range as well as being the closest to the FWHM of collaborating PET systems, DRO_{7mm} was elected as the optimal reference in our study. DROs smoothed with different FWHMs may also be references for another multicenter study depending on the spatial resolution of the systems. Finally, we collected PET/CT data of the NEMA phantom with only two low LBR values. Whether the results would be applicable for other contrasts such as high uptake lesions remains unknown. Examining this issue could be one of the top priorities for further research.

Conclusion

This study evaluated variation in PET SUVs among different scanners with a focus on low uptake lesions. Harmonization based on applying appropriate postreconstruction smoothing filters minimized misleading differences in SUVs across systems. As a result, two sets of images including original and filtered images were used for diagnosis and quantification respectively. This reduction improved reproducibility for quantification purposes while preserving image quality at an acceptable level. Using DRO alongside standard accepted EARL ranges yields more accurate harmonization. At the same time, the current EARL bandwidths need to be updated, considering higher recovery coefficient values compatible with developments in recent PET/CT scanners.

Acknowledgements

The authors thank the staff of PET/CT centers at three collaborating university hospitals given their valuable assistance.

This work was supported by the Tehran University of Medical Sciences, Grant No. 40272.

Conflicts of interest

There are no conflicts of interest.

References

1 Kinahan P, Fletcher JW. Ultrasound JF-S in, MRI C and, 2010 undefined. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. Elsevier [Internet]. [cited 2018 Oct 7]; Available from: https://www.sciencedirect.com/science/article/pii/S0887217110000880

2 Byrd D, Doot R, Allberg KC, MacDonald LR, McDougald WA, Elston BF. A Journal For Imaging, 2016 undefined. Evaluation of cross-calibrated 68Ge/68Ga phantoms for assessing PET/CT measurement bias in oncology imaging for single-and multicenter trials. ncbi.nlm.nih.gov [Internet]. [cited 2018 Oct 7]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5214172/

3 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.

4 Kurland BF, Peterson LM, Shields AT, Lee JH, Byrd DW, Novakova-Jiresova A, *et al.* Test-Retest reproducibility of 18F-FDG PET/CT uptake in cancer patients within a qualified and calibrated local network. *J Nucl Med* 2019; **60**:608–614.

5 Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med* 2009; **50**(Suppl 1):122S–150S.

6 Tsutsui Y, Awamoto S, Himuro K, Umezu Y, Baba S, Sasaki M. Characteristics of smoothing filters to achieve the guideline recommended positron emission tomography image without harmonization. *Asia Ocean J Nucl Med Biol* 2018; **6**:15–23.

7 Visser EP, Boerman OC, Oyen WJ. SUV: from silly useless value to smart uptake value. *J Nucl Med* 2010; **51**:173–175.

8 Huang SC. Anatomy of SUV. Standardized uptake value. *Nucl Med Biol* 2000; **27**:643–646.

9 Kelly MD, Declerck JM. SUVref: reducing reconstruction-dependent variation in PET SUV. *EJNMMI Res* 2011; **1**:16.

10 Kaalep A, Sera T, Rijnsdorp S, Yaqub M, Talsma A, Lodge MA, Boellaard R. Feasibility of state of the art PET/CT systems performance harmonisation. *Eur J Nucl Med Mol Imaging* 2018; **45**:1344–1361.

11 Lodge MA. Repeatability of SUV in oncologic 18F-FDG PET. *J Nucl Med* 2017; **58**:523–532.

12 Doot RK, Kurland BF, Kinahan PE, Mankoff DA. Design considerations for using PET as a response measure in single site and multicenter clinical trials. *Acad Radiol* 2012; **19**:184–190.

13 Lasnon C, Desmonts C, Quak E, Gervais R, Do P, Dubos-Arvis C, Aide N. 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.

14 Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys* 2018; **102**:1143–1158.

15 Sullivan DC, Obuchowski NA, Kessler LG, Raunig DL, Gatsonis C, Huang EP, *et al.*; RSNA-QIBA Metrology Working Group. Metrology standards for quantitative imaging biomarkers. *Radiology* 2015; **277**:813–825.

16 Young H, Baum R, Cremerius U. Position paper. Measurement of clinical and subclinical tumour response using -fluorodeoxyglucose and positron emission tomography: review and 1999. *Eur J [Internet]* 1999; **35**:1773–1782. Available from: https://www.sciencedirect.com/science/article/pii/S0959804999002294?via%3Dihub

17 Tsutsui Y, Daisaki H, Akamatsu G, Umeda T, Ogawa M, Kajiwara H, *et al.* Multicentre analysis of PET SUV using vendor-neutral software: the Japanese Harmonization Technology (J-Hart) study. *EJNMMI Res* 2018; **8**:1–10.

18 NEMA Standards Publication NU 2. NEMA Standards Publication NU 2-2018. Performance measurements of positron emission tomographs. *Natl Electr Manuf Assoc* 2012. https://psec.uchicago.edu/library/applications/PET/chien_min_NEMA_NU2_2007.pdf.

19 Pierce LA 2nd, Elston BF, Clunie DA, Nelson D, Kinahan PE. A digital reference object to analyze calculation accuracy of PET standardized uptake value. *Radiology* 2020; **294**:647–657.

20 Quantitative Imaging Biomarkers Alliance. QIBA Profile: 18F-FDG PET/CT UPICT Protocol Writing Committee. *Radiology* 2017; **44**:17–31.

21 Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *Eur J Nucl Med Mol Imaging* 2017; **44**:17–31.

22 JSNM. Standard PET imaging protocols and phantom test procedures and criteria: executive summary. *Japanese Soc Nucl Med [Internet]* 2014;1–6. Available from: https://www.semanticscholar.org/paper/Standard-PET-imaging-protocols-and-phantom-test-and/d58897dd0b10a63fb1aa1c5c6c545982bde5b84a#citing-papers

23 Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol* 2017; **27**:4498–4509.

24 Burger IA, Huser DM, Burger C, von Schulthess GK, Buck A. Repeatability of FDG quantification in tumor imaging: averaged SUVs are superior to SUVmax. *Nucl Med Biol* 2012; **39**:666–670.

25 de Langen AJ, Vincent A, Velasquez LM, van Tinteren H, Boellaard R, Shankar LK, *et al*. Repeatability of 18F-FDG uptake measurements in tumors: a metaanalysis. *J Nucl Med* 2012; **53**:701–708.

26 Sher A, Lacoeuille F, Fosse P, Vervueren L, Cahouet-Vannier A, Dabli D, *et al*. For avid glucose tumors, the SUV peak is the most reliable parameter for [18F]FDG-PET/CT quantification, regardless of acquisition time. EJNMMI Res [Internet]. *EJNMMI Research* 2016; **6**:4–9.

27 Quak E, Le Roux PY, Hofman MS, Robin P, Bourhis D, Callahan J, *et al*. Harmonizing FDG PET quantification while maintaining optimal lesion detection: prospective multicentre validation in 517 oncology patients. *Eur J Nucl Med Mol Imaging* 2015; **42**:2072–2082.

28 Kramer GM, Liu Y, de Langen AJ, Jansma EP, Trigonis I, Asselin MC, *et al*.; QuIC-ConCePT consortium. Repeatability of quantitative 18F-FLT uptake measurements in solid tumors: an individual patient data multi-center meta-analysis. *Eur J Nucl Med Mol Imaging* 2018; **45**:951–961.

29 Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. *J Nucl Med* 2012; **53**:1041–1047.

30 Nakahara T, Daisaki H, Yamamoto Y, Iimori T, Miyagawa K, Okamoto T, *et al*. Use of a digital phantom developed by QIBA for harmonizing SUVs obtained from the state-of-the-art SPECT/CT systems: a multicenter study. *EJNMMI Res* 2017; **7**:53.