Annals of Nuclear Medicine (2022) 36:144–161
https://doi.org/10.1007/s12149-021-01709-1

SPECIAL ARTICLE

New standards for phantom image quality and SUV harmonization range for multicenter oncology PET studies

Go Akamatsu[1] · Naoki Shimada[2] · Keiichi Matsumoto[3] · Hiromitsu Daisaki[4] · Kazufumi Suzuki[5] · Hiroshi Watabe[6] · Keiichi Oda[7] · Michio Senda[8] · Takashi Terauchi[2] · Ukihide Tateishi[9]

Received: 14 October 2021 / Accepted: 5 December 2021 / Published online: 14 January 2022

Abstract

Not only visual interpretation for lesion detection, staging, and characterization, but also quantitative treatment response assessment are key roles for ^{18}F-FDG PET in oncology. In multicenter oncology PET studies, image quality standardization and SUV harmonization are essential to obtain reliable study outcomes. Standards for image quality and SUV harmonization range should be regularly updated according to progress in scanner performance. Accordingly, the first aim of this study was to propose new image quality reference levels to ensure small lesion detectability. The second aim was to propose a new SUV harmonization range and an image noise criterion to minimize the inter-scanner and intra-scanner SUV variabilities. We collected a total of 37 patterns of images from 23 recent PET/CT scanner models using the NEMA NU2 image quality phantom. PET images with various acquisition durations of 30–300 s and 1800 s were analyzed visually and quantitatively to derive visual detectability scores of the 10-mm-diameter hot sphere, noise-equivalent count ($NEC_{phantom}$), 10-mm sphere contrast ($Q_{H,10\ mm}$), background variability ($N_{10\ mm}$), contrast-to-noise ratio ($Q_{H,10\ mm}/N_{10\ mm}$), image noise level (CV_{BG}), and SUVmax and SUVpeak for hot spheres (10–37 mm diameters). We calculated a reference level for each image quality metric, so that the 10-mm sphere can be visually detected. The SUV harmonization range and the image noise criterion were proposed with consideration of overshoot due to point-spread function (PSF) reconstruction. We proposed image quality reference levels as follows: $Q_{H,10\ mm}/N_{10\ mm} \geq 2.5$ and $CV_{BG} \leq 14.1\%$. The 10th–90th percentiles in the SUV distributions were defined as the new SUV harmonization range. $CV_{BG} \leq 10\%$ was proposed as the image noise criterion, because the intra-scanner SUV variability significantly depended on CV_{BG}. We proposed new image quality reference levels to ensure small lesion detectability. A new SUV harmonization range (in which PSF reconstruction is applicable) and the image noise criterion were also proposed for minimizing the SUV variabilities. Our proposed new standards will facilitate image quality standardization and SUV harmonization of multicenter oncology PET studies. The reliability of multicenter oncology PET studies will be improved by satisfying the new standards.

Keywords PET · Image quality · SUV · Standardization · Harmonization

Introduction

Whole-body ^{18}F-fluorodeoxyglucose (FDG) PET imaging has been widely used in the management of various malignant cancers [1–3]. Not only lesion detection, staging, and characterization, but also therapy response assessment are key roles for FDG PET in oncology [4]. With the advent of molecular targeted therapy and immunotherapy, metabolic activity of tumors is frequently assessed by quantitative FDG PET imaging. FDG PET has become a *quantitative* imaging biomarker, moving beyond a *qualitative* functional imaging tool [5, 6].

Go Akamatsu and Naoki Shimada have contributed equally to this work.

✉ Go Akamatsu
akamatsu.go@qst.go.jp

✉ Naoki Shimada
naoki.shimada@jfcr.or.jp

Extended author information available on the last page of the article

For measuring responses to therapy by FDG PET, major methodologies such as the EORTC criteria and PERCIST have been proposed [7, 8]. In these methodologies, tumor response is assessed by visual interpretation as well as percentage change in standardized uptake values (SUVs), and then classified into the following four definitions: complete metabolic response (CMR), partial metabolic response (PMR), stable metabolic disease (SMD), and progressive metabolic disease (PMD). In this manner, maximum and peak SUVs (SUVmax, SUVpeak) and SUVs normalized by lean body mass (SULs) have been used as quantitative markers for primary and secondary endpoints in FDG PET studies and trials in oncology [9–11].

However, PET image quality and quantitative accuracy are considerably affected by numerous factors such as injection activity, uptake duration, subject body size, scanner specifications, and image reconstruction parameters [12, 13]. Figure 1 overviews the factors affecting diagnostic accuracy in FDG PET. Small lesion detectability and tumor SUVs are easily made variable owing to these many factors. This variability may not have a significant impact on results in the case of a single-scanner study. In multicenter studies using multiple scanners, however, the inter-scanner variability might seriously degrade the reliability of the study outcomes [14]. Therefore, in multicenter oncology FDG PET studies, imaging protocols and image characteristics should be verified and standardized using an appropriate phantom before starting the study. As stated by Boellaard [12], the required level of standardization depends on the intended use of FDG PET. When PET is used for visual interpretation such as lesion detection and characterization, image quality should be verified and standardized to ensure detectability of small lesions. On the other hand, more strict standards are required for quantitative PET. When using lesion SUVs to measure responses to certain therapies [8], harmonization of SUVs is essential to minimize the inter-scanner variability in SUVs [15]. Groups led by Kinahan have reported that reducing variability to measure true metabolic change can greatly reduce the required sample size and study costs [16, 17]. Therefore, image quality standardization and SUV harmonization are essential to improve the reliability of multicenter oncology PET studies.

Motivated by this issue, several organizations such as EANM/EARL, RSNA/QIBA, ACR/ACRIN, and SNMMI/CTN have provided their own criteria for optimizing image quality as well as reducing SUV variability [18–26]. In Japan, the Japanese Society of Nuclear Medicine (JSNM) provides the standard PET imaging protocol and phantom test procedures with the NEMA NU2 image quality phantom (NEMA body phantom) [27, 28]. The JSNM presents image quality reference levels and an SUV harmonization range for each sphere of the phantom (10–37 mm diameters). However, the reference levels and specified range were determined by the phantom data that had been acquired in the early 2010s with the PET scanners available at that time [29]. In the meantime, clinical PET scanner performance has been improved by recent novel technologies such as the point-spread function (PSF) modeling [30, 31], time-of-flight (TOF) measurements [32, 33], and the penalized likelihood reconstruction algorithm [34]. In particular, TOF coincidence timing resolution has been greatly improved by replacing the conventional photomultiplier tube (PMT) with a newer silicon photomultiplier (SiPM) [35–38]. With such new technologies, recent PET scanners can visualize small spheres with higher SUVs (a smaller partial volume effect). Because their SUVmax recovery curves often exceed the upper range, downsmoothing is required to satisfy the current

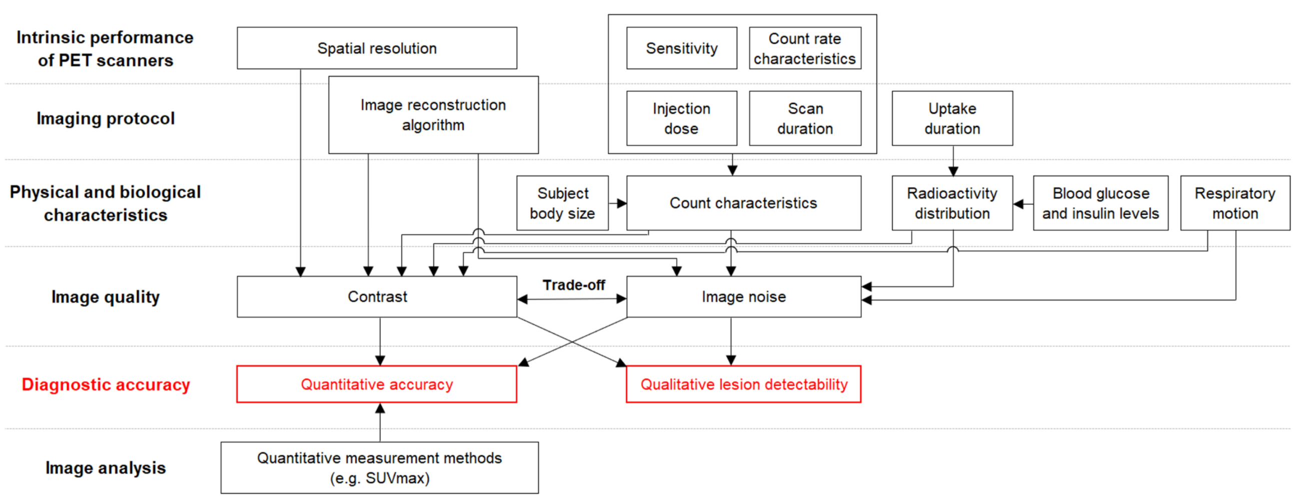

Fig. 1 Factors affecting the diagnostic accuracy of FDG PET in oncology

range. Although downsmoothing of the images is a simple way to harmonize, it spoils the image contrast and may degrade the visual detectability of small lesions. To adapt to advanced PET scanners with better performance, image quality reference levels and the range for SUVmax should be updated accordingly [12]. Also, a harmonization range for SUVpeak should be established, because this term has been widely used in many clinical studies [12, 39–42].

In addition to SUV harmonization (minimizing the *inter-scanner* variability), image noise levels should be lowered to reduce the *intra-scanner* variability. Increasing image noise levels (e.g., short scan duration) would provide a positive bias for SUVs [43]. A sufficient scan duration is needed to reduce uncertainties in SUV measurements as much as possible [44]. The relationship between SUV variability and image noise levels should be investigated in detail to establish reasonable criteria for image noise levels. The combination of SUV harmonization and image noise management can lead to significant improvement in the value and reliability of quantitative FDG PET studies (Fig. 2).

Motivated by these backgrounds, we investigated image quality and SUV variability in hot spheres of almost all recent PET/CT scanner models using an image quality phantom. The first aim of this study was to propose new image quality reference levels with a focus on 10 mm sphere detectability. The second aim was to propose a new SUV harmonization range and an image noise criterion for minimizing the inter-scanner and intra-scanner SUV variabilities.

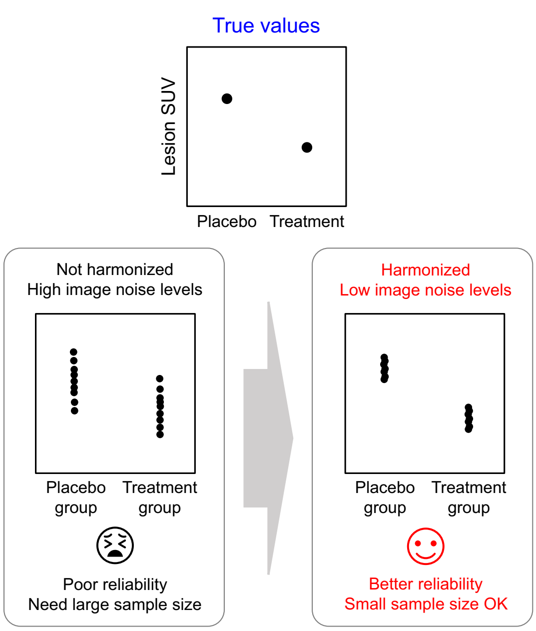

Fig. 2 Significance of SUV harmonization and image noise management in multicenter quantitative PET studies

Materials and methods

PET/CT scanners

Table 1 lists the PET/CT scanner models and image reconstruction parameters used in this study. Detailed scanner specifications and correction methods are summarized in Supplemental Table 1 [45–61]. We evaluated the 23 scanner models (16 PMT-based scanners and 7 SiPM-based scanners) used at 19 clinical sites. Phantom data were acquired from November 2018 to May 2020. This study did not include human data or any personal information.

Phantom experiments

Phantom measurements were performed according to the JSNM phantom test procedures [27]. The NEMA NU2 image quality phantom (NEMA body phantom) was used for all evaluations. We provided the phantom test procedure manual to all sites, and we visited several sites and supported the phantom test, if necessary. The phantom contains six spheres, having diameters of 10, 13, 17, 22, 28, and 37 mm. All spheres were filled with ^{18}F-FDG solutions, so that the sphere-to-background activity ratio was 4. The activity concentration in the background area was 2.53 ± 0.13 ($\pm 5\%$) kBq/mL, which was determined by the following equation:

$$A_x = \frac{a}{60} \times \exp\left(\frac{-60}{109.8} \times \ln(2)\right) \times S \text{ [kBq/mL]}, \quad (1)$$

where A_x (kBq/mL) is the activity concentration in the background area, a (MBq) is the assumed injection activity for 60-kg subjects, and S is the assumed specific gravity of a human body, that is 1.0 (g/mL). Since the assumed injection dose was 3.7 MBq/kg in this study, a was 222 MBq (3.7×60). The patient's weight section (0010, 1030) of the DICOM header was filled with the phantom background volume, so that the true SUV was 1.00 in the background area.

Data acquisition and image reconstruction

Emission data were acquired for 1800s in list mode. PET images were reconstructed with various acquisition durations of 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, and 1800s. For each acquisition duration except 1800s, three image datasets were reconstructed by changing the data start time of 0, 60, and 120 s. Table 1 shows the image reconstruction parameter, which is the setting for clinical whole-body FDG PET imaging used at each site. For the scanner models with PSF reconstruction, both PET images were reconstructed with and without PSF modeling. A total of 37 patterns of images were obtained. In the data analyses

Table 1 PET/CT scanners and image reconstruction settings

Scanner model	Reconstruction algorithm			Reconstruction parameter	Smoothing	Matrix size	Voxel size (mm^3)
	TOF + PSF	TOF	PSF				
Celesteion	✓	✓		it. 3, sub. 10	Gauss 6.0 mm	272 × 272 × 96 (TOF + PSF) 144 × 144 × 48 (TOF)	2.04 × 2.04 × 2.04 (TOF + PSF) 4.08 × 4.08 × 4.08 (TOF)
Biograph Vision 600	✓	✓		it. 4, sub. 5	N/A	220 × 220 × 53	3.30 × 3.30 × 5
Biograph Vision 450	✓	✓		it. 3, sub. 5	N/A	440 × 440 × 99	1.65 × 1.65 × 2
Biograph Horizon 4R	✓	✓		it. 4, sub. 10	Gauss 5.0 mm	360 × 360 × 111	2.06 × 2.06 × 3
Biograph Horizon 3R	✓	✓		it. 4, sub. 10	Gauss 5.0 mm	180 × 180 × 55	4.11 × 4.11 × 5
Biograph mCT 4R	✓	✓		it. 2, sub. 21	Gauss 5.0 mm	400 × 400 × 111	2.04 × 2.04 × 3
Biograph mCT 3R	✓	✓		it. 2, sub. 21	Gauss 6.0 mm	256 × 256 × 55	3.18 × 3.18 × 5
Discovery MI 5R	✓(Q.Clear)			β350	N/A	384 × 384 × 89	1.30 × 1.30 × 2.80
Discovery MI 4R	✓(Q.Clear)			β600	N/A	256 × 256 × 71	1.95 × 1.95 × 2.79
Discovery MI 3R	✓(Q.Clear)			β900	N/A	256 × 256 × 53	2.34 × 2.34 × 2.78
Discovery MI/DR	✓(Q.Clear)			β650	N/A	192 × 192 × 47	3.13 × 3.13 × 3.27
Discovery IQ 5R			✓(Q.Clear)	β300	N/A	192 × 192 × 79	2.60 × 2.60 × 3.26
Discovery IQ 4R			✓(Q.Clear)	β450	N/A	192 × 192 × 63	3.13 × 3.13 × 3.26
Discovery IQ 3R			✓(Q.Clear) ✓(OSEM)	β500 (Q.Clear) it. 3, sub. 12 (PSF)	N/A (Q.Clear) Gauss 5.0 mm (PSF)	192 × 192 × 47	3.65 × 3.65 × 3.26
Discovery 710	✓	✓		it. 4, sub. 12 (TOF + PSF) it. 3, sub. 12 (TOF)	Gauss 4.0 mm	192 × 192 × 47	3.65 × 3.65 × 3.27
Discovery 690	✓	✓		it. 3, sub. 16 (TOF + PSF) it. 3, sub. 8 (TOF)	Gauss 6.0 mm (TOF + PSF) Gauss 4.0 mm (TOF)	192 × 192 × 47	3.13 × 3.13 × 3.27
Discovery 610			✓(Q.Clear)	β300	N/A	192 × 192 × 47	2.60 × 2.60 × 3.27
Vereos	✓	✓		it. 3, sub. 15	6.0 mm FWHM PSF iteration = 1, PSF regularization	144 × 144 × 41	4 × 4 × 4
Ingenuity TF	✓	✓		it. 3, sub. 33	Smooth A (relaxation parameter 0.6)	144 × 144 × 45	4 × 4 × 4
Gemini TF (TR: 495 ps)		✓		it. 3, sub. 33	Smooth A (relaxation parameter 0.6)	144 × 144 × 45	4 × 4 × 4
Gemini TF (TR: 585 ps)		✓		it. 3, sub. 33	Relaxation parameter 0.6	144 × 144 × 45	4 × 4 × 4
uMI510	✓	✓		it. 3, sub. 24 (TOF + PSF) it. 2, sub. 24 (TOF)	Smooth 3, NLM + Gauss 6.0 mm	256 × 256 × 117	2.73 × 2.73 × 2
uMI780	✓	✓		it. 2, sub. 20	NLM + Gauss 4.0 mm	192 × 192 × 101	3.13 × 3.13 × 3

described below, the data were classified into four groups: overall ($n = 37$), TOF + PSF ($n = 17$), TOF ($n = 15$), and PSF ($n = 5$).

Average SUV in the background area ($SUV_{B,ave}$)

To confirm the quantitative accuracy of data, we examined the average SUV in the background area ($SUV_{B,ave}$) on PET images with 1800-s acquisition. Image analysis was performed with the PETquactIE Ver. 3 software (Nihon Medi-Physics Co., Ltd) [62]. On the axial slice of the sphere center, 12 circular regions-of-interest (ROIs) with a 37-mm diameter were placed over the background area [63]. The ROIs were also placed on the slices ± 1 and ± 2 cm away from the central slice (60 ROIs in total). The $SUV_{B,ave}$ was calculated by the following equation:

$$\mathrm{SUV}_{\mathrm{B,ave}} = \frac{\sum_{k=1}^{K} \mathrm{SUV}_{\mathrm{B,37\,mm,k}}}{K}, \quad (2)$$

where $SUV_{B,37\,mm}$ is the average SUV for the 37-mm ROIs and K is the number of ROIs, that is 60. An acceptable range of the $SUV_{B,ave}$ was defined as 0.95–1.05. When the $SUV_{B,ave}$ did not meet this acceptable range, re-testing was done after cross calibration and, if necessary, scanner maintenance.

Part I: image quality with a focus on 10 mm sphere detectability

Visual detectability score

Detectability of the 10-mm-diameter hot sphere was visually assessed by five nuclear medicine technologists in a 3-step scale (0, not visualized; 1, visualized, but similar hot spots are observed; and 2, identifiable). The VOX-BASE/MANAGER (J-MAC SYSTEM, INC., Japan) was used to display PET images using an inverted gray scale with an upper level of 4 and a lower level of 0 (SUV-scaled). The score was averaged across the three image sets and then averaged across the five raters. A score of 1.5 was defined as an acceptable level (i.e., the 10 mm hot sphere can be detected by half or more of the raters) [29].

$\mathrm{NEC_{phantom}}$

To examine coincidence count data quality, the noise-equivalent count for phantom ($\mathrm{NEC_{phantom}}$) was calculated by the following equations [29, 64, 65]:

$$\mathrm{NEC_{phantom}} = (1-\mathrm{SF})^2 \frac{(T+S)^2}{(T+S)+(1+k)fR}\ [\mathrm{Mcounts}] \quad (3)$$

$$f = \frac{S_a}{\pi r^2}, \quad (4)$$

where SF represents scatter fraction, and T, S, and R are true, scatter and random coincidence counts. $T+S$ was calculated by subtracting estimated random coincidence counts (R) from prompt coincidence counts ($T+S+R$). k is a random scaling factor, depending on the random correction method used [66]. We simply set $k=1$ for a delayed coincidence-based method, and $k=0$ for a singles-based method. f is the ratio of object size to the transaxial field-of-view, S_a is the cross-sectional area of the phantom, and r is the radius of the detector ring. The scatter fraction (SF) for each scanner, according to NEMA NU2 standards, is shown in Supplemental Table 1. The SF values were obtained from previous publications or scanner specification sheets or measured at the clinical site.

Image quality [10-mm-sphere contrast ($Q_{\mathrm{H,10\,mm}}$), background variability ($N_{\mathrm{10\,mm}}$), and image noise level ($\mathrm{CV_{BG}}$)]

For image quality assessment, we evaluated the contrast for the 10 mm hot sphere, background variability and image noise level in the background area using the PETquactIE Ver.3 software [62]. On the axial slice of the sphere center, we placed a circular ROI on the 10 mm sphere. In addition, we placed twelve 10-mm-diameter circular ROIs on the background area on the slice of the sphere center and on slices ± 1 cm and ± 2 cm away from the central slice (60 ROIs in total). The percent contrast for the 10 mm hot sphere ($Q_{\mathrm{H,10\,mm}}$) was calculated as follows:

$$Q_{\mathrm{H,10\,mm}} = \frac{C_{\mathrm{H,10\,mm}}/C_{\mathrm{B,10\,mm}} - 1}{a_{\mathrm{H}}/a_{\mathrm{B}} - 1} \times 100\ (\%), \quad (5)$$

where $C_{\mathrm{H,10\,mm}}$ and $C_{\mathrm{B,10\,mm}}$ are the average activity in the ROI for the 10 mm sphere and the average activity in all the background 10-mm-diameter ROIs, respectively. $a_{\mathrm{H}}/a_{\mathrm{B}}$ is the activity concentration ratio between the hot spheres and the background. The percent background variability ($N_{\mathrm{10\,mm}}$) for the 10 mm circular ROIs was calculated as follows:

$$N_{\mathrm{10\,mm}} = \frac{\mathrm{SD_{10\,mm}}}{C_{\mathrm{B,10\,mm}}} \times 100\ (\%) \quad (6)$$

$$\mathrm{SD_{10\,mm}} = \sqrt{\frac{\sum_{k=1}^{K}\left(C_{\mathrm{b,10\,mm,k}} - C_{\mathrm{B,10\,mm}}\right)^2}{K-1}}, K = 60, \quad (7)$$

where $\mathrm{SD_{10\,mm}}$ is the standard deviation of the mean activity for the background 60 ROIs. For image noise assessment, we placed 37-mm-diameter circular ROIs on the background area in the same manner as for the background variability assessment (60 ROIs). The coefficient of variation on the background area ($\mathrm{CV_{BG}}$) (image noise levels) was calculated by the following equation:

$$\mathrm{CV_{BG}} = \text{mean of}\left(\frac{\mathrm{SD_{37\,mm}}}{C_{\mathrm{B,37\,mm}}} \times 100\right)[\%], [n = 60], \quad (8)$$

where $\mathrm{SD_{37\,mm}}$ and $C_{\mathrm{B,37\,mm}}$ are the standard deviation and average of the activity in each 37-mm-diameter ROI, respectively. The $Q_{\mathrm{H,10\,mm}}$, $N_{\mathrm{10\,mm}}$ and $\mathrm{CV_{BG}}$ were measured and averaged by five nuclear medicine technologists.

Investigation of image quality reference levels allowing the 10 mm sphere to be visible

The relationships between each image quality metric and visual detectability score were examined to explore an appropriate image quality level for 10 mm sphere detection. The $\mathrm{NEC_{phantom}}$, $Q_{\mathrm{H,10\,mm}}$, $N_{\mathrm{10\,mm}}$, $Q_{\mathrm{H,10\,mm}}/N_{\mathrm{10\,mm}}$, $\mathrm{CV_{BG}}$, and visual detectability score are shown as a function of acquisition duration (30–300 s). As mentioned earlier, a visual detectability score of 1.5 was defined as an acceptable level. Figure 3 shows the workflow to determine a reference level for each image quality metric. For

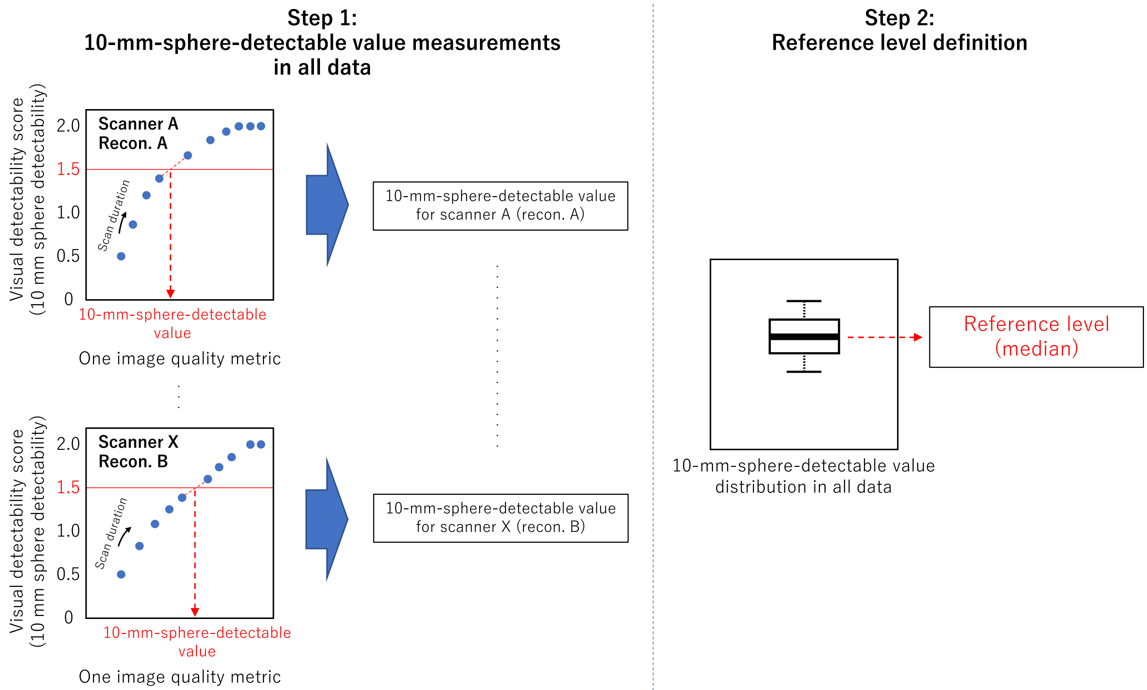

Fig. 3 The two-step workflow to determine a reference level for each image quality metric

each image quality metric and each dataset, we measured a 10-mm-sphere-detectable value so as to achieve the visual detectability score of 1.5 (Fig. 3, step 1). For all data, the acquisition duration corresponding to the visual detectability score of 1.5 was calculated by linear interpolation between the nearest data. If the visual detectability score was higher than 1.5 at the minimum acquisition duration of 30 s, the data with the acquisition duration of 30 s were used as the 10-mm-sphere-detectable value. Subsequently, the reference level for each image quality metric ($NEC_{phantom}$, $N_{10\ mm}$, $Q_{H,10\ mm}/N_{10\ mm}$ and CV_{BG}) was calculated (Fig. 3, step 2). The reference level was defined as the median for all 10-mm-sphere-detectable values.

Inter-rater variability in each image quality metric

To evaluate the inter-rater variability in $Q_{H,10\ mm}$, $N_{10\ mm}$ and CV_{BG}, we calculated the respective coefficient of variation across five raters (*inter-rater variability*) as follows:

$$\text{Inter-rater variability} = \frac{\sigma}{\mu} \times 100\ (\%), \tag{9}$$

where σ and μ are the standard deviation and mean of the measurement values, respectively. To remove the effect of statistical noise, the PET images with 300 s acquisition were used for this evaluation.

Part II: SUV variability

SUVs of hot spheres

On PET images with 1800-s acquisition, SUVmax and SUVpeak for the hot spheres were measured using PETquactIE Ver. 3 and RAVAT, respectively (Nihon Medi-Physics Co., Ltd.) [15, 62]. To measure SUVmax for each sphere, a circular ROI was placed with a diameter equal to the inner diameter of the sphere. To measure SUVpeak for each sphere, a volume-of-interest (VOI) was placed, so that the VOI covered the whole uptake. The SUVpeak was defined as the average value within a 1 mL spherical VOI (12-mm-diameter) that was placed so as to maximize the average SUV [18]. Considering this definition, we did not measure the SUVpeak of the 10-mm sphere. When showing recovery coefficient curves, the SUVs were normalized by the true value of 4.

SUV harmonization range

SUVs of the hot spheres among all images with 1800-s acquisition ($n = 37$) were investigated for all-size spheres. To investigate feasible lower and upper limits, 0–30th percentiles and 70th–100th percentiles were calculated in a fifth percentile step. On PET images with PSF reconstruction, the SUVs of 13–22 mm spheres were often overestimated by edge artifact [67, 68]. Here, the maximum overshoot rate in SUVs (MOR) was calculated by the following equation:

$$\mathrm{MOR} = \frac{\mathrm{SUV}_i - \mathrm{SUV}_{37\ \mathrm{mm}}}{\mathrm{SUV}_{37\ \mathrm{mm}}} \times 100(\%), \tag{10}$$

where SUV_i is the SUV of the i-mm diameter sphere that shows the highest SUV among 13–22 mm spheres, and $\mathrm{SUV}_{37\ \mathrm{mm}}$ is the SUV of the 37-mm-diameter sphere. Based on these data, we investigated a feasible SUV harmonization range. The upper limit was determined, so that the MOR was lower than 5%. For the lower limit, we considered that it should be lower than the true SUV of 4 for all spheres.

Relationships between SUVs of hot spheres and image noise levels ($\mathrm{CV}_{\mathrm{BG}}$)

On PET images with 30–300 s acquisition, we investigated relationships between SUVs of the hot spheres and image noise levels. In this evaluation, SUVmax of the hot spheres was measured using spherical VOIs that sufficiently covered the whole uptake, assuming realistic tumor uptake measurements. Each SUV of the hot spheres on PET images with 1800-s acquisition was defined as a reference, because the images were in low noise conditions. Then, on PET images with 30–300 s acquisition, relative differences of SUVs were plotted as a function of $\mathrm{CV}_{\mathrm{BG}}$. The measurement procedure of the $\mathrm{CV}_{\mathrm{BG}}$ was described above (Eq. 8). The relative differences of SUVs ($\mathrm{RD}_{\mathrm{SUV}}$) were calculated by the following equation:

$$\mathrm{RD}_{\mathrm{SUV}} = \frac{\mathrm{SUV}_i - \mathrm{SUV}_{i,\mathrm{ref}}}{\mathrm{SUV}_{i,\mathrm{ref}}} \times 100(\%), \tag{11}$$

where SUV_i is the SUV of the i-mm diameter sphere on each PET image and $\mathrm{SUV}_{i,\mathrm{ref}}$ is the SUV of the i-mm-diameter sphere on PET images with 1800-s acquisition. The $\mathrm{RD}_{\mathrm{SUV}}$ was calculated for SUVmax and SUVpeak. To investigate the effect of the uptake volume, the $\mathrm{RD}_{\mathrm{SUV}}$ values were classified into two groups based on the sphere diameter (diameter: < 20 mm and ≥ 20 mm). This was based on the recommendation by the QIBA and PERCIST that the minimum lesion size was 2 cm in diameter for the target lesion at the baseline [8, 18].

Statistical analysis

All statistical analyses were performed with EZR (Saitama Medical Center, Jichi Medical University, Saitama, Japan) [69], which is a graphical user interface for R (The R Foundation for Statistical Computing, Vienna, Austria). Comparisons of values between two groups were performed with the Mann–Whitney U test. Comparisons of values among three or more groups were performed using the Kruskal–Wallis test, followed by the Steel–Dwass pair-wise multiple comparison test. Spearman's correlation test was used to investigate the correlation of each image quality metric with the visual detectability score. Correlations between $\mathrm{RD}_{\mathrm{SUV}}$ and $\mathrm{CV}_{\mathrm{BG}}$ were examined with Pearson's correlation test. In all analyses, $P < 0.05$ was defined as statistically significant.

Results

Average SUV in the background area ($\mathrm{SUV}_{\mathrm{B,ave}}$)

The mean ± SD of the $\mathrm{SUV}_{\mathrm{B,ave}}$ was 1.00 ± 0.03 and all values were within 0.95–1.05. Supplemental Fig. 1 shows $\mathrm{SUV}_{\mathrm{B,ave}}$ for all scanner models. There was no significant difference among reconstruction algorithms ($P = 0.56$).

Part I: image quality

Figure 4 shows PET images with 120-s acquisition, which were reconstructed with clinical settings. There were no artifacts in any images, but large differences were found in visual contrasts of the smallest 10 mm sphere among scanners. Figure 5 shows $\mathrm{NEC}_{\mathrm{phantom}}$, $Q_{\mathrm{H,10\ mm}}$, $N_{10\ \mathrm{mm}}$, $Q_{\mathrm{H,10\ mm}}/N_{10\ \mathrm{mm}}$, $\mathrm{CV}_{\mathrm{BG}}$ and visual detectability score as a function of scan duration. The $\mathrm{NEC}_{\mathrm{phantom}}$, $Q_{\mathrm{H,10\ mm}}/N_{10\ \mathrm{mm}}$, and visual detectability score increased with acquisition duration, while $N_{10\ \mathrm{mm}}$ and $\mathrm{CV}_{\mathrm{BG}}$ decreased with it. The $Q_{\mathrm{H,10\ mm}}$ did not correlate with acquisition duration.

Figure 6 shows distributions of 10-mm-sphere-detectable values (i.e., corresponding to visual detectability score = 1.5) for $\mathrm{NEC}_{\mathrm{phantom}}$, $N_{10\ \mathrm{mm}}$, $Q_{\mathrm{H,10\ mm}}/N_{10\ \mathrm{mm}}$ and $\mathrm{CV}_{\mathrm{BG}}$. The data were classified into four groups by image reconstruction methods as follows: Overall ($n = 37$), TOF + PSF ($n = 17$), TOF ($n = 15$), and PSF ($n = 5$). The medians [min, max] of the 10-mm-sphere-detectable values were 3.2 [0.5, 6.8] for $\mathrm{NEC}_{\mathrm{phantom}}$, 10.6 [7.3, 19.6] for $N_{10\ \mathrm{mm}}$, 2.5 [0.3, 3.5] for $Q_{\mathrm{H,10\ mm}}/N_{10\ \mathrm{mm}}$, and 14.1% [8.8, 33.5] for $\mathrm{CV}_{\mathrm{BG}}$. For $\mathrm{NEC}_{\mathrm{phantom}}$ and $N_{10\ \mathrm{mm}}$, significant differences were observed in the 10-mm-sphere-detectable values among the three groups. For more detailed information, the relationships between each image quality metric and visual detectability score are shown in the supplemental data (Supplemental Figs. 2–5). Each image quality metric was significantly correlated with the visual detectability score ($P < 0.001$) (Supplemental Table 2).

Medians [min, max] of the inter-rater variability in $Q_{\mathrm{H,10\ mm}}$, $N_{10\ \mathrm{mm}}$ and $\mathrm{CV}_{\mathrm{BG}}$ were 4.0 [1.0, 9.4], 5.6 [2.1, 13.3], and 0.8 [0.3, 5.6], respectively (Fig. 7). Inter-rater variability was significantly lower for $\mathrm{CV}_{\mathrm{BG}}$ compared to $Q_{\mathrm{H,10\ mm}}$ and $N_{10\ \mathrm{mm}}$ ($P < 0.001$).

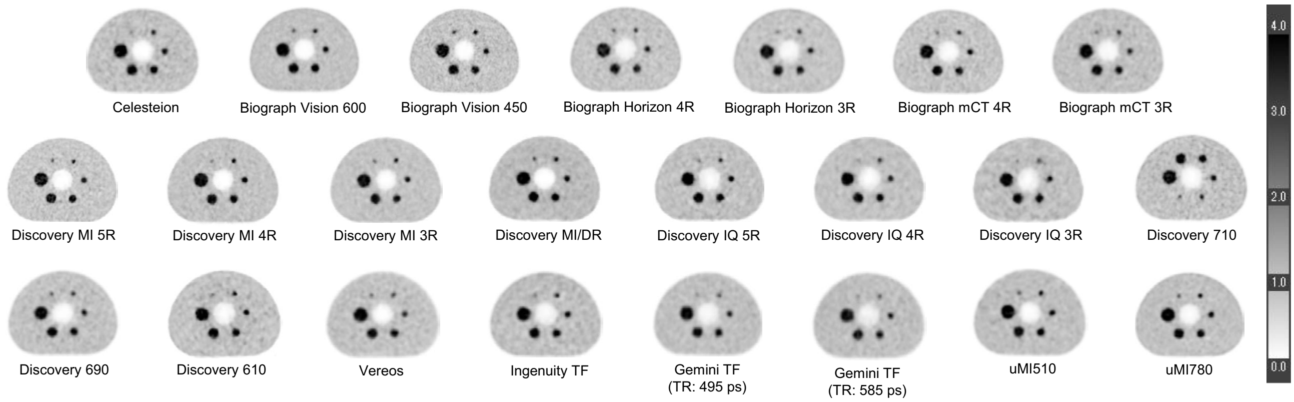

Fig. 4 PET images obtained with 120-s acquisition, which were reconstructed with the clinical settings at each site. For the scanners with PSF reconstruction, the PET images reconstructed with PSF modeling are shown. They are displayed with an upper level of SUV = 4, which equals the activity concentration of the hot spheres, and a lower level of SUV = 0

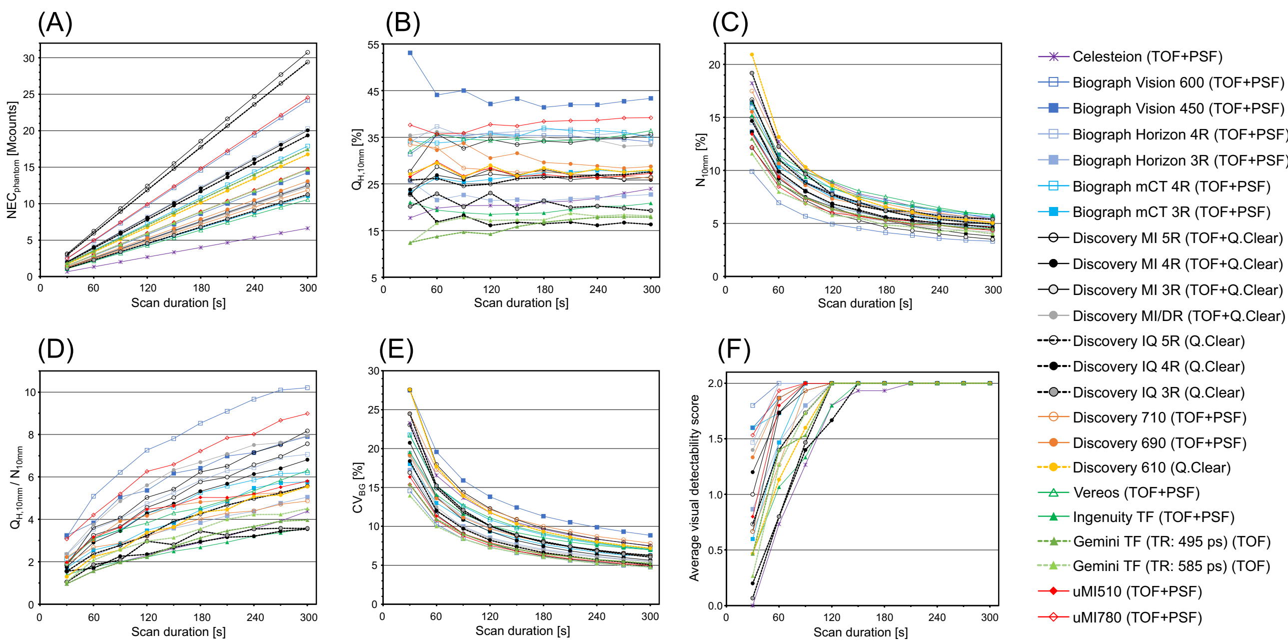

Fig. 5 **A** $\mathrm{NEC}_{\mathrm{phantom}}$, **B** $Q_{\mathrm{H,10\,mm}}$, **C** $N_{\mathrm{10\,mm}}$, **D** $Q_{\mathrm{H,10\,mm}}/N_{\mathrm{10\,mm}}$, **E** $\mathrm{CV}_{\mathrm{BG}}$, and **F** visual detectability score as a function of scan duration

Part II: SUV variability

Figure 8 shows recovery coefficients for SUVmax and SUVpeak on PET images with 1800-s acquisition. A large variability was observed especially for the 13 mm sphere. Table 2 summarizes median, minimum, and maximum values of SUVmax and SUVpeak on PET images with 1800-s acquisition. For the small spheres (10–17 mm diameter spheres), the inter-scanner variability in SUVpeak was smaller than that in SUVmax.

The mean ± SD and various (0–30th and 70th–100th) percentile values for SUVmax and SUVpeak of all spheres are shown in Table 3 for PET images with 1800-s acquisition. The MOR for each upper range of 70th–100th percentiles is also given in that table. Using the 100th percentile, we obtained MORs for SUVmax and SUVpeak of 11.0% and 2.3%, respectively.

The MOR for SUVmax was lower than 5% when using ≤ 90th percentile values as the upper limit (Table 3). Therefore, the 90th percentile values were defined as the upper limit for the SUV harmonization range (Fig. 9). Then, the 10th percentile values were defined as the lower limit. This was selected, because the lower limit for all spheres was lower than the true SUV of 4, and the exclusion rate was the same as the upper limit.

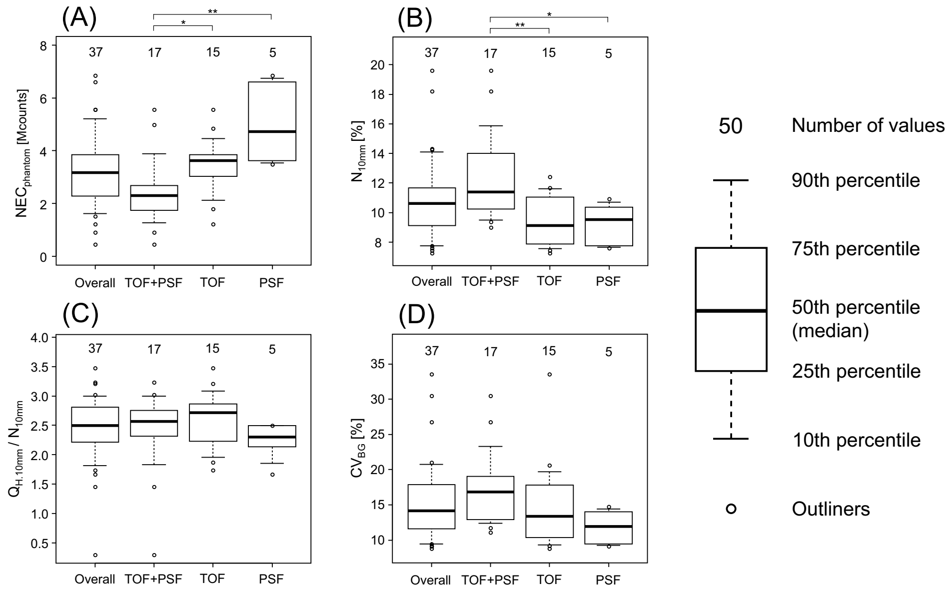

Fig. 6 Box plots of 10-mm-sphere-detectable values (i.e., corresponding visual detectability score = 1.5) for **A** $NEC_{phantom}$, **B** $N_{10\ mm}$, **C** $Q_{H,10\ mm}/N_{10\ mm}$ and **D** CV_{BG}. The data were classified into four groups by image reconstruction algorithms. The midline indicates the median, the box indicates the first and third quartiles of the distribution, whiskers indicate the 10% and 90% values, and circles represent outliers. * Indicates $P < 0.05$ and ** indicates $P < 0.01$

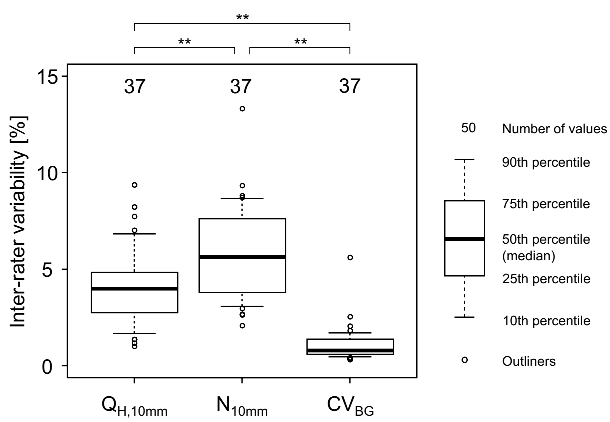

Fig. 7 Box plots of inter-rater variability for $Q_{H,10\ mm}$, $N_{10\ mm}$ and CV_{BG}. The midline indicates the median, the box indicates the first and third quartiles of the distribution, whiskers indicate the 10% and 90% values, and circles represent outliers. ** Indicates $P < 0.01$

For SUVmax and SUVpeak for the hot spheres on PET images with 30–300 s acquisition, RD_{SUV} in relation to CV_{BG} are shown in Fig. 10. In SUVmax for the small spheres (10–17 mm diameter), a positive bias was observed in RD_{SUV}. Table 4 shows median, minimum, and maximum values for the RD_{SUV}. The median [min, max] of the RD_{SUV} for SUVmax and SUVpeak in all spheres were 5.3% [− 30.6%, 340.7%] and 1.1% [− 17.8%, 49.8%], respectively. There was a significant difference in the RD_{SUV} between SUVmax and SUVpeak ($P < 0.001$). The RD_{SUV} for both the SUVmax and SUVpeak significantly depended on sphere diameter (< 20 mm and ≥ 20 mm) and CV_{BG} (≤ 10% and > 10%) ($P < 0.001$).

Discussion

We investigated image quality and SUV variability in hot spheres using 23 recent PET scanner models. Since almost all recent PET/CT scanner models were included in this study, the data precisely reflect current PET image characteristics available at clinical sites. Based on the data, we have proposed a reference level for each image quality metric ($NEC_{phantom}$, $N_{10\ mm}$, $Q_{H,10\ mm}/N_{10\ mm}$ and CV_{BG}) with a focus on 10 mm sphere detectability. In addition, we have proposed a new SUV harmonization range and image noise criterion with a focus on the inter-scanner and intra-scanner SUV variabilities. Our proposed new standards will be useful for image quality standardization and SUV harmonization of PET studies in oncology.

Part I: image quality

Figures 4 and 5 show PET images and image quality metrics under clinical image reconstruction conditions. Because standardization of PET image quality was not performed, there was a large difference in 10-mm-sphere contrasts among scanners. As theoretically expected, longer scan durations provided lower image noise levels and better visual detectability scores. The results indicate that a simple way to obtain better image quality is to extend scan duration [70]. Looking at the 180-s scan data, which is the standard scan duration recommended by the JSNM [27], almost all scanners achieved the visual detectability score of 2.0 (Fig. 5). Therefore, a 180-s scan for each bed position would be reasonable as a reference standard.

For each image quality metric, we have proposed a reference level that makes the 10 mm sphere visible. The

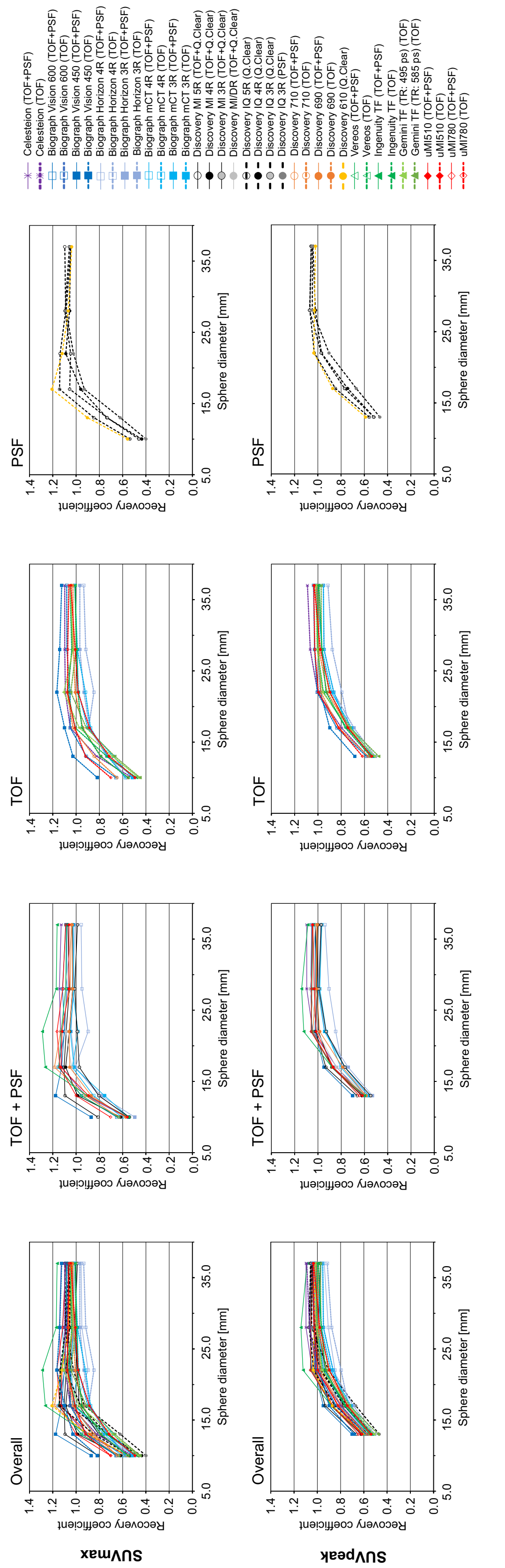

Fig. 8 Recovery coefficients of SUVmax and SUVpeak for four groups: Overall, TOF + PSF, TOF, and PSF

Table 2 Median, minimum, and maximum values of SUVmax, SUVpeak, and CV_{BG} in 1800-s PET images

Metric	Sphere diameter	Median [min, max]
SUVmax	10 mm	2.2 [1.6, 3.5]
	13 mm	3.4 [2.5, 4.7]
	17 mm	4.1 [3.5, 5.1]
	22 mm	4.3 [3.4, 5.2]
	28 mm	4.2 [3.7, 4.7]
	37 mm	4.2 [3.7, 4.7]
SUVpeak	13 mm	2.3 [1.9, 2.8]
	17 mm	3.2 [2.7, 3.8]
	22 mm	3.9 [3.2, 4.5]
	28 mm	4.0 [3.5, 4.6]
	37 mm	4.1 [3.7, 4.4]
CV_{BG}		2.8% [2.1%, 7.2%]

calculation procedure for the reference level (Fig. 3) was the same as that of the previous work in 2014 [29], in which the reference levels were proposed as follows: $NEC_{phantom} > 10.8$ Mcounts, $N_{10\ mm} < 5.6\%$, $Q_{H,10\ mm}/N_{10\ mm} > 2.8$. On the other hand, we have provided reference levels as follows: $NEC_{phantom} \geq 3.2$ Mcounts, $N_{10\ mm} \leq 10.6\%$, $Q_{H,10\ mm}/N_{10\ mm} \geq 2.5$, $CV_{BG} \leq 14.1\%$. The CV_{BG} has been newly added to the image quality metrics.

The proposed new reference level for the $NEC_{phantom}$ was lower than that in the 2014 study [29]. This result suggests that recent PET scanners can visualize the 10 mm sphere even with a low $NEC_{phantom}$ value. This is mainly because significant progress has been made in developing image reconstruction algorithms. The NEC is a count-based metric, and independent of image reconstruction algorithms [65]. Because PET image quality is determined by detected coincidence count quality (e.g., NEC), image reconstruction algorithms, and so on (Fig. 1), the $NEC_{phantom}$ would not be suitable for the use for image quality standardization [71, 72].

The $N_{10\ mm}$, which is a metric of background variability, had similar results to those of $NEC_{phantom}$. The proposed reference level for the $N_{10\ mm}$ was higher than that in the previous study. This is also probably due to advances in image reconstruction algorithm. Specifically, PSF and TOF would contribute mainly to improving contrast for the 10 mm sphere [73]. These new techniques allow recent PET scanners to visualize the 10 mm sphere even with higher background variability. In addition, smaller voxel sizes were used in this study (1.3–4.1 mm) compared with those in the previous study (3.1–5.3 mm) [29]. Higher background variability might be derived from smaller voxel size.

On the other hand, the reference level for $Q_{H,10\ mm}/N_{10\ mm}$ (contrast-to-noise ratio) was almost the same as that in the previous study. In addition, there was no significant

Table 3 Mean ± SD and various percentile values for SUVmax and SUVpeak for all spheres

Metric	Sphere diameter	Mean ± SD	Percentile 0	5th	10th	15th	20th	25th	30th	~	70th	75th	80th	85th	90th	95th	100th
SUVmax	10 mm	2.3 ± 0.4	1.6	1.8	1.8	1.9	2.0	2.1	2.1		2.5	2.6	2.6	2.6	2.8	3.3	3.5
	13 mm	3.4 ± 0.5	2.5	2.7	2.8	2.9	2.9	2.9	3.0		3.7	3.7	3.7	3.8	4.0	4.2	4.7
	17 mm	4.1 ± 0.4	3.5	3.5	3.6	3.7	3.8	3.8	3.9		4.4	4.5	4.5	4.6	4.6	4.8	5.1
	22 mm	4.2 ± 0.3	3.4	3.7	3.8	3.9	3.9	4.0	4.2		4.4	4.4	4.5	4.5	4.5	4.7	5.2
	28 mm	4.2 ± 0.2	3.7	3.7	4.0	4.0	4.0	4.1	4.1		4.3	4.4	4.4	4.4	4.5	4.6	4.7
	37 mm	4.2 ± 0.2	3.7	3.9	4.0	4.0	4.0	4.1	4.1		4.3	4.3	4.3	4.4	4.4	4.5	4.7
Maximum overshoot rate	–	–	–	–	–	–	–	–	–		3.4%	4.0%	4.3%	4.8%	4.7%	5.8%	11.0%
SUVpeak	13 mm	2.3 ± 0.2	1.9	2.0	2.1	2.1	2.1	2.1	2.2		2.4	2.4	2.5	2.5	2.5	2.7	2.8
	17 mm	3.2 ± 0.3	2.7	2.8	2.9	2.9	3.0	3.0	3.0		3.4	3.5	3.5	3.5	3.5	3.6	3.8
	22 mm	3.9 ± 0.3	3.2	3.4	3.5	3.6	3.6	3.7	3.7		4.0	4.1	4.1	4.2	4.2	4.2	4.5
	28 mm	4.0 ± 0.2	3.5	3.8	3.8	3.9	3.9	4.0	4.0		4.1	4.2	4.2	4.3	4.3	4.3	4.6
	37 mm	4.1 ± 0.2	3.7	3.8	3.8	3.9	3.9	4.0	4.0		4.2	4.2	4.2	4.2	4.2	4.3	4.4
Maximum overshoot rate	–	–	–	–	–	–	–	–	–		–	–	–	–	–	–	2.3%

difference in the 10-mm-sphere-detectable values for $Q_{H,10\ mm}/N_{10\ mm}$ among the image reconstruction algorithms (Fig. 6). These results suggest that the $Q_{H,10\ mm}/N_{10\ mm}$ would be a useful metric for assuring 10 mm sphere visibility, irrespective of PET scanner models and image reconstruction algorithms. The $Q_{H,10\ mm}/N_{10\ mm}$ includes information on both the 10 mm-sphere-contrast and background variability, and the balance of contrast and noise might be a key component for visual detectability of small hot lesions.

As for the CV_{BG} (image noise levels), there was no significant difference in the 10-mm-sphere-detectable values among image reconstruction algorithms (Fig. 6). Additionally, the CV_{BG} has some advantages compared with other metrics. The CV_{BG} showed the lowest inter-rater variability among all image quality metrics (Fig. 7). The reason for its low variability is that the large 37 mm ROIs were used to measure the CV_{BG} (10 mm ROIs were used for $Q_{H,10\ mm}$ and $N_{10\ mm}$ measurements). The CV_{BG} is therefore more reproducible than $Q_{H,10\ mm}$ and $N_{10\ mm}$ are. Furthermore, the CV_{BG} has been widely used for standardization of FDG PET in oncology. RSNA/QIBA and EANM/EARL specify that image noise levels are assessed by measuring the CV in the uniform background area as part of their standardization strategies [18, 22]. They have provided an acceptable level of 15% that is close to our proposed reference level (14.1%), although the phantom and ROI conditions are somewhat different. The CV_{BG} and its reference level are compatible with other international standards. The use of CV_{BG} may facilitate international standardization and global PET studies. What should be taken account for the CV_{BG} is not considering the image contrast. Not only the CV_{BG} also other image contrast-related metrics such as $Q_{H,10\ mm}/N_{10\ mm}$ and recovery coefficients [29] should be evaluated to assure small lesion detectability.

Part II: SUV variability

As shown in Supplemental Fig. 1, the $SUV_{B,ave}$ of all scanner models were within 0.95–1.05. This result indicated that all scanners were well calibrated, and their quantitative accuracy was within ± 5% error. Therefore, our phantom data are sufficiently reliable to establish an SUV harmonization range. In the previous report on 2013, the $SUV_{B,ave}$ of 16 scanners were distributed from 0.87 to 1.14 [74]. Quantitative accuracy of PET scanners would have been improved by scanner performance progress. As described in the Materials and methods section, we visited several sites and supported the phantom test when requested. Such support might be effective in minimizing any technical errors in the process of phantom preparation.

Subsequently, we investigated inter-scanner SUV variability in each sphere on PET images with 1800-s acquisition (in noise-less conditions). Most scanner models showed

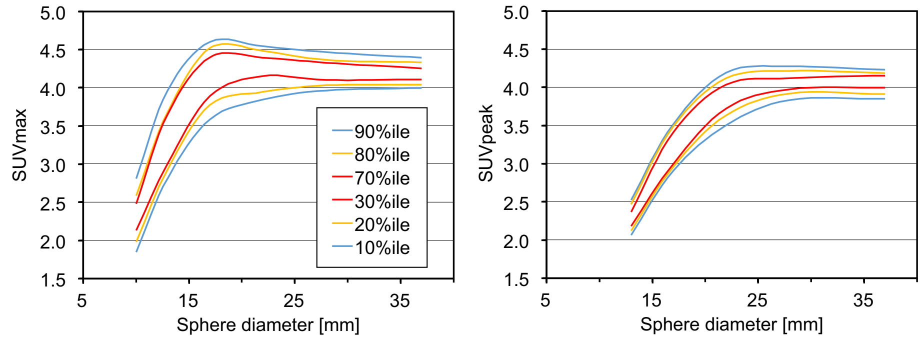

Fig. 9 Representative percentiles in SUVmax (left) and SUVpeak (right). The range of the 10th-to-90th percentiles was proposed as the SUV harmonization range

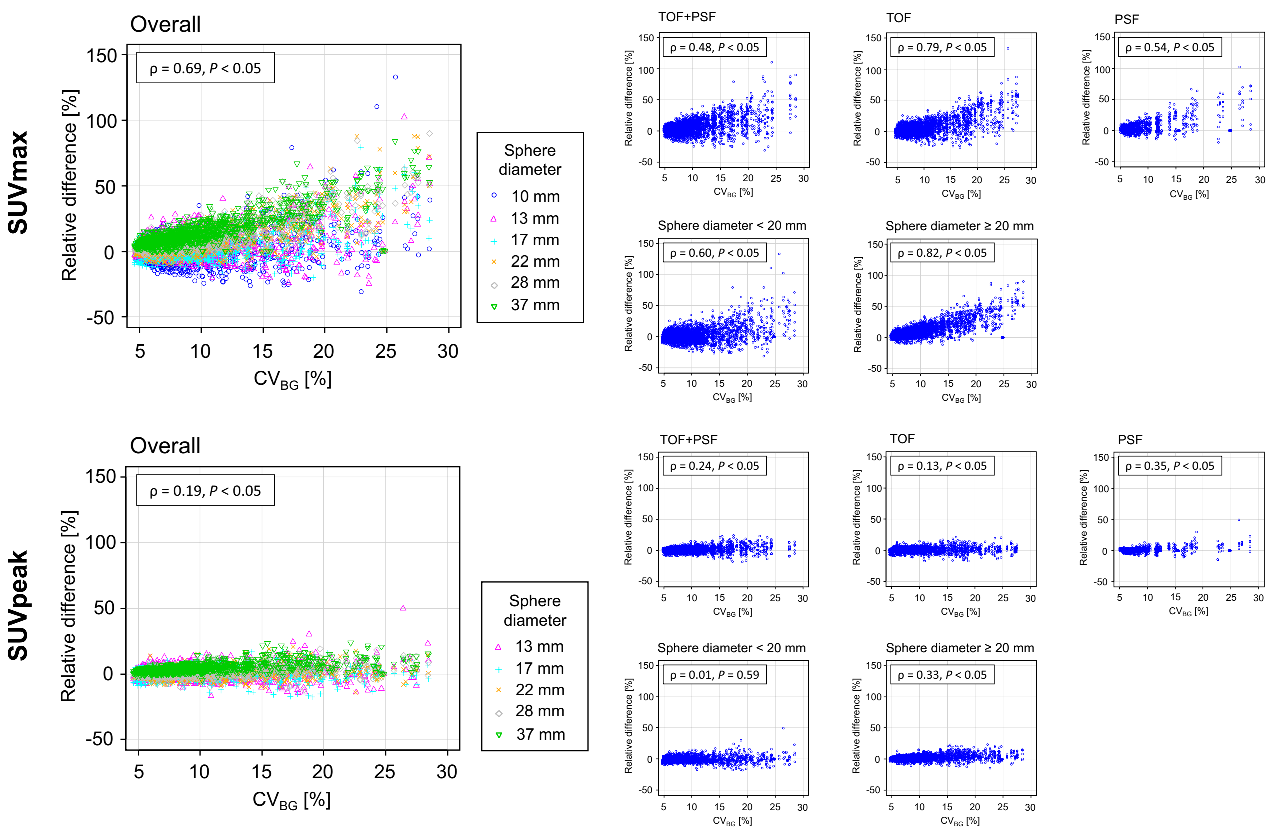

Fig. 10 Scatter plots of relative differences for SUVmax (upper) and SUVpeak (lower). Each reference was a corresponding SUV for an 1800-s PET image. Five different categorizations (TOF + PSF, TOF, PSF, sphere diameter < 20 mm and ≥ 20 mm) are shown on the right

higher SUVmax recovery coefficients than their upper limit provided by JSNM (Supplemental Fig. 6). This result suggested that the SUV harmonization range should be regularly updated according to the performance improvement of commercial scanners [12]. In comparison to the large spheres (28–37 mm diameters), the small spheres (10–22 mm diameters) had larger SUV variability (Fig. 8). Many studies have reported that TOF PET scanners provided higher SUVs for small lesions compared with those without TOF [26, 75, 76]. Since this study used both TOF and non-TOF scanner models (19 TOF PET scanner models and 4 non-TOF PET scanner models), the SUV variability in the small spheres would result in large variability.

Comparing TOF + PSF and TOF groups (Fig. 8), higher SUVs were obtained for the 17-mm sphere when using PSF reconstruction. Furthermore, in most cases, SUVmax of the 17-mm sphere was higher than that of the 37-mm sphere. This overshoot would be derived from the edge artifact [31, 67, 68]. If we use the SUVmax of a small lesion on PSF-based PET images for monitoring treatment response, this overshoot must be suppressed by SUV harmonization [77]. For SUVpeak, on the other hand, the

Table 4 Median, minimum, and maximum values for the RD_{SUV} with various categorizations

Metric	*n*	Sphere diameter	Categorization	Median [min, max] of the RD_{SUV} (%)	*P* value
SUVmax	6660	All	–	5.3 [− 30.6, 340.7]	< 0.001
SUVpeak	5550	All	–	1.1 [− 17.8, 49.8]	
SUVmax	1110	10 mm	–	0.5 [− 30.6, 169.4]	< 0.001
	1110	13 mm	–	2.0 [− 24.7, 165.2]	
	1110	17 mm	–	1.2 [− 19.9, 249.0]	
	1110	22 mm	–	6.0 [− 11.3, 257.4]	
	1110	28 mm	–	7.7 [− 7.9, 340.7]	
	1110	37 mm	–	11.3 [0.0, 279.1]	
	3330	All	Sphere diameter < 20 mm	1.2 [− 30.6, 249.0]	< 0.001
	3330	All	Sphere diameter ≥ 20 mm	8.6 [− 11.3, 340.7]	
	4224	All	CV_{BG} ≤ 10%	3.6 [− 22.3, 35.3]	< 0.001
	2436	All	CV_{BG} > 10%	12.4 [− 30.6, 340.7]	
SUVpeak	1110	13 mm	–	0.0 [− 16.9, 49.8]	< 0.001
	1110	17 mm	–	− 0.3 [− 17.8, 17.0]	
	1110	22 mm	–	0.4 [− 14.2, 20.9]	
	1110	28 mm	–	1.2 [− 12.2, 19.2]	
	1110	37 mm	–	3.6 [− 4.9, 23.7]	
	2220	All	Sphere diameter < 20 mm	− 0.1 [− 17.8, 49.8]	< 0.001
	3330	All	Sphere diameter ≥ 20 mm	1.9 [− 14.2, 23.7]	
	3520	All	CV_{BG} ≤ 10%	0.8 [− 10.8, 15.4]	< 0.001
	2030	All	CV_{BG} > 10%	2.0 [− 17.8, 49.8]	

overshoot was suppressed even in PSF-based PET images, and the inter-scanner variability was lower than that for SUVmax.

Based on various percentile values for SUVmax and SUVpeak of all spheres, we proposed a new SUV harmonization range (Fig. 9, 10th–90th percentile). To address the overshoot due to PSF reconstruction [77], we determined the upper limit, so that the MOR was lower than 5% (Table 3). By satisfying our proposed harmonization range, PET images can be used for both lesion detection and quantification even if PSF reconstruction is applied; and feasible and practical SUV harmonization is possible using this harmonization range. Compared with the SUV recovery coefficients for EANM/EARL standards 2 [22, 78], our proposed SUVmax harmonization range is lower (Supplemental Table 3). This is probably due to differences in the phantom test conditions. Because of the low activity concentration, the short scan duration, and high sphere-to-background contrast, the EANM/EARL standards 2 provided a higher bandwidth for SUVmax recovery coefficients. Taking the difference in phantom test conditions into consideration, there would be no big differences between the SUV recovery coefficient harmonization ranges. Interestingly, the differences in SUVpeak recovery coefficient ranges were exceedingly small despite the different phantom test conditions. International harmonization may be possible, although further investigations are required.

Then, we investigated intra-scanner SUV variability in relation to image noise levels. For all data ($n = 37$), three images each with the same acquisition time (30–300 s) were reconstructed. The number of images ($n = 1110$) would be adequate to investigate the relationships. For SUVmax, the variability increased as the CV_{BG} increased. Because SUVmax is derived from a single maximum voxel value, its variability depends considerably on image noise levels [44]. For the large spheres (≥ 20 mm diameter), a positive bias was clearly observed ($\rho = 0.82$). This noise-dependent bias was also reported by Lodge et al. [43]. On the other hand, for the small spheres (< 20 mm diameter), the positive bias was weaker ($\rho = 0.60$) and the numbers of negative values were increased (Fig. 10). When measuring a sequential percentage change in SUVs between two time points, the variability may be large for small lesions. Low image noise is essential for accurate quantitative evaluation, especially for small lesions.

As shown in Table 4, the RD_{SUV} values for SUVmax were distributed from − 30.6 to 340.7% on the PET images with CV_{BG} of higher than 10%. Meanwhile, on the PET images with CV_{BG} of 10% or lower, the RD_{SUV} were distributed from − 22.3 to 35.3%. In the QIBA/UPICT, the CV in the uniform area should be lower than 15% as a target level, and ideally, it should be lower than 10% [18, 79]. The SNMMI/CTN also uses CV in the uniform area as an image noise metric, and it is recommended that CV be 10% or lower

[80, 81]. Akamatsu et al. [44] examined the relationships between image noise levels and SUVs using a phantom and a single PET scanner, and suggested the CV in the uniform area should be below 10% to minimize the SUVmax fluctuation. Considering the results in this study and the standards set by the major nuclear medicine societies, $CV_{BG} \leq 10\%$ would be reasonable and feasible as the image noise criterion.

Comparison of SUVmax and SUVpeak showed that each has its own advantages and disadvantages. SUVmax has been most commonly used to measure lesion uptakes in FDG PET, because its measurement is easy and observer-independent [8, 13]. The partial volume effect is relatively small even in small lesions [82]. Furthermore, SUVmax reflects the highest metabolically active area inside potentially heterogeneous tumors. This is important, because the highest metabolic activity might be critical information clinically. The most challenging issue is the variability in SUVmax (Figs. 8 and 9). Because the inter-scanner and intra-scanner variabilities in SUVmax are problematic, SUV harmonization and image noise management are essential in multicenter studies. In contrast to SUVmax, SUVpeak has lower intra-scanner variability (Fig. 10). SUVpeak was less sensitive to image noise levels than SUVmax. On the PET images with CV_{BG} of 10% or lower, the RD_{SUV} values for SUVpeak were distributed from − 10.8 to 15.4%. Makris et al. [25] also reported that the SUVpeak was less sensitive to variability in image characteristics and might be less affected by noise-dependent bias in comparison to SUVmax. Since SUVpeak may provide lower inter-scanner and intra-scanner variabilities than SUVmax, it is more suitable for use in multicenter studies. However, there are some considerations if SUVpeak is to be used. Because SUVpeak is derived from the 12-mm-diameter spherical VOI, lesion uptakes might be underestimated due to the partial volume effect, particularly in lesions smaller than 20 mm, and it is not applicable to lesions smaller than 12 mm. In addition, there are various definitions for SUVpeak itself [83] and variability will be introduced depending on the image analysis software. To compare the values derived from multiple software codes, VOI definitions should be verified and standardized among image analysis software codes. The appropriate quantitative measure (SUVmax, SUVpeak, etc.) should be selected according to each study's purpose and the characteristics of the target lesion.

Limitations and future issues

The image quality reference levels that we proposed are not appropriate for all FDG PET studies. We focused on 10-mm-sphere detectability; however, if sub-centimeter lesions are the study target, smaller spheres should be evaluated for more effective standardization. In addition, the NEMA image quality phantom mimics an average human body size. In some cases, such as pediatric studies or studies on overweight patients, phantoms of corresponding size would be suitable. Fukukita et al. [29] evaluated larger size body phantoms, and demonstrated that a longer scan time was required for larger phantoms to keep the 10 mm sphere visual detectability. Appropriate evaluations and quality controls should be made according to the purposes of the individual FDG-PET studies [12].

Regarding FDG distributions, intra-tumoral FDG uptakes are not homogeneous but heterogeneous in some types of tumor [84–86]. SUVmax and SUVpeak reflect only the amount of FDG uptakes in specified regions. Recently, other quantitative measures to characterize lesion FDG uptakes have been used, such as metabolic tumor volumes, total lesion glycolysis, and textural features [86–88]. If these quantitative metrics are being used in multicenter studies, the inter-scanner and intra-scanner variabilities should be verified using an appropriate phantom to move toward harmonization.

Conclusions

We experimentally investigated image quality and SUV variability in hot spheres using 23 recent PET scanner models and the NEMA image quality phantom. Then, we investigated appropriate image quality reference levels, so that a 10 mm sphere is visible. The reference levels were newly proposed as: $Q_{H,10\,mm}/N_{10\,mm} \geq 2.5$ and $CV_{BG} \leq 14.1\%$. CV_{BG} is the most reliable and useful, because it has the lowest inter-rater variability (Fig. 7) and is compatible with other international standards such as RSNA/QIBA and EANM/EARL. In addition, we investigated the inter-scanner and intra-scanner SUV variabilities. The new SUV harmonization range (in which PSF reconstruction is applicable) and the image noise criterion ($CV_{BG} \leq 10\%$) were proposed based on these data. Then, our study results supported that SUVpeak is a useful quantitative metric, because it provided reduced inter-scanner and intra-scanner variabilities compared with SUVmax. International SUV harmonization may be facilitated using SUVpeak, although further investigations are needed.

Our proposed new standards are useful for image quality standardization and SUV harmonization of whole-body FDG PET studies in oncology. The reliability of multicenter PET studies will be improved by satisfying the standards before starting the study. We believe that the new standards will help facilitate research and development of new treatments for cancers.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12149-021-01709-1.

Acknowledgements The authors acknowledge the following colleagues for their kind support regarding the phantom experiments: Masafumi Ban, Ryuji Ikeda, Yuji Kojima, Akihito Kuroki, Takamasa Maeda, Yukito Maeda, Hiroyuki Nishida, Kazuki Nitta, Shinji Ochi, Hiroyoshi Okajima, Koji Osanai, Kazuhiro Otani, Shota Sakimoto, Minoru Sakurai, Takahiro Shiraishi, Yuji Tsutsui, Masaki Uno, Kei Wagatsuma, and Masanori Watanabe. The authors would like to thank Shohei Fukai, Noriaki Miyaji, Kazuki Motegi, and Takuro Umeda for visual assessment of phantom images. The authors appreciate the following two committees of the Japanese Society of Nuclear Medicine for their valuable support: the Expert Committee of Standardization of PET imaging (Members: Hiroshi Ito, Setsu Sakamoto, Tohru Shiga, Keiichi Matsumoto, and Hiroshi Watabe) and the PET Nuclear Medicine Committee (Members: Makoto Hosono, Masayuki Sasaki, Ukihide Tateishi, Kenji Ishii, Kengo Ito, Hiroshi Ito, Terue Okamura, Masami Kawamoto, Yuji Kuge, Ichiei Kuji, Michio Senda, Tadaki Nakahara, Yasuhiro Magata, Keiichi Matsumoto, Koji Murakami, Tsuyoshi Yoshida, and Atsuo Waki).

Funding This study was supported in part by the National Cancer Center Research and Development Fund (2020-J-3), Foundation for Promotion of Cancer Research in Japan, the Japanese Society of Nuclear Medicine (JSNM) Working Group, and JSPS KAKENHI Grant Number JP20K08091.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest. All authors are members of the working group of the JSNM (no payment received).

References

1. Rohren EM, Turkington TG, Coleman RE. Clinical applications of PET in oncology. Radiology. 2004;231:305–32.
2. Delbeke D. Oncological applications of FDG PET imaging: brain tumors, colorectal cancer, lymphoma and melanoma. J Nucl Med. 1999;40:591–603.
3. Fletcher JW, Djulbegovic B, Soares HP, Siegel BA, Lowe VJ, Lyman GH, et al. Recommendations on the use of 18F-FDG PET in oncology. J Nucl Med. 2008;49:480–508.
4. Weber WA. Assessing tumor response to therapy. J Nucl Med. 2009;50:1S-10S.
5. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14:169–86.
6. Meikle SR, Sossi V, Roncali E, Cherry SR, Banati R, Mankoff D, et al. Quantitative PET in the 2020s: a roadmap. Phys Med Biol. 2021;66:06RM01.
7. Young H, Baum R, Cremerius U, Herholz K, Hoekstra O, Lammertsma AA, et al. Measurement of clinical and subclinical tumour response using [18F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. Eur J Cancer. 1999;35:1773–82.
8. 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-S150.
9. Judson I, Scurr M, Gardner K, Barquin E, Marotti M, Collins B, et al. Phase II study of cediranib in patients with advanced gastrointestinal stromal tumors or soft-tissue sarcoma. Clin Cancer Res. 2014;20:3603–12.
10. Yap TA, Arkenau H-T, Camidge DR, George S, Serkova NJ, Gwyther SJ, et al. First-in-human phase I trial of two schedules of OSI-930, a novel multikinase inhibitor, incorporating translational proof-of-mechanism studies. Clin Cancer Res. 2013;19:909–19.
11. Connolly RM, Leal JP, Solnes L, Huang C-Y, Carpenter A, Gaffney K, et al. TBCRC026: phase II trial correlating standardized uptake value with pathologic complete response to pertuzumab and trastuzumab in breast cancer. J Clin Oncol. 2019;37:714–22.
12. Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med. 2009;50(Suppl 1):11S-20S.
13. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. AJR Am J Roentgenol. 2010;195:310–20.
14. Fahey FH, Kinahan PE, Doot RK, Kocak M, Thurston H, Poussaint TY. Variability in PET quantitation within a multicenter consortium. Med Phys. 2010;37:3660–6.
15. Daisaki H, Kitajima K, Nakajo M, Watabe T, Ito K, Sakamoto F, et al. Usefulness of semi-automatic harmonization strategy of standardized uptake values for multicenter PET studies. Sci Rep. 2021;11:8517.
16. Kinahan PE, Doot RK, Wanner-Roybal M, Bidaut LM, Armato SG, Meyer CR, et al. PET/CT assessment of response to therapy: tumor change measurement, truth data, and error. Transl Oncol. 2009;2:223–30.
17. 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;19:184–90.
18. FDG-PET/CT Technical Committee. FDG-PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy, Quantitative Imaging Biomarkers Alliance, Version 1.13, Technically Confirmed Version. QIBA, November 18, 2016. Available from: RSNA.ORG/QIBA.
19. Graham MM, Wahl RL, Hoffman JM, Yap JT, Sunderland JJ, Boellaard R, et al. Summary of the UPICT protocol for 18F-FDG PET/CT imaging in oncology clinical trials. J Nucl Med. 2015;56:955–61.
20. 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;50:1187–93.
21. Christian P. 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. J Nucl Med. 2012;53(Suppl 1):437.
22. Kaalep A, Sera T, Rijnsdorp S, Yaqub M, Talsma A, Lodge MA, et al. Feasibility of state of the art PET/CT systems performance harmonisation. Eur J Nucl Med Mol Imaging. 2018;45:1344–61.
23. Kaalep A, Sera T, Oyen W, Krause BJ, Chiti A, Liu Y, et al. EANM/EARL FDG-PET/CT accreditation—summary results from the first 200 accredited imaging systems. Eur J Nucl Med Mol Imaging. 2018;45:412–22.
24. Kinahan PE, Perlman ES, Sunderland JJ, Subramaniam R, Wollenweber SD, Turkington TG, et al. The QIBA profile for FDG PET/CT as an imaging biomarker measuring response to cancer therapy. Radiology. 2020;294:647–57.
25. Makris NE, Huisman MC, Kinahan PE, Lammertsma AA, Boellaard R. Evaluation of strategies towards harmonization of FDG PET/CT studies in multicentre trials: comparison of scanner validation phantoms and data analysis procedures. Eur J Nucl Med Mol Imaging. 2013;40:1507–15.
26. Sunderland JJ, Christian PE. Quantitative PET/CT scanner performance characterization based upon the society of nuclear medicine and molecular imaging clinical trials network oncology clinical simulator phantom. J Nucl Med. 2015;56:145–52.

27. Japanese Society of Nuclear Medicine. Standard PET imaging protocols and phantom test procedures and criteria: executive summary. 2017. http://jsnm.sakura.ne.jp/wp_jsnm/wp-content/themes/theme_jsnm/doc/StandardPETProtocolPhantom20170201.pdf. Accessed 14 Aug 2021.
28. Senda M. Standardization of PET imaging and site qualification program by JSNM: collaboration with EANM/EARL. Ann Nucl Med. 2020;34:873–4.
29. Fukukita H, Suzuki K, Matsumoto K, Terauchi T, Daisaki H, Ikari Y, et al. Japanese guideline for the oncology FDG-PET/CT data acquisition protocol: synopsis of Version 2.0. Ann Nucl Med. 2014;28:693–705.
30. Panin VY, Kehren F, Michel C, Casey M. Fully 3-D PET reconstruction with system matrix derived from point source measurements. IEEE Trans Med Imaging. 2006;25:907–21.
31. Rahmim A, Qi J, Sossi V. Resolution modeling in PET imaging: theory, practice, benefits, and pitfalls. Med Phys. 2013;40:064301.
32. Vandenberghe S, Mikhaylova E, D'Hoe E, Mollet P, Karp JS. Recent developments in time-of-flight PET. EJNMMI Physics. 2016;3:3.
33. Surti S, Karp JS. Update on latest advances in time-of-flight PET. Phys Med. 2020;80:251–8.
34. Teoh EJ, McGowan DR, Bradley KM, Belcher E, Black E, Gleeson FV. Novel penalised likelihood reconstruction of PET in the assessment of histologically verified small pulmonary nodules. Eur Radiol. 2016;26:576–84.
35. van Sluis J, Boellaard R, Somasundaram A, van Snick PH, Borra RJH, Dierckx RAJO, et al. Image quality and semiquantitative measurements on the Biograph Vision PET/CT System: Initial experiences and comparison with the Biograph mCT. J Nucl Med. 2020;61:129–35.
36. Kataoka J, Kishimoto A, Fujita T, Nishiyama T, Kurei Y, Tsujikawa T, et al. Recent progress of MPPC-based scintillation detectors in high precision X-ray and gamma-ray imaging. Nucl Instruments Method Phys Res Sect A Accel Spectrometers, Detect Assoc Equip. 2015;784:248–54.
37. Wagatsuma K, Miwa K, Sakata M, Oda K, Ono H, Kameyama M, et al. Comparison between new-generation SiPM-based and conventional PMT-based TOF-PET/CT. Phys Med. 2017;42:203–10.
38. Ota R. Photon counting detectors and their applications ranging from particle physics experiments to environmental radiation monitoring and medical imaging. Radiol Phys Technol. 2021;14:134–48.
39. Boellaard R, Sera T, Kaalep A, Hoekstra OS, Barrington SF, Zijlstra JM. Updating PET/CT performance standards and PET/CT interpretation criteria should go hand in hand. EJNMMI Res. 2019;9:5–6.
40. Weber WA, Gatsonis CA, Mozley PD, Hanna LG, Shields AF, Aberle DR, et al. Repeatability of 18F-FDG PET/CT in advanced non-small cell lung cancer: prospective assessment in 2 multicenter trials. J Nucl Med. 2015;56:1137–43.
41. Machtay M, Duan F, Siegel BA, Snyder BS, Gorelick JJ, Reddin JS, et al. Prediction of survival by [18F]fluorodeoxyglucose positron emission tomography in patients with locally advanced non-small-cell lung cancer undergoing definitive chemoradiation therapy: results of the ACRIN 6668/RTOG 0235 trial. J Clin Oncol. 2013;31:3823–30.
42. Kahraman D, Scheffler M, Zander T, Nogova L, Lammertsma AA, Boellaard R, et al. Quantitative analysis of response to treatment with erlotinib in advanced non-small cell lung cancer using 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET. J Nucl Med. 2011;52:1871–7.
43. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. J Nucl Med. 2012;53:1041–7.
44. Akamatsu G, Ikari Y, Nishida H, Nishio T, Ohnishi A, Maebatake A, et al. Influence of statistical fluctuation on reproducibility and accuracy of SUVmax and SUVpeak: a phantom study. J Nucl Med Technol. 2015;43:222–6.
45. Kaneta T, Ogawa M, Motomura N, Iizuka H, Arisawa T, Hino-Shishikura A, et al. Initial evaluation of the Celesteion large-bore PET/CT scanner in accordance with the NEMA NU2–2012 standard and the Japanese guideline for oncology FDG PET/CT data acquisition protocol version 2.0. EJNMMI Res. 2017;7:83.
46. Reddin JS, Scheuermann JS, Bharkhada D, Smith AM, Casey ME, Conti M, et al. Performance evaluation of the SiPM-based siemens biograph vision PET/CT system. IEEE Nucl Sci Symp Med Imaging Conf Rec. 2018;2018:1–5.
47. Rausch I, Cal-González J, Dapra D, Gallowitsch HJ, Lind P, Beyer T, et al. Performance evaluation of the Biograph mCT Flow PET/CT system according to the NEMA NU2-2012 standard. EJNMMI Phys. 2015;2:26.
48. Jakoby BW, Bercier Y, Conti M, Casey ME, Bendriem B, Townsend DW. Physical and clinical performance of the mCT time-of-flight PET/CT scanner. Phys Med Biol. 2011;56:2375–89.
49. Pan T, Einstein SA, Kappadath SC, Grogg KS, Lois Gomez C, Alessio AM, et al. Performance evaluation of the 5-Ring GE Discovery MI PET/CT system using the national electrical manufacturers association NU 2–2012 Standard. Med Phys. 2019;46:3025–33.
50. Hsu DFC, Ilan E, Peterson WT, Uribe J, Lubberink M, Levin CS. Studies of a next-generation silicon-photomultiplier–based time-of-flight PET/CT system. J Nucl Med. 2017;58:1511–8.
51. Vandendriessche D, Uribe J, Bertin H, De Geeter F. Performance characteristics of silicon photomultiplier based 15-cm AFOV TOF PET/CT. EJNMMI Phys. 2019;6:8.
52. Michopoulou S, O'Shaughnessy E, Thomson K, Guy MJ. Discovery molecular imaging digital ready PET/CT performance evaluation according to the NEMA NU2-2012 standard. Nucl Med Commun. 2019;40:270–7.
53. Reynés-Llompart G, Gámez-Cenzano C, Romero-Zayas I, Rodríguez-Bel L, Vercher-Conejero JL, Martí-Climent JM. Performance characteristics of the whole-body discovery IQ PET/CT system. J Nucl Med. 2017;58:1155–61.
54. Demir M, Toklu T, Abuqbeitah M, Çetin H, Sezgin HS, Yeyin N, et al. Evaluation of PET scanner performance in PET/MR and PET/CT systems: NEMA tests. Mol Imaging Radionucl Ther. 2018;27:10–8.
55. Bettinardi V, Presotto L, Rapisarda E, Picchio M, Gianolli L, Gilardi MC. Physical performance of the new hybrid PET∕CT Discovery-690. Med Phys. 2011;38:5394–411.
56. De Ponti E, Morzenti S, Guerra L, Pasquali C, Arosio M, Bettinardi V, et al. Performance measurements for the PET/CT Discovery-600 using NEMA NU 2–2007 standards. Med Phys. 2011;38:968–74.
57. Zhang J, Maniawski P, Knopp MV. Performance evaluation of the next generation solid-state digital photon counting PET/CT system. EJNMMI Res. 2018. https://doi.org/10.1186/s13550-018-0448-7.
58. Kolthammer JA, Su K, Grover A, Narayanan M, Jordan DW, Muzic RF. Performance evaluation of the Ingenuity TF PET/CT scanner with a focus on high count-rate conditions. Phys Med Biol. 2014;59:3843–59.
59. Surti S, Kuhn A, Werner ME, Perkins AE, Kolthammer J, Karp JS. Performance of Philips Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities. J Nucl Med. 2007;48:471–80.
60. Xu B, Changbin L, Yun D, Renming T, Yachao L, Hui Y, et al. Performance evaluation of a high-resolution TOF clinical PET/CT. J Nucl Med. 2016;57(Suppl 2):202.

61. Teoh EJ, McGowan DR, Macpherson RE, Bradley KM, Gleeson FV. Phantom and clinical evaluation of the bayesian penalized likelihood reconstruction algorithm Q.Clear on an LYSO PET/CT system. J Nucl Med. 2015;56:1447–52.
62. Matsumoto K, Endo K. Development of analysis software package for the two kinds of Japanese Fluoro-D-glucose-positron emission tomography guideline. Japanese J Radiol Technol. 2013;69:648–54.
63. NEMA. NEMA Standards Publication NU 2–2018: performance measurements of positron emission tomographs. Rosslyn, VA: National Electrical Manufacturers Association; 2018.
64. Strother SC, Casey ME, Hoffman EJ. Measuring PET scanner sensitivity: relating countrates to image signal-to-noise ratios using noise equivalents counts. IEEE Trans Nucl Sci. 1990;37:783–8.
65. Badawi RD, Dahlbom M. NEC: some coincidences are more equivalent than others. J Nucl Med. 2005;46:1767–8.
66. Brasse D, Kinahan PE, Lartizien C, Comtat C, Casey M, Michel C. Correction methods for random coincidences in fully 3D whole-body PET: impact on data and image quality. J Nucl Med. 2005;46:859–67.
67. Reader AJ, Julyan PJ, Williams H, Hastings DL, Zweit J. EM algorithm resolution modeling by image-space convolution for PET reconstruction. IEEE Nucl Sci Symp Conf Rec. 2002;2002:1221–5.
68. Kidera D, Kihara K, Akamatsu G, Mikasa S, Taniguchi T, Tsutsui Y, et al. The edge artifact in the point-spread function-based PET reconstruction at different sphere-to-background ratios of radioactivity. Ann Nucl Med. 2016;30:97–103.
69. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. Bone Marrow Transplant. 2013;48:452–8.
70. Masuda Y, Kondo C, Matsuo Y, Uetani M, Kusakabe K. Comparison of imaging protocols for 18F-FDG PET/CT in overweight patients: optimizing scan duration versus administered dose. J Nucl Med. 2009;50:844–8.
71. Chang T, Chang G, Clark JW, Diab RH, Rohren E, Mawlawi OR. Reliability of predicting image signal-to-noise ratio using noise equivalent count rate in PET imaging. Med Phys. 2012;39:5891–900.
72. Maebatake A, Akamatsu G, Miwa K, Tsutsui Y, Himuro K, Baba S, et al. Relationship between the image quality and noise-equivalent count in time-of-flight positron emission tomography. Ann Nucl Med. 2016;30:68–74.
73. Akamatsu G, Ishikawa K, Mitsumoto K, Taniguchi T, Ohya N, Baba S, et al. Improvement in PET/CT image quality with a combination of point-spread function and time-of-flight in relation to reconstruction parameters. J Nucl Med. 2012;53:1716–22.
74. Matsumoto K, Suzuki K, Fukukita H, Ikari Y, Oda K, Kimura Y, et al. Variability in PET quantitation within a multicenter studies in Japan. Eur J Nucl Med Mol Imaging. 2013;40(Suppl 2):S305.
75. El Fakhri G, Surti S, Trott CM, Scheuermann J, Karp JS. Improvement in lesion detection with whole-body oncologic time-of-flight PET. J Nucl Med. 2011;52:347–53.
76. Akamatsu G, Mitsumoto K, Taniguchi T, Tsutsui Y, Baba S, Sasaki M. Influences of point-spread function and time-of-flight reconstructions on standardized uptake value of lymph node metastases in FDG-PET. Eur J Radiol. 2014;83:226–30.
77. Munk OL, Tolbod LP, Hansen SB, Bogsrud TV. Point-spread function reconstructed PET images of sub-centimeter lesions are not quantitative. EJNMMI Phys. 2017;4:5.
78. Kaalep A, Burggraaff CN, Pieplenbosch S, Verwer EE, Sera T, Zijlstra J, et al. Quantitative implications of the updated EARL 2019 PET–CT performance standards. EJNMMI Phys. 2019;6:1–16.
79. 18F-FDG PET/CT UPICT Protocol Writing Committee. UPICT Oncology FDG-PET CT Protocol. http://qibawiki.rsna.org/images/d/de/UPICT_Oncologic_FDG-PETCTProtocol_6-07-13.pdf.
80. Ulrich EJ, Sunderland JJ, Smith BJ, Mohiuddin I, Parkhurst J, Plichta KA, et al. Automated model-based quantitative analysis of phantoms with spherical inserts in FDG PET scans. Med Phys. 2018;45:258–76.
81. SNMMI Phantom Analysis Toolkit (PAT). https://www.snmmi.org/PAT. Accessed 14 Aug 2021.
82. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. J Nucl Med. 2007;48:932–45.
83. Vanderhoek M, Perlman SB, Jeraj R. Impact of the definition of peak standardized uptake value on quantification of treatment response. J Nucl Med. 2012;53:4–11.
84. Watabe T, Tatsumi M, Watabe H, Isohashi K, Kato H, Yanagawa M, et al. Intratumoral heterogeneity of F-18 FDG uptake differentiates between gastrointestinal stromal tumors and abdominal malignant lymphomas on PET/CT. Ann Nucl Med. 2012;26:222–7.
85. Miwa K, Inubushi M, Wagatsuma K, Nagao M, Murata T, Koyama M, et al. FDG uptake heterogeneity evaluated by fractal analysis improves the differential diagnosis of pulmonary nodules. Eur J Radiol. 2014;83:715–9.
86. Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJR. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. Eur J Nucl Med Mol Imaging. 2013;40:133–40.
87. Lim R, Eaton A, Lee NY, Setton J, Ohri N, Rao S, et al. 18F-FDG PET/CT metabolic tumor volume and total lesion glycolysis predict outcome in oropharyngeal squamous cell carcinoma. J Nucl Med. 2012;53:1506–13.
88. Kitajima K, Miyoshi Y, Sekine T, Takei H, Ito K, Suto A, et al. Harmonized pretreatment quantitative volume-based FDG-PET/CT parameters for prognosis of stage I-III breast cancer: Multicenter study. Oncotarget. 2021;12:95–105.

Authors and Affiliations

Go Akamatsu[1] · Naoki Shimada[2] · Keiichi Matsumoto[3] · Hiromitsu Daisaki[4] · Kazufumi Suzuki[5] · Hiroshi Watabe[6] · Keiichi Oda[7] · Michio Senda[8] · Takashi Terauchi[2] · Ukihide Tateishi[9]

[1] National Institutes for Quantum Science and Technology (QST), 4-9-1 Anagawa, Inage-ku, Chiba 263-8555, Japan

[2] Cancer Institute Hospital, 3-8-31 Ariake, Koto, Tokyo 135-8550, Japan

[3] Kyoto College of Medical Science, 1-3 Imakita, Oyamahigashi-cho, Sonobe-cho, Nantan, Kyoto 622-0041, Japan

[4] Gunma Prefectural College of Health Sciences, 323-1 Kamioki-machi, Maebashi, Gunma 371-0052, Japan

[5] Dokkyo Medical University Hospital, 880 Kitakobayashi, Mibu, Shimotsugagun, Tochigi 321-0293, Japan

[6] Tohoku University, 6-3 Aoba, Aramaki, Aoba-ku, Sendai, Miyagi 980-8578, Japan

[7] Hokkaido University of Science, 7-Jo 15-4-1 Maeda, Teine, Sapporo, Hokkaido 006-8585, Japan

[8] Kobe City Medical Center General Hospital, 2-1-1 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

[9] Tokyo Medical and Dental University School of Medicine, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan