Am J Nucl Med Mol Imaging 2017;7(5):204-211
www.ajnmmi.us /ISSN:2160-8407/ajnmmi0061851

Original Article

Variation and repeatability of measured standardized uptake values depending on actual values: a phantom study

Tomohiro Kaneta[1], Na Sun[1,2], Matsuyoshi Ogawa[1], Hitoshi Iizuka[1], Tetsu Arisawa[1], Ayako Hino-Shishikura[1], Keisuke Yoshida[1], Tomio Inoue[1]

[1]*Department of Radiology, Yokohama City University, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan;* [2]*Department of Nuclear Medicine, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, 800 Dongchuan Road, Shanghai 200240, China*

Received July 19, 2017; Accepted September 12, 2017; Epub November 1, 2017; Published November 15, 2017

Abstract: Standardized uptake values (SUVs) are the most widely used quantitative imaging biomarkers in positron emission tomography (PET); however, little is known about the changes in variation and repeatability of SUVs depending on the magnitude of the values. We hypothesized that low SUVs have larger variations than high SUVs, and attempted various kinds of experimental PET scans using a phantom. By adjusting the ratio of F-18 solution to tap water, a NEMA IEC body phantom was set for SUVs of 2.0, 4.0, and 8.0 inside six hot spheres. PET data were obtained for 4 hours, and the data reconstructed every 2 min. The SUVmax and SUVpeak of the spheres in all images were recorded. The relative SUVs were calculated by dividing the measured SUV by actual SUV, and used for the Bland-Altman plots. Some variation was observed for the measured SUVs. The measured SUVs for the actual SUV of 2.0 showed the largest variation among those of 2.0, 4.0, and 8.0, and those of 8.0 showed the smallest. Similarly, the relative SUVs showed significantly larger variations for lower values. In addition, the relative SUVmax showed larger variation and value than the relative SUVpeak. The Bland-Altman plots showed considerable variation and little agreement, but the degree of variation decreased as the measured value increased. We demonstrated some variation of the measured SUVs, which decreased for larger measured values. Clinicians should consider the inaccuracy of low SUVs not only in daily practice, but also for multi-institutional studies.

Keywords: PET, SUV, variation, repeatability, QIBA

Introduction

The standardized uptake value (SUV) is the most widely used quantitative imaging biomarker (QIB) in the field of positron emission tomography (PET). This quantitative value has been used for the evaluation of differential diagnosis, therapeutic effect, and prognostic prediction. Among SUVs, the maximum SUV (SUVmax) is the most common QIB in daily practice and clinical research [1-5], because it is easily calculated and is not influenced by the size or shape of the region of interest (ROI). It is well known that SUVs are influenced by many factors, such as scanner, workstation, protocol, body habitus, disease distribution, image noise, and radioactivity [6-8]. However, the changes in variation and repeatability of measured SUVs depending on their magnitude have not been well researched. Variation in clinical practice results in poorer outcomes and higher costs. As Quantitative Imaging Biomarkers Alliance (QIBA) organized by the Radiological Society of North America (RSNA) claims, reducing variation will improve the value and practicability of quantitative imaging biomarkers (https://www.rsna.org/qiba/). In this study, our hypothesis is that low SUVs have larger variations than high SUVs. To verify this hypothesis, we attempted various kinds of PET scans using a National Electrical Manufacturers Association (NEMA) International Electrotechnical Commission (IEC) body phantom [9], and assessed the changes in variation and repeatability of measured SUVs depending on the actual SUVs.

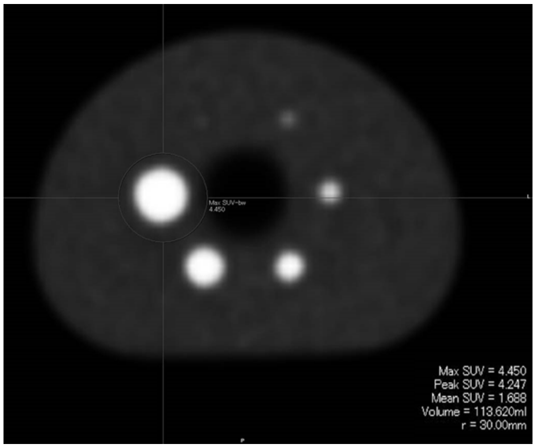

Figure 1. A sample PET images of a NEMA phantom with a VOI. Both the SUVmax and SUVpeak of all six spheres (inner diameter 37, 28, 22, 17, 13, and 10 mm) in the phantom were measured using a VOI that covered the whole sphere and was recorded for all reconstructed images.

Methods

Phantom preparation

An image-quality IEC body phantom of the type described in the NEMA NU-2 2012 Standard [9] was used for the experiments. We made phantoms for the evaluation of an SUV of 2.0, 4.0, and 8.0, as follows: First, we measured the background volume of the phantoms beforehand. Then, using a regularly checked dose calibrator and taking decay into consideration, F-18 solution was prepared at an average concentration of 3.7 kBq/ml (assuming a standard injected dose for adults in Japan: 185 MBq per 50 kg) for at least one hour before the start of data acquisition. For example, when making a phantom for the evaluation of an SUV of 4.0, exactly one-fourth of the background volume was filled with tap water, and a precise amount of F-18 fluorodeoxyglucose (FDG) was added to make a hot solution. An aliquot of this solution was added into all six (10-, 13-, 17-, 22-, 28-, and 37-mm diameter) hot spheres. The phantom background was filled with tap water and stirred to make a warm solution. Similarly, to make a phantom of SUV 2.0 and 8.0, one-half and one-eighth, respectively, of the background volume was filled with tap water first. The radioactivity of the FDG used for the experiments was measured and corrected using a standard radiation source (Ge-68 Dose Calibrator Source (37MBq) BM06S-CE, RadQual, NH, USA).

PET/CT scan

We used a *Celesteion* PET/CT scanner (Toshiba Medical Systems, Tochigi, Japan), which combines a high-speed helical 16-slice CT scanner and a lutetium-yttrium oxyorthosilicate (LYSO) scintillator block detector PET scanner. The PET scanner can acquire data in three-dimensional (3D) configurations. The energy window of the system was set to 425-650 keV, while the coincidence time window was set to 2.7 ns. The TOF temporal resolution was reported to be < 450 ps by Toshiba.

A CT transmission scan was performed using a 120-kV tube voltage. The FOV was set to 550 mm. Subsequently, a PET emission scan for 2 h was performed using list-mode for 2 times in a row (total 4 h). Two hours is the maximum duration of continuous scanning for this scanner.

PET images were reconstructed every 2 and 5 min by Time-of-flight (TOF) list-mode ordered subsets expectation maximization (TOF-LM-OSEM) using a 450-ps TOF temporal resolution kernel [10]. The TOF-LM-OSEM method is a TOF-OSEM algorithm using the Area-Simulating-Volume, which calculates the geometric probabilities in the system matrix of 3D PET systems. For the reconstruction, a 208 × 208 matrix size (pixel size 2 mm) was used and a post-reconstruction Gaussian filtering with 6-mm FWHM was applied.

SUV measurements

The SUVmax and SUVpeak of the reconstructed images for every 2 and 5 min were calculated using Vox-base II (J-MAC, Sapporo, Japan). SUVpeak is defined as the average SUV within a small sphere (10 mm in diameter) centered on the highest uptake region in the volume of interest (VOI), as opposed to SUVmax, which is the most intense voxel within the VOI. Both the SUVmax and SUVpeak of all six spheres (inner diameter 37, 28, 22, 17, 13, and 10 mm) in the phantom were measured using a VOI that covered the whole sphere and was recorded for all reconstructed images. A sample PET image with a VOI is shown in **Figure 1**. The relative SUVs were calculated by dividing the measured SUV by the actual SUV.

To evaluate the repeatability of the measured SUVs, Bland-Altman plot analyses were per-

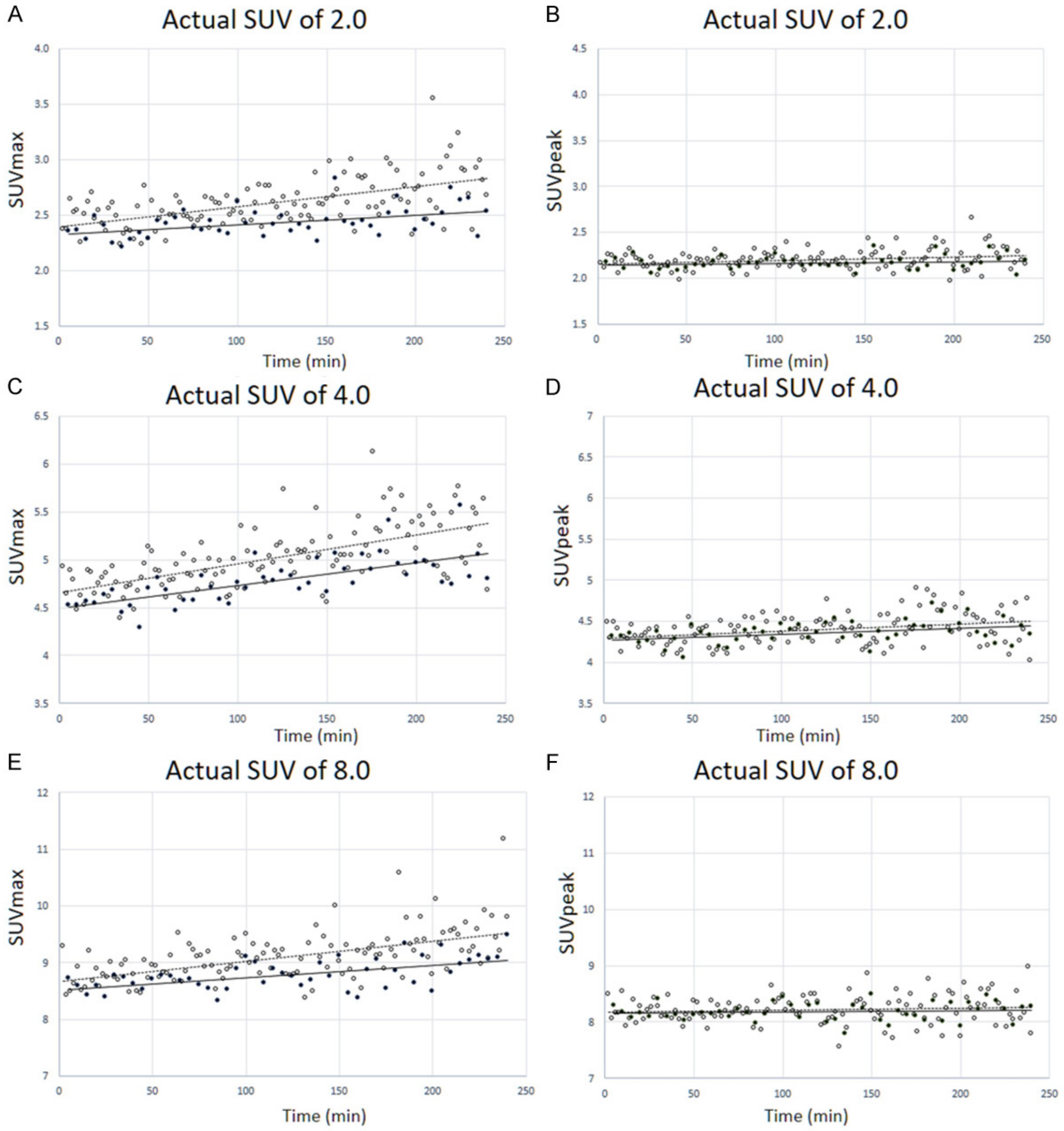

Figure 2. Changes of the measured standardized uptake values (SUVs) and their variations over time. SUVs were measured for 2-min (open circle, dotted line) and 5-min (black circle, solid line) scans. The measured SUVmax (A, C and E) and SUVpeak (B, D and F) correspond to the actual SUVs of 2.0, 4.0 and 8.0, respectively. The approximated lines of the plots are also indicated.

Table 1. Mean and SD of measured SUVs

		Actual SUV					
		2.0		4.0		8.0	
		SUVmax	SUVpeak	SUVmax	SUVpeak	SUVmax	SUVpeak
2 min*	Mean	2.61	2.20	5.02	4.40	9.10	8.22
	SD	0.17	0.08	0.33	0.19	0.32	0.19
5 min*	Mean	2.43	2.16	4.79	4.36	8.71	8.18
	SD	0.10	0.06	0.24	0.13	0.21	0.13

*acquisition time. SD: standard deviation.

formed. For the comparison of the data obtained from the actual SUV of 2.0, 4.0 and 8.0, the relative SUVs were used. The agreements between the relative SUVs and those measured 2 min later were evaluated. One hot sphere, which was the minimum size but not influ-

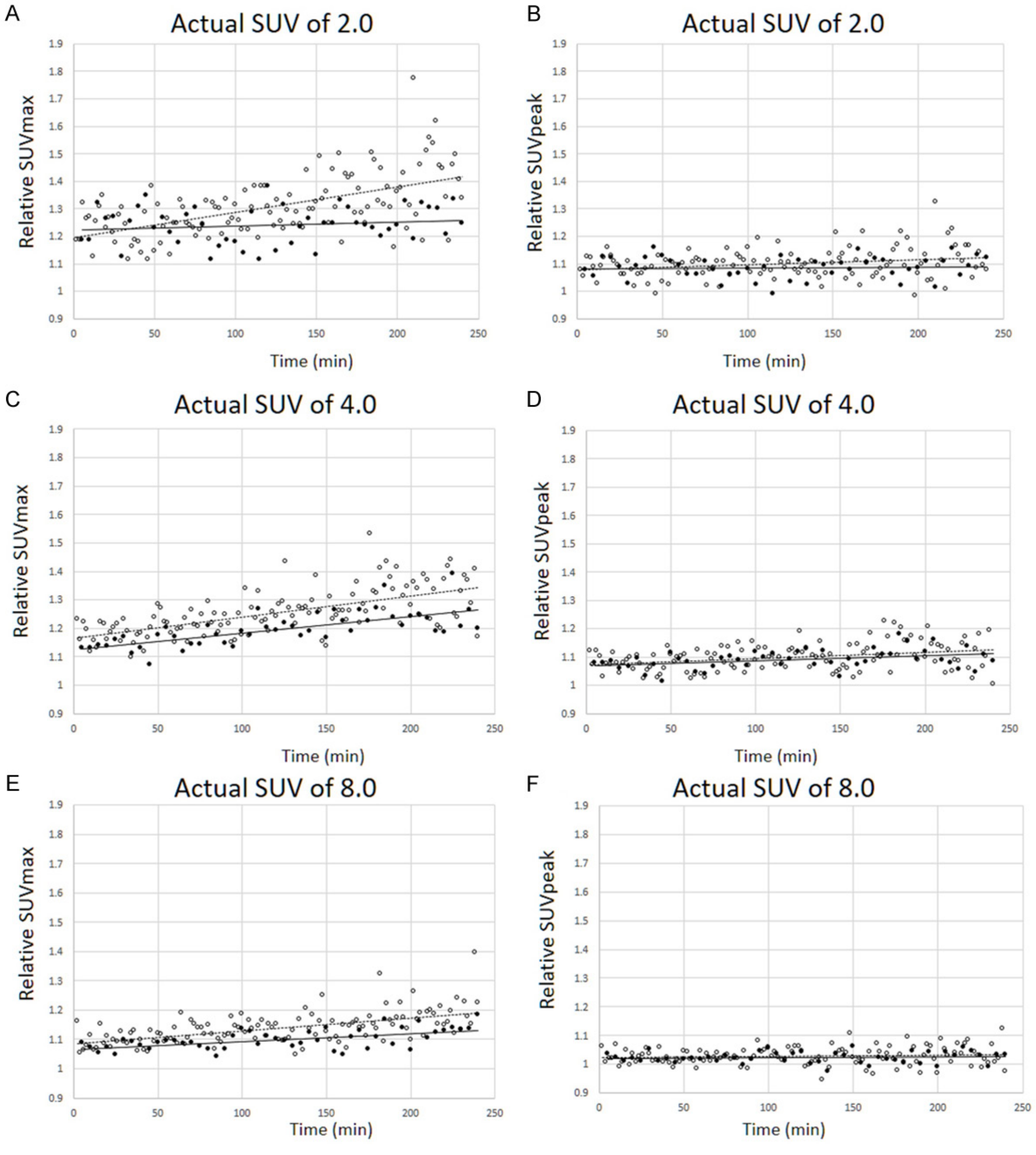

Figure 3. Changes of the relative SUVs and their variations over time. The relative SUVs were measured for 2-min (open circle, dotted line) and 5-min (black circle, solid line) scans. The relative SUVmax (A, C and E) and SUVpeak (B, D and F) correspond to the actual SUV of 2.0, 4.0 and 8.0, respectively. The approximated lines of the plots are also indicated.

enced by count recovery, was measured for the SUVs used in these analyses.

Statistical analysis

Levine's test for equality of variance was used for the evaluation of the variances among measured SUVs. A *p* value of < 0.01 was considered statistically significant.

Results

The radioactivity of the F-18 solution used to obtain SUVs of 2.0, 4.0, and 8.0 was 39.8, 37.9, and 32.7 MBq at the start of scanning, respectively. Among all spheres in the phantom, we selected the minimum size that was unlikely to be influenced by count recovery. The SUVmax and SUVpeak of the 22-mm diameter spheres

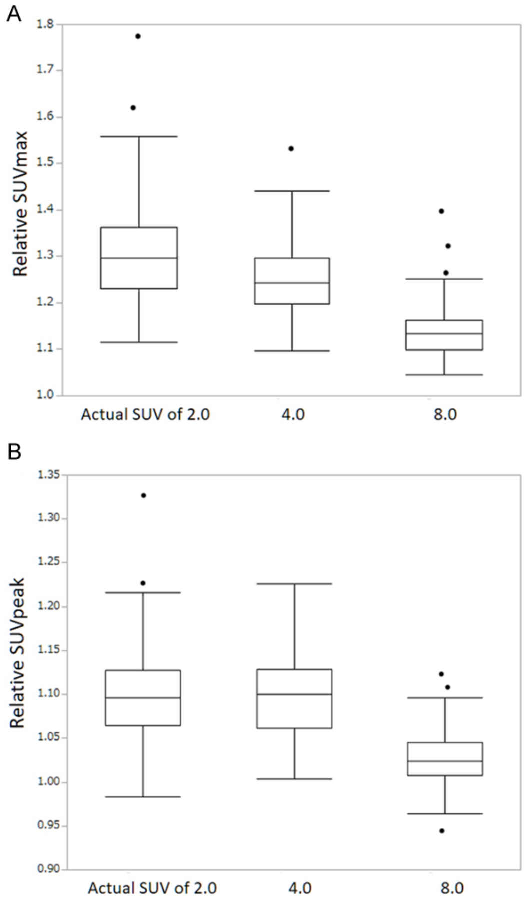

Figure 4. Boxplots of the SUVmax and SUVpeak for the actual SUVs of 2.0, 4.0 and 8.0, presenting five sample statistics - the minimum, the lower quartile, the median, the upper quartile and the maximum. There were significant differences between all groups (Levene test, p < 0.0001).

sometimes fell below the true values, but those of the 28-mm diameter spheres did not. Thus, the sphere with 28-mm diameter was used for the following analyses.

Changes in measured SUVs and their variations

Figure 2 shows the changes over time in SUVmax and SUVpeak for the images acquired at 2 and 5 min. SUVmax showed a gradual increase in both value and variation over time. SUVpeak also showed an increase, but to a lesser extent than SUVmax. The approximated lines of SUVmax and SUVpeak were calculated, and all lines demonstrated positive slopes. The slopes of the lines for the 2-min acquisition were steeper than those for the 5-min acquisition times, but the differences were smaller for SUVpeak than SUVmax. **Table 1** summarizes the mean and standard deviation (SD) of the measured SUVs.

Figure 3 shows the plots of the relative SUVs over time, and clearly demonstrated the differences among the actual SUVs of 2.0, 4.0 and 8.0. The measured SUVs for higher actual SUVs showed smaller variation and were closer to the actual values than those of smaller SUVs. This tendency was seen in SUVmax more clearly than in SUVpeak.

Figure 4 shows the boxplots of the SUVmax and SUVpeak for the actual SUV of 2.0, 4.0 and 8.0. There were significant differences in variation between each of the two groups (Levene test, p < 00001). In addition, there was significantly larger variation in the value of SUVmax and SUVpeak for smaller SUVs than for higher SUVs.

Repeatability of SUVs

Figure 5 shows the Bland-Altman plots of the relative SUVs for the actual SUVs of 2.0, 4.0 and 8.0. Both the relative SUVmax and SUVpeak showed small mean differences, but considerable variations and little agreements. The variations of both SUVmax and SUVpeak were largest for the actual SUV of 2.0, and smallest for that of 8.0.

Discussion

Our results clearly demonstrated the differences in variations and repeatability of the measured SUVs depending on the magnitude of the values, and succeeded in verifying our hypothesis that low SUVs have larger measurement variations than high SUVs. To our knowledge, this is the first paper that reported such changes in the measured SUVs depending on the magnitude of the actual SUV. Our results may provide some important suggestions for clinical PET practices and multi-institutional studies.

First, our results suggested that the reliability of the measured SUVs is not always the same; specifically, measurements of low SUVs are less reliable. There have been a few clinical studies which support our results. Doot *et al.* reported that the quantified changes in SUV before and after therapy were less accurate with decreases in the magnitude of SUV at the baseline scan [11]. Moreover, McDermott *et al.* reported that changes in serial SUV could not

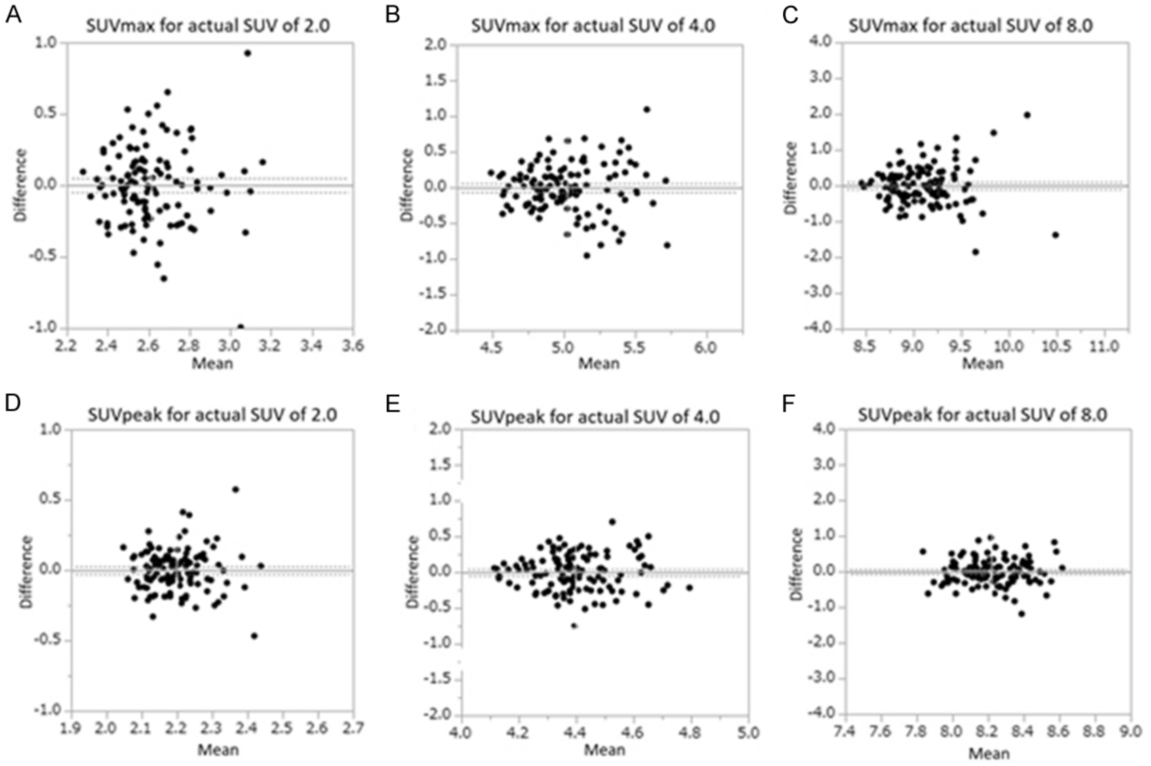

Figure 5. The Bland-Altman plots of relative SUVs with 2-min intervals. The relative SUVmax and SUVpeak for the actual SUV of 2.0, 4.0 and 8.0 are shown. The dotted horizontal lines represent the 95% confidence limits (limits of agreement), and the solid lines represent the mean of the differences.

differentiate between high and low responding groups to therapy when the tumor to background ratio was less than 5 [12]. However, this effect has never been verified by phantom studies. In addition, there have been several studies using SUVs in the liver, mediastinum (blood), or muscle as a reference; these regions commonly show low SUVs, around 2.0 [13, 14]. An interim PET evaluation using a Deauville 5-point scale [15] for malignant lymphoma used FDG uptake of the liver and mediastinum as a reference. The results of visual assessment may differ from those of SUV measurement. Due to a large variation of low SUVs, the use of SUVs in the liver or mediastinum as a reference may lead to unreliable results.

Next, the measured SUVs showed an increase in value and variation with time. This is thought to be caused by the increase of statistical noise over time [16-19] due to the decay of the radioactivity. It is well known that images reconstructed using a small amount of data due to a small administered dose or short acquisition time become noisy [16]. Statistical noise increases the SUV, and this effect is especially apparent in SUVmax. This may affect the interpretation of results in both clinical and research settings. For example, an increase of SUV in the delayed scan (2 h after injection) compared with that in the ordinary scan (1 h after injection) is thought to indicate a malignant tumor [18]. However, as our results suggested, SUVs, especially SUVmax, tend to increase with time. Interpreters should be aware of this issue when assessing delayed scans.

In clinical settings, a 2-min scan per bed position is common, and a 5-min scan may be rarely performed. Our results showed considerable differences for SUVmax between 2-min and 5-min scans, which were also clearly seen for low SUVs. In addition, there were much smaller differences in SUVpeak between the two acquisition times, compared to SUVmax. SUVmax is measured from a single voxel, regardless of whether this is a true value or noise. Some studies reported that mean SUV is less subject to noise than SUVmax [19]. Our results support the advantage of SUVpeak compared with SUVmax, as a quantitative imaging biomarker.

Our Bland-Altman plot analyses showed considerable variation and little agreement of measured SUVs. A considerable number of the differences within 2 min of each other were outside the 95% confidence intervals. These results may help determine what differences in SUV should be considered as significant. From our results, for example, there may not be a meaningful difference between an SUVmax of 2.4 and 2.8, though there is a 20% disparity between them. Further, there might be significant differences among institutions, scanners, protocols, etc.

This study was limited in that we did not consider the differences in scanners, workstations, reconstruction methods, corrections of attenuation and scatter, temporal changes of detectors, and so on. Further studies are required to optimize the SUV measurements.

In conclusion, we demonstrated a certain amount of variation of the measured SUVs, and the degree of variation decreased as the measured values increased. We should consider the inaccuracy of low SUVs not only for daily practice, but also for large multi-institutional studies.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Tomohiro Kaneta, Department of Radiology, Yokohama City University, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan. Tel: +81-45-787-2696; Fax: +81-45-786-0369; E-mail: kaneta@yokohama-cu.ac.jp

References

[1] Dehdashti F, Siegel BA, Griffeth LK, Fusselman MJ, Trask DD, McGuire AH, McGuire DJ. Benign versus malignant intraosseous lesions: discrimination by means of PET with 2-[F-18]fluoro-2-deoxy-D-glucose. Radiology 1996; 200: 243-247.

[2] Song BI, Lee SW, Jeong SY, Chae YS, Lee WK, Ahn BC, Lee J. 18F-FDG uptake by metastatic axillary lymph nodes on pretreatment PET/CT as a prognostic factor for recurrence in patients with invasive ductal breast cancer. J Nucl Med 2012; 53: 1337-1344.

[3] Song BI, Kim HW, Won KS, Ryu SW, Sohn SS, Kang YN. Preoperative standardized uptake value of metastatic lymph nodes measured by 18F-FDG PET/CT improves the prediction of prognosis in gastric cancer. Medicine (Baltimore) 2015; 94: e1037.

[4] Kitajima K, Suenaga Y, Kanda T, Miyawaki D, Yoshida K, Ejima Y, Sasaki R, Komatsu H, Saito M, Otsuki N, Nibu K, Kiyota N, Minamikawa T, Sugimura K. Prognostic value of FDG PET imaging in patients with laryngeal cancer. PLoS One 2014; 9: e96999.

[5] Berghmans T, Dusart M, Paesmans M, Hossein-Foucher C, Buvat I, Castaigne C, Scherpereel A, Mascaux C, Moreau M, Roelandts M, Alard S, Meert AP, Patz EF Jr, Lafitte JJ, Sculier JP; European Lung Cancer Working Party for the IASLC Lung Cancer Staging Project. Primary tumor standardized uptake value (SUVmax) measured on fluorodeoxyglucose positron emission tomography (FDG-PET) is of prognostic value for survival in non-small cell lung cancer (NSCLC): a systematic review and meta-analysis (MA) by the european lung cancer working party for the IASLC lung cancer staging project. J Thorac Oncol 2008; 3: 6-12.

[6] Beaulieu S, Kinahan P, Tseng J, Dunnwald LK, Schubert EK, Pham P, Lewellen B, Mankoff DA. SUV varies with time after injection in (18)F-FDG PET of breast cancer: characterization and method to adjust for time differences. J Nucl Med 2003; 44: 1044-1050.

[7] Keyes JW Jr. SUV: standard uptake or silly useless value? J Nucl Med 1995; 36: 1836-1839.

[8] 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.

[9] National Electrical Manufacturers Association. Performance measurements of positron emission tomographs. NEMA Standards Publication NU 2-2012. Rosslyn, USA: National Electical Manufacturers Association. 2012.

[10] Ye H, Niu X, Wang W. Improved list-mode reconstruction with an area simulating-volume projector in 3D PET: IEEE Medical Imaging Conference Record 2012.

[11] Doot RK, Dunnwald LK, Schubert EK, Muzi M, Peterson LM, Kinahan PE, Kurland BF, Mankoff DA. Dynamic and static approaches to quantifying 18F-FDG uptake for measuring cancer response to therapy, including the effect of granulocyte CSF. J Nucl Med 2007; 48: 920-925.

[12] McDermott GM, Welch A, Staff RT, Gilbert FJ, Schweiger L, Semple SI, Smith TA, Hutcheon AW, Miller ID, Smith IC, Heys SD. Monitoring primary breast cancer throughout chemotherapy using FDG-PET. Breast Cancer Res Treat 2007; 102: 75-84.

[13] Bütof R, Hofheinz F, Zöphel K, Stadelmann T, Schmollack J, Jentsch C, Löck S, Kotzerke J, Baumann M, van den Hoff J. Prognostic value

of pretherapeutic tumor-to-blood standardized uptake ratio in patients with esophageal carcinoma. J Nucl Med 2015; 56: 1150-1156.

[14] Hofheinz F, Bütof R, Apostolova I, Zöphel K, Steffen IG, Amthauer H, Kotzerke J, Baumann M, van den Hoff J. An investigation of the relation between tumor-to-liver ratio (TLR) and tumor-to-blood standard uptake ratio (SUR) in oncological FDG PET. EJNMMI Res 2016; 6: 19.

[15] Dupuis J, Berriolo-Riedinger A, Julian A, Brice P, Tychyj-Pinel C, Tilly H, Mounier N, Gallamini A, Feugier P, Soubeyran P, Colombat P, Laurent G, Berenger N, Casasnovas RO, Vera P, Paone G, Xerri L, Salles G, Haioun C, Meignan M. Impact of [(18)F]fluorodeoxyglucose positron emission tomography response evaluation in patients with high-tumor burden follicular lymphoma treated with immunochemotherapy: a prospective study from the Groupe d'Etudes des Lymphomes de l'Adulte and GOELAMS. J Clin Oncol 2012; 30: 4317-4322.

[16] Boellaard R, Krak NC, Hoekstra OS, Lammertsma AA. Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study. J Nucl Med 2004; 45: 1519-1527.

[17] Chen MK, Menard DH 3rd, Cheng DW. Determining the minimal required radioactivity of 18F-FDG for reliable semiquantification in PET/CT Imaging: a phantom study. J Nucl Med Technol 2016; 44: 26-30.

[18] Nakamoto Y, Higashi T, Sakahara H, Tamaki N, Kogire M, Doi R, Hosotani R, Imamura M, Konishi J. Delayed (18)F-fluoro-2-deoxy-D-glucose positron emission tomography scan for differentiation between malignant and benign lesions in the pancreas. Cancer 2000; 89: 2547-2554.

[19] Kumar V, Nath K, Berman CG, Kim J, Tanvetyanon T, Chiappori AA, Gatenby RA, Gillies RJ, Eikman EA. Variance of SUVs for FDG-PET/CT is greater in clinical practice than under ideal study settings. Clin Nucl Med 2013; 38: 175-182.