Am J Nucl Med Mol Imaging 2018;8(6):373-386
www.ajnmmi.us /ISSN:2160-8407/ajnmmi0087887

Original Article

Multi-level otsu method to define metabolic tumor volume in positron emission tomography

Hyung-Jun Im[1,5], Meiyappan Solaiyappan[4], Inki Lee[1,6], Tyler Bradshaw[2], Najat C Daw[7], Fariba Navid[8,9], Barry L Shulkin[10], Steve Y Cho[1,3]

[1]Radiology, [2]Medical Physics, University of Wisconsin, Madison, WI, USA; [3]University of Wisconsin Carbone Cancer Center, Madison, WI, USA; [4]Radiology, Johns Hopkins School of Medicine, Baltimore, MD, USA; [5]Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea; [6]Department of Nuclear Medicine, Korea Cancer Center Hospital, Korea Institutes of Radiological and Medical Sciences, Seoul, Republic of Korea; [7]Division of Pediatrics, MD Anderson Cancer Center, Houston, TX, USA; [8]Department of Pediatrics, Children's Hospital Los Angeles, Los Angeles, CA, USA; [9]Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; [10]Department of Diagnostic Imaging, St. Jude Children's Research Hospital, Memphis, TN, USA

Received November 5, 2018; Accepted December 4, 2018; Epub December 20, 2018; Published December 30, 2018

Abstract: This study was to validate reliability and clinical utility of a PET tumor segmentation method using multi-level Otsu (MO-PET) in standard National Electrical Manufacturers Association (NEMA) image quality (IQ) phantom and patients with osteosarcoma. The NEMA IQ phantom was prepared with a lesion-to-background ratio (LBR) of either 8:1, 4:1, or 1.5:1. The artificial lesions in the phantom were segmented using MO-PET, gradient-based method (PETedge), relative threshold methods, and background threshold methods. Metabolic tumor volumes (MTVs) using MO-PET and PETedge were named as MTV (MO-PET) and MTV (PETedge), respectively. Among the MTVs using multiple methods, only MTV (MO-PET) and MTV (PETedge) showed excellent agreements with the actual volume of NEMA IQ phantom across the different LBRs (intraclass correlation coefficient, ICC = 0.987, 0.985 in LBR 8:1, 0.981, 0.993 in LBR 4:1 and 0.947, 0.994 in LBR 1.5:1). Repeated measurements of MTV (MO-PET) of the primary tumors showed excellent reproducibility with ICC of 0.994 (0.989-0.997) in patients with osteosarcoma. Also, MTV (MO-PET) was found to be predictive of Event Free Survival (EFS) [Hazard ratio (95% CI) = 6.1 (2.1-17.2), log rank P = 0.0003] in patients with osteosarcoma. We have validated in NEMA IQ phantom that the MTV (MO-PET) is accurate, and importantly, stable and consistent across a range of lesion sizes and LBRs representative of clinical tumor lesions. Furthermore, MTV (MO-PET) showed excellent reproducibility and was predictive for EFS in patients with osteosarcoma.

Keywords: Multi-level otsu, metabolic tumor volume, ^{18}F-fluorodeoxyglucose, positron emission tomography, segmentation

Introduction

Metabolic tumor volume (MTV) and total lesion glycolysis (TLG) are radiomic parameters that represent metabolic tumor burden using ^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography/computed tomography (PET/CT). MTV is a measurement of tumor volume with increased glucose metabolism, while TLG is the product of MTV and the mean standardized uptake value (SUV) of the volume. MTV and TLG have been reported being useful markers for predicting clinical outcome and evaluation of response to treatment in multiple types of malignancies [1-7]. Moreover, MTV and TLG are considered to be better prognostic factors than simple metabolic parameter such as maximum SUV (SUVmax) [3, 8, 9]. However, MTV and TLG are not incorporated in the clinical practice yet, mainly because the optimal tumor segmentation method to measure the values has not been established [10, 11].

Osteosarcoma is the most common primary bone malignancy in pediatric and young adult patients, with a five-year survival of 65-75% for localized tumor and < 30% for metastatic tumor [12]. The standard of care treatment for high-grade osteosarcoma includes neoadjuvant chemotherapy (NCT) and subsequent surgical

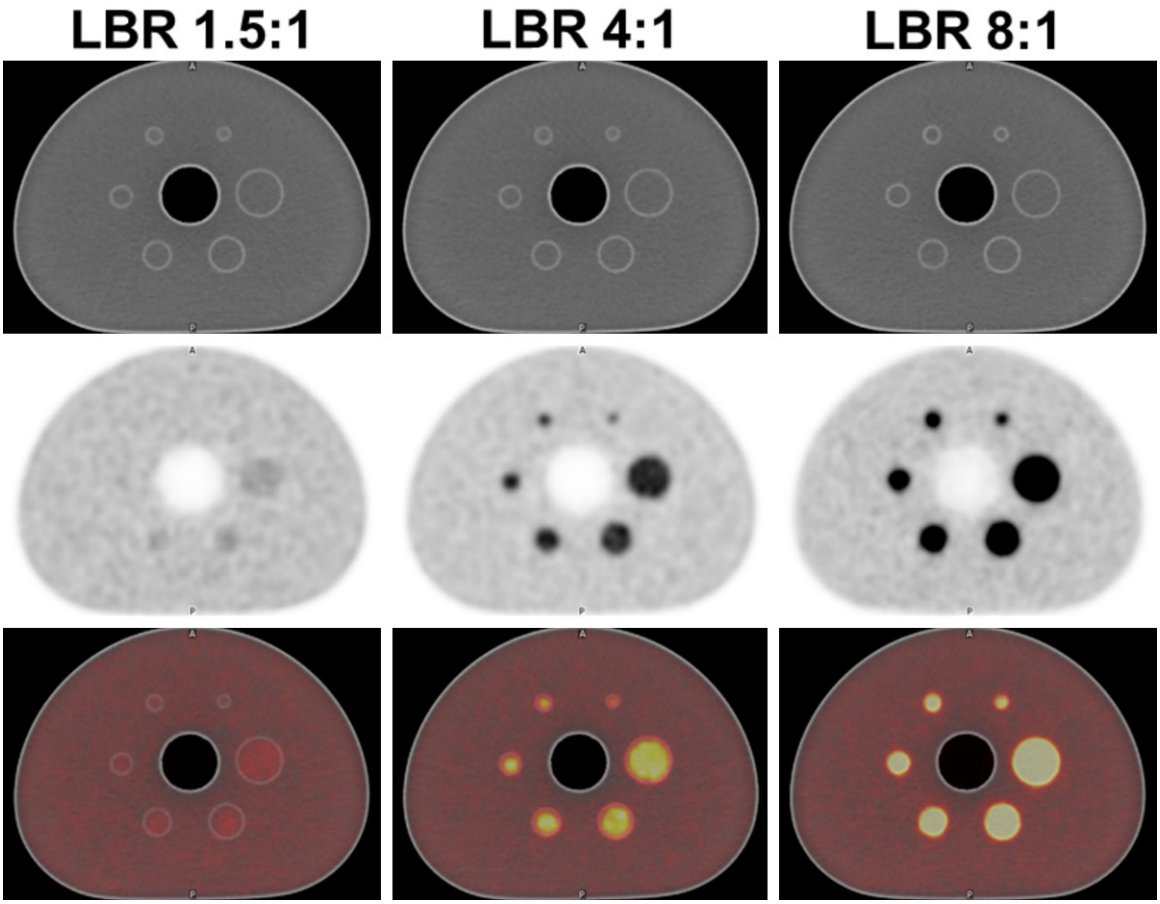

Figure 1. PET image of NEMA IQ phantom. NEMA IQ phantom was prepared to have three different lesion to background ratios (LBRs) of 1.5:1, 4:1 and 8:1. Of note, three smaller lesions were not visible in the phantom with 1.5:1 LBR. Upper row: CT, middle row: PET, lower row: PET/CT.

resection [13, 14]. In osteosarcoma, MTV has been reported to have the predictive value for clinical outcome [15-20]. However, the consensus has not been reached on the selection of an optimal tumor segmentation method to predict survival in osteosarcoma.

Recently we have developed multi-level Otsu threshold methods for tumor segmentation using FDG PET/CT (MO-PET) [21, 22]. Multi-level Otsu is a thresholding algorithm based on Otsu threshold strategy that classifies pixels of an image into two classes by searching for a threshold value that minimizes the intra-class variance defined as the weighted sum of the two classes. The weighting factor is the class-probabilities determined from the histogram. As a result, the threshold effectively separates the image pixels into background features and foreground features (corresponding to high metabolism). A tumor with high FDG uptake can be segmented with very minimal user interaction using the algorithm. The method showed stable and consistent delineation across a range of tumor sizes and SUV values using clinical PET scans of patients with melanoma [23]. However, MO-PET has not been previously tested in phantom images with known values for lesion volumes. Furthermore, the reproducibility and clinical usefulness of MO-PET in patients have not previously been evaluated.

Herein, we validated MO-PET as follows; 1) the accuracy and reliability using standard National Electrical Manufacturers Association (NEMA) image quality (IQ) phantom, 2) the reproducibility using FDG PET/CT scans of patients with osteosarcoma, and 3) the clinical usefulness, particularly, the ability for prediction of event free survival (EFS) in patients with osteosarcoma.

Table 1. Agreement between MTVs and actual volumes in NEMA IQ phantom

LBR	Segmentation method	ICC	95% CI	*P* value
8:1	20%	0.808	-0.233-0.974	0.005
	40%	0.998	0.946-1	< 0.0001
	60%	0.971	0.657-0.996	< 0.0001
	80%	0.816	-0.083-0.973	0.018
	BG+SD	0.676	-0.34-0.95	0.036
	BG+2SD	0.564	-0.349-0.926	0.063
	2BG	0.945	0.428-0.993	0.001
	2.5BG	0.979	0.786-0.997	< 0.0001
	MO-PET	0.987	0.877-0.998	< 0.0001
	PETedge	0.985	0.907-0.998	< 0.0001
4:1	20%	0.427	-0.235-0.884	0.071
	40%	0.995	0.947-0.999	< 0.0001
	60%	0.971	0.566-0.996	< 0.0001
	80%	0.816	-0.125-0.974	0.015
	BG+SD	0.862	-0.108-0.981	0.005
	BG+2SD	0.756	-0.285-0.965	0.016
	2BG	1.000	0.998-1	< 0.0001
	2.5BG	0.986	0.659-0.998	< 0.0001
	MO-PET	0.981	0.84-0.997	< 0.0001
	PETedge	0.993	0.954-0.999	< 0.0001
1.5:1	20%	0.245	-0.13-0.946	0.219
	40%	0.251	-0.13-0.947	0.216
	60%	0.823	-0.233-0.995	0.047
	80%	0.468	-0.476-0.98	0.233
	BG+SD	0.925	-0.144-0.998	0.017
	BG+2SD	0.989	0.643-1	0.005
	2BG	N/A	N/A	N/A
	2.5BG	N/A	N/A	N/A
	MO-PET	0.959	-0.600-0.999	0.039
	PETedge	0.994	0.906-1	0.006

20%, 40%, 60%, 80% = relative threshold methods (20%, 40%, 60%, or 80% of the maximum activity); BG+SD, BG+2SD, 2BG, 2.5BG = background threshold methods (mean background +1 or +2 standard deviations, and mean background × 2 or × 2.5); LBR: lesion to background ratio; ICC = Intra-class correlation coefficient.

Methods

PET imaging of NEMA IQ phantom

The NEMA IQ phantom was filled with an ^{18}F solution to have a uniform background activity of 3.05 KBq/ml. The six spherical artificial lesions in the phantom [diameter (cm)/volume (cm^3): 1/0.52, 1.3/1.15, 1.7/2.57, 2.2/5.57, 2.8/11.5, and 3.7/26.5] were filled with ^{18}F activity to have a lesion-to-background ratio (LBR) of either 8:1, 4:1, or 1.5:1 (**Figure 1**). The phantom was imaged using a GE Discovery 710 PET/CT scanner (GE Healthcare, Milwaukee, USA). Matrix size was 256 × 256. A size of voxel was 2.73 × 2.73 × 3.27 mm. The PET scans were reconstructed using VUE-point HD (VP-HD) which uses a three-dimensional maximum likelihood ordered subsets expectation maximization (OSEM) with corrections for scatter, random coincidences, dead time, and attenuation.

Patients

Thirty-four patients with newly diagnosed high-grade osteosarcoma were prospectively enrolled from June 2008 to May 2012 in a clinical trial (NCT00667342). The study was approved by the Institutional Review Board and written informed consent was received from all enrolled patients. All patients received a uniform protocol of neoadjuvant chemotherapy (NCT) and subsequent surgical resection in St. Jude Children's Research Hospital [24, 25]. FDG-PET/CT scans were done before (baseline scan), during (5 weeks after initiation of NCT, interim scan) and after the NCT (10 weeks after initiation of NCT, post-therapy scan). All patients were regularly assessed after surgery including bone scan and contrast-enhanced computed tomography (CT). When recurrence was suspected, further examinations were performed to confirm recurrence which included FDG-PET/CT, magnetic resonance imaging (MRI), and/or biopsy. EFS was defined as the time interval from study enrollment to date of first event (EFS: recurrence, progression, death), or to date of the last contact for patients without events. The survival data were censored at the time of the last visit if patients were alive or free of disease recurrence or progression. In this study, only baseline FDG PET/CT scans were used to evaluate the reproducibility and predictive value of MTV using the MO-PET for clinical outcome, particularly, EFS.

Detailed FDG-PET/CT image protocol is described previously [26]. Briefly, FDG (5.4 MBq per kg body weight; maximum 12 mCi) was injected after patients had fasted at least 4 hours. Transmission CT and PET emission scans were acquired one hour after the injection from the top of the skull to the feet. PET emission scans were acquired for 5 min per bed position in two-dimensional mode correct-

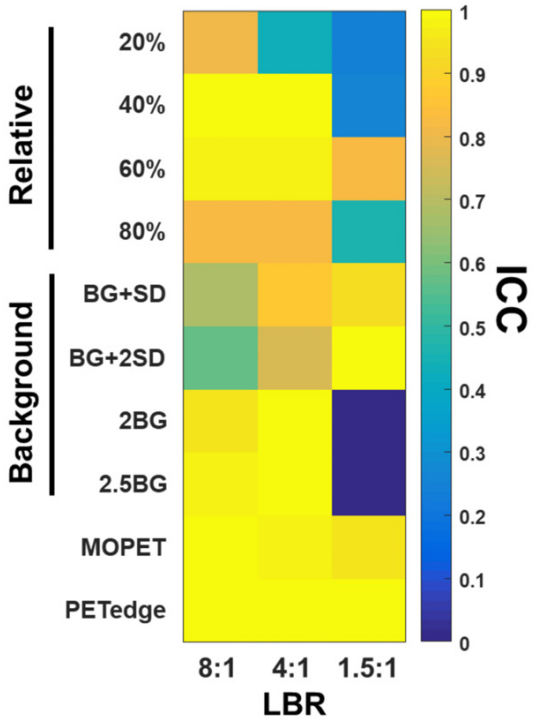

Figure 2. Intra-class correlation coefficient (ICC) between actual lesion volumes and MTVs using various segmentation methods. ICCs between actual lesion volumes and MTVs using various methods in different lesion to background ratios were color coded. Segmentation methods to measure MTV were listed on y axis and lesion to background ratios (LBRs) were in x axis. 20%, 40%, 60%, 80% = relative threshold methods (20%, 40%, 60%, or 80% of the maximum activity); BG+SD, BG+2SD, 2BG, 2.5BG = background threshold methods (mean background +1 or +2 standard deviations, and mean background × 2 or × 2.5); LBR: lesion to background ratio; ICC = Intra-class correlation coefficient.

ed for attenuation using CT scan. CT scans were acquired using parameters as follows: 120 kilovoltage peak (kVp), milliamperes per second (mAs) adjusted for body weight (maximum 90 mAs), slice thickness of 5 mm, tube rotation time of 0.8 seconds, a table speed of 1.5 centimeters per rotation, and pitch of 1.5 to 1.

Analysis of PET image

All image analysis was done using MIRADA XD3 program (MIRADA medical, Denver, CO, USA) or MIMvista software (MIMvista Corp, Cleveland, OH, USA). MO-PET segmentation was done by MO-PET plug-in for ImageJ (NIH, Bethesda, MD, USA). The lesions in the phantom were segmented using 4 approaches: 1) MO-PET, 2) gradient-based method using PETedge tool in MIMvista software, 3) conventional threshold methods which are relative threshold methods (20%, 40%, 60%, or 80% of the maximum activity), and 4) four background threshold methods (mean background +1 or +2 standard deviations, and mean background × 2 or × 2.5). Background uptake was measured using a spherical ROI with a diameter of 3 cm. The above defined MTVs are named as MTV (MO-PET), MTV (PETedge), MTV (20%), MTV (40%), MTV (60%), MTV (80%), MTV (BG+SD), MTV (BG+2SD), MTV (2BG), and MTV (2.5BG). Three small lesions were not evaluated in LBR of 1.5:1, because they were not well distinguishable from the background at this low LBR (**Figure 1**). To evaluate PET segmentation accuracy, we compared volume the ratio (VR) of MTV to the actual volume of the phantom lesions. VR closer to 1 indicated a better segmentation performance.

Baseline FDG PET/CT scans of thirty-four patients with osteosarcoma were used in this study. MTV (MO-PET) and MTV (PETedge) of primary tumors were measured. MTV (MO-PET) and MTV (PETedge) were measured twice, the initial observation separated by at least 2 months from the second measurement, by one nuclear medicine physician with seven years of experience.

Statistical analysis

The intra-class correlation coefficient (ICC), and Bland-Altman analysis were used to test the agreement of the measurements. The optimized cut-off was defined to produce the maximum Youden index from receiver operating characteristic (ROC) analysis. EFS distributions between high and low MTV groups were compared using the log-rank test. All statistical analyses were performed using SPSS software (SPSS Inc., Chicago, IL, USA). *P* values of less than 0.05 were considered statistically significant.

Results

Reliability of MO-PET segmentation in NEMA IQ phantom

Intra-class correlation coefficients (ICCs) were assessed between actual lesion volumes of the

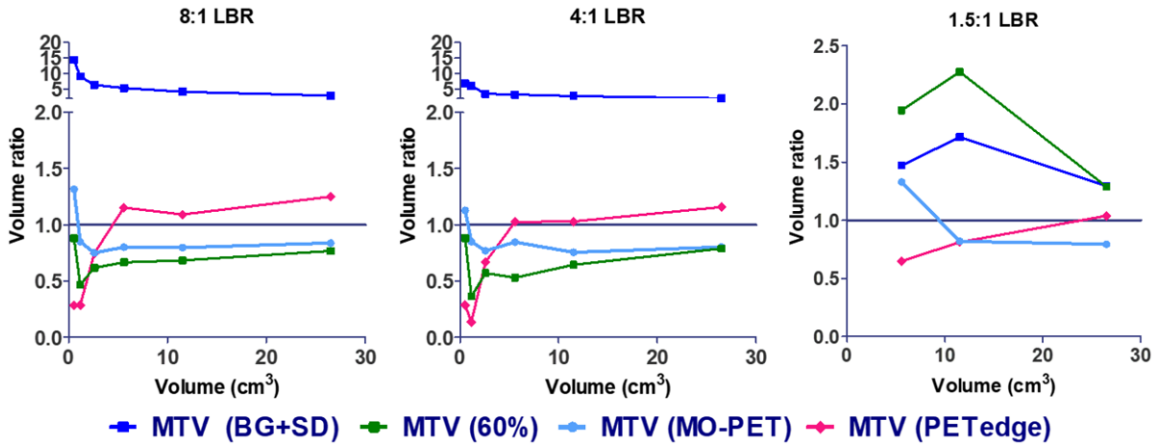

Figure 3. The volume ratio between selected MTVs and actual lesion volumes in NEMA IQ phantom. Volume ratio (VR) of MTV to the actual volume of the phantom lesions are displayed. VR closer to 1 indicated a better segmentation strategy. Three small lesions were not evaluated in LBR of 1.5:1, because they were not well distinguishable from the background at this low LBR. LBR: lesion to background ratio.

NEMA IQ phantom and measured MTVs using various segmentation methods. Among the MTVs using various methods, only MTV (MO-PET) and MTV (PETedge) showed excellent agreements with the actual lesion volumes in all three different LBRs (ICC = 0.987, 0.985 in LBR 8:1, 0.981, 0.993 in LBR 4:1 and 0.947, 0.994 in LBR 1.5:1). Among relative threshold MTVs, MTV (60%) was the most consistently reliable method across the different LBRs. MTV (BG+SD) was the most consistently reliable method among background based threshold MTV (**Table 1**; **Figure 2**). Thus MTV (MO-PET), MTV (PETedge), MTV (60%) and MTV (BG+SD) were selected for further analysis.

The VRs (volume ratio of MTV to the actual volume of the phantom lesion) were compared between the four selected methods. In 8:1, 4:1 LBR, MTV (60%), MTV (MO-PET) and MTV (PETedge) were accurately reflecting the actual lesion volumes (mean ± SD of VR in 8:1 LBR = 0.68 ± 0.14, 0.89 ± 0.21, and 0.80 ± 0.43, respectively; mean ± SD of VR in 4:1 LBR = 0.63 ± 0.19, 0.86 ± 0.14, and 0.72 ± 0.43, respectively), while MTV (BG+SD) overestimated the actual lesion volumes (mean ± SD of VR in 8:1 and 4:1 LBR = 7.08 ± 4.17, 4.08 ± 1.86, respectively). In 1.5:1 LBR, MTV (MO-PET) and MTV (PETedge) were stably reflecting the actual lesion volume (mean ± SD of VR = 0.98 ± 0.30 and 0.83 ± 0.19, respectively), however MTV (BG+SD) and MTV (60%) overestimated the lesion volume (mean ± SD of VR = 1.49 ± 0.21, and 1.84 ± 0.50, respectively) (**Figures 3** and **4A**). Furthermore, MTV (MO-PET) showed better lesion segmentation in the two smallest lesions in the phantom (diameters of 1 and 1.4 cm, respectively) in LBR 4:1 and LBR 8:1 than MTV (PETedge) (**Figures 3** and **4B**). In summary, MTV (MO-PET) demonstrated the most reliable volume estimation across the different LBRs and lesion sizes, and MTV (PETedge) showed comparable reliability with MTV (MO-PET) except for the two smallest lesions. Thus, MTV (MO-PET) and MTV (PETedge) were selected for further evaluation in clinical PET scans from patients with osteosarcoma.

Reproducibility of MTV (MO-PET) and MTV (PETedge) in patients with osteosarcoma

Agreements of repeated measurements using MO-PET and PETedge segmentation were assessed using baseline FDG PET/CT scans of patients with osteosarcoma. Repeated measurement of MTV (MO-PET) and MTV (PETedge) of the primary tumors showed excellent reproducibility with ICC of 0.994 and 0.997, respectively (**Figure 5**; **Table 2**). Also, measurements of SUVmean, TLG showed excellent reproducibility using both MO-PET and PETedge segmentation methods (**Table 2**). Repeated measurements of both MTV (MO-PET) and MTV (PETedge) showed minimal bias (1.1% and -0.3%) and relatively narrow lower and upper limits (-22.2%/24.3% and -24.5%/23.9%, respectively) in Bland-Altman analysis. Also,

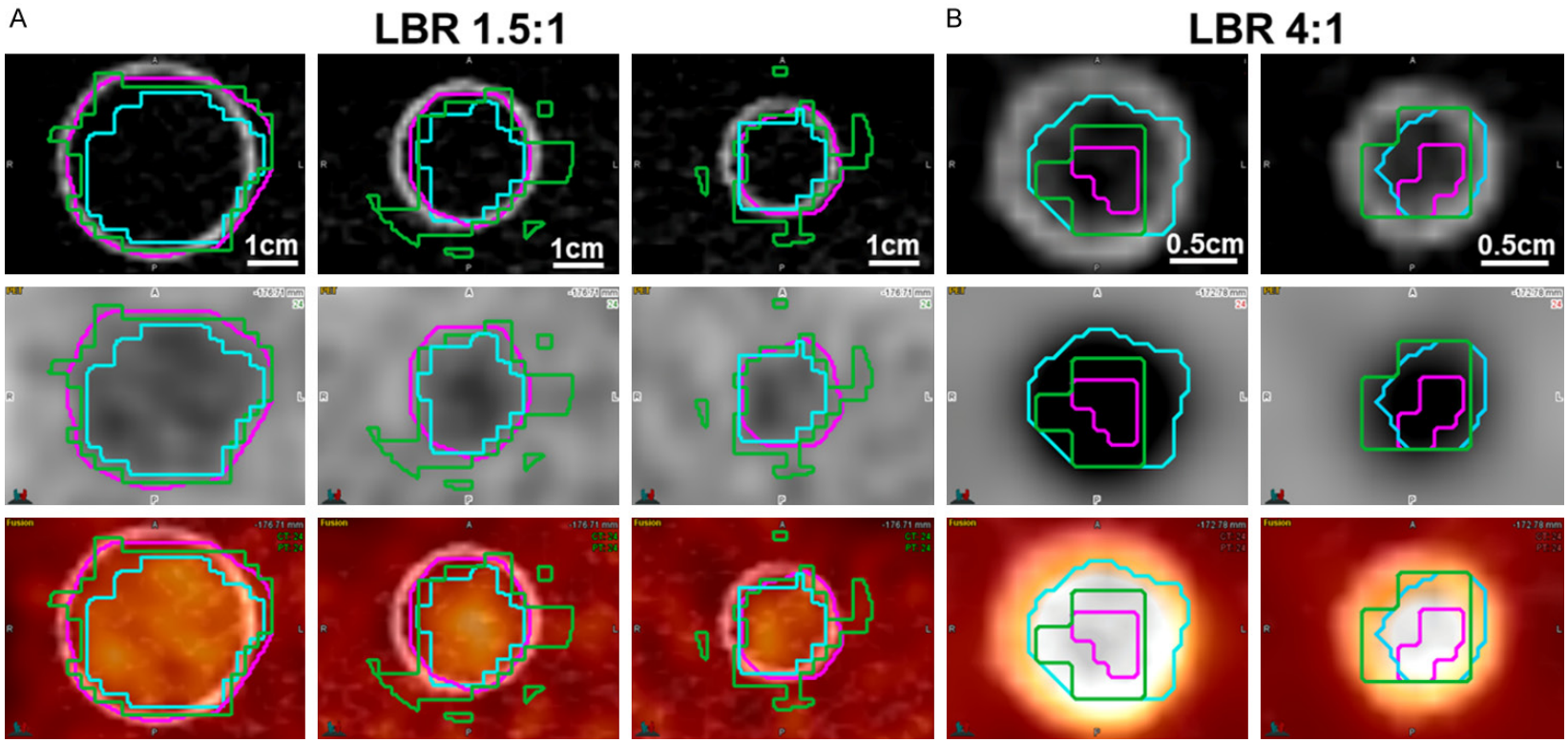

Figure 4. Representative NEMA IQ phantom PET image to compare accuracy between segmentation methods. A. In LBR 1.5:1, MO-PET and PETedge showed relatively accurate segmentation in three largest lesions in the phantom (diameter = 3.7, 2.8 and 2.2 cm) but 60% relative threshold overestimated the volumes in two lesions with diameters of 2.8 and 2.2 cm. B. In LBR 4:1 phantom, MO-PET (light blue colored ROI), and 60% relative threshold (green colored ROI) showed relatively accurate segmentation in two smallest lesions (diameter = 1.3, and 1 cm). However, PETedge (pink colored ROI) underestimated the actual volume of the two lesions. Upper row: CT, middle row: PET, lower row: PET/CT.

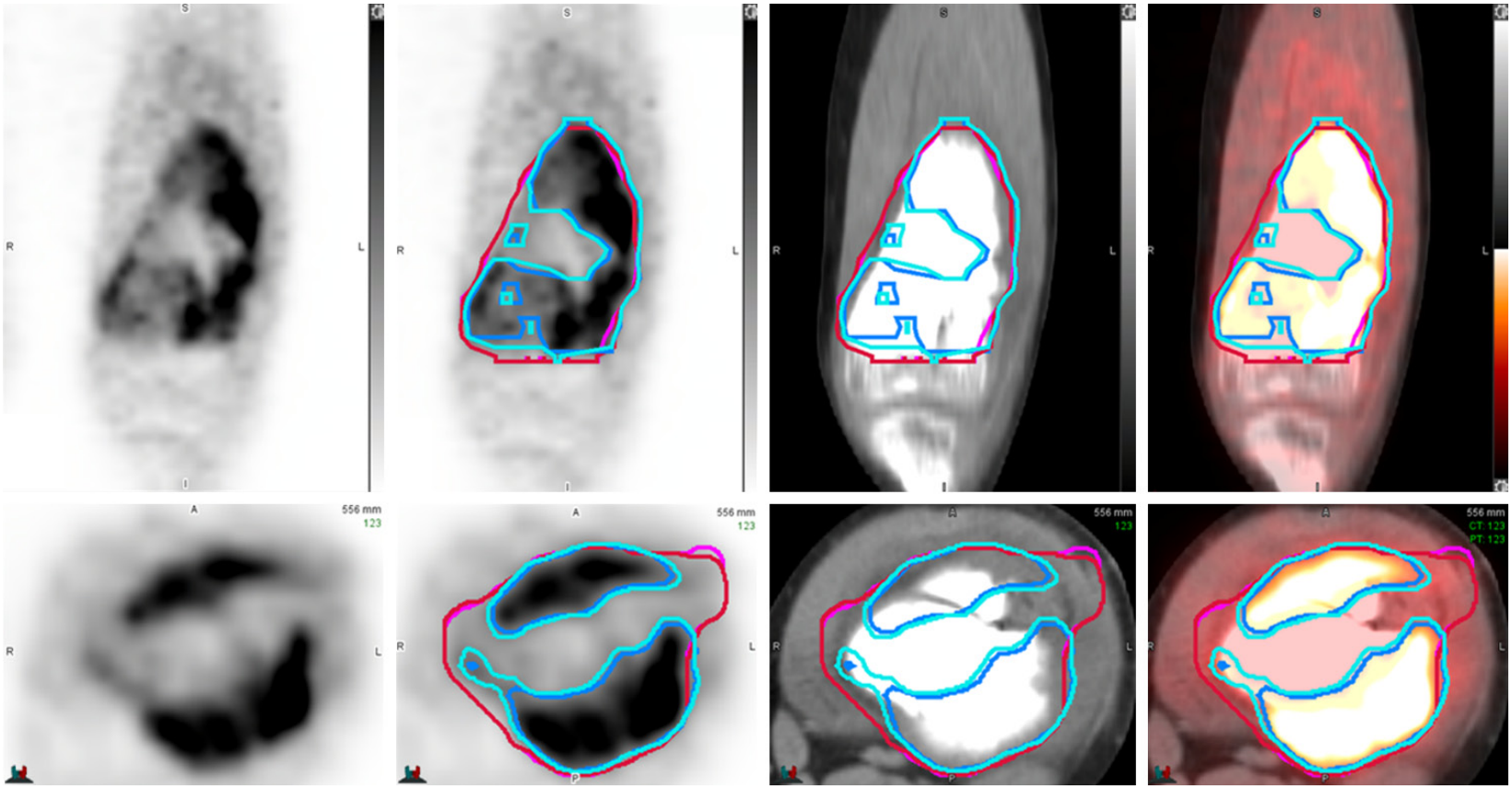

Figure 5. MO-PET and PETedge segmentation in a patient with osteosarcoma. Repeated independent measurements using MO-PET (blue and light blue colored ROIs) and PETedge (red and pink colored ROIs) showed excellent intra-method agreements with the minor discrepancy. Note that the tumor portion with a low FDG uptake was included in PETedge segmentation but excluded in MO-PET segmentation.

Table 2. Intra-class correlation and Bland-Altman analysis of measurements using MO-PET and PETedge segmentation

Segmentation method	Parameter	Intra-class correlation analysis		Bland-Altman analysis			
		ICC	95% CI	Bias (%)	SD (%)	Lower limit (%)	Upper limit (%)
MO-PET vs. MO-PET	MTV	0.994	0.989-0.997	1.1	11.9	-22.2	24.3
	SUVmean	0.998	0.997-0.999	-0.6	3.4	-7.2	5.9
	TLG	0.999	0.998-0.999	0.4	8.6	-16.5	17.3
PETedge vs. PETedge	MTV	0.997	0.993-0.998	-0.3	12.3	-24.5	23.9
	SUVmean	0.996	0.992-0.998	-0.1	5.5	-10.9	10.7
	TLG	0.999	0.998-0.999	-0.4	6.6	-16.3	15.6
MO-PET vs. PETedge	MTV	0.754	-0.014-0.914	-60.2	50.1	-110.2	-10.1
	SUVmean	0.795	-0.055-0.935	27.9	32.1	-4.3	60.0
	TLG	0.956	0.749-0.985	-34.5	31.3	-65.8	-3.1

ICC = Intra-class correlation coefficient.

SUVmean, TLG using both segmentation methods showed minimal bias and narrow upper and lower limits (**Figure 6A**, **6B** and **Table 2**).

Agreement between MTV (MO-PET) and MTV (PETedge) was also evaluated. Two methods to measure MTV showed a moderate level of agreement with ICC of 0.759. Bland-Altman analysis showed a substantial bias [MTV (MO-PET)-MTV (PETedge)] of -60.2% (upper and lower limit = -110.2%/-10.1%, respectively), indicating that MTV is measured larger by PETedge than by MO-PET. Also, SUVmean, TLG between the two segmentation methods showed substantial bias (**Figure 6C**; **Table 3**). The bias is partly caused by the difference in managing inner structure between PETedge and MO-PET. PETedge does not allow hollow inner structure while MO-PET does (**Figure 5**).

Predictive value of MTV (MO-PET) and MTV (PETedge) for EFS in patients with osteosarcoma

Finally, the predictive values of MTV (MO-PET) and MTV (PETedge) at baseline for EFS were evaluated in patients with osteosarcoma. Predictive values of MTVs using conventional threshold methods were extensively evaluated in our previous study using the same patient data [20], thus we evaluated only MTV (MO-PET) and MTV (PETedge) in this study. In ROC analysis, both MTV (MO-PET) and MTV (PETedge) could predict the event at two year with areas under curve (AUC) of 0.771 (0.595-0.897, P = 0.002) and 0.781 (0.610-0.907, P = 0.0007), respectively. However, there was no statistically significant difference between the two ROC curves (P = 0.77) (**Figure 7A**). The patients were further divided into two groups (High MTV and Low MTV groups) using an optimized cut-offs of MTV (MO-PET) and MTV (PETedge), respectively. The numbers of patients in high and low MTV groups were fifteen (44%) and nineteen (56%), respectively by both MTV (MO-PET) and MTV (PETedge). High MTV group by MO-PET showed worse prognosis than low MTV group with a hazard ratio of 6.07 (log rank P = 0.0003, **Figure 7B**). Similarly, high MTV group by PETedge demonstrated worse prognosis than low MTV group with a hazard ratio of 5.55 (log rank P = 0.0007) (**Table 3**; **Figure 7C**). Thus both MTV (MO-PET) and MTV (PETedge) were predictive of EFS, and MTV (MO-PET) showed higher hazard ratio than MTV (PETedge) although the difference was not statistically significant.

Four patients (11.8%) were classified differently by MTV (MO-PET) and MTV (PETedge). More specifically, the two patients were in the high MTV group by MTV (MO-PET) but in the low MTV group by MTV (PETedge), and the other two patients in reverse. When we further compared the EFS between high MTV group by both methods [high MTV (MO-PET and PETedge)] and high MTV group by only one method [high MTV (MO-PET or PETedge)] and low MTV group by both methods [low MTV (MO-PET and PETedge)], we found that the EFS could be stratified according to the three groups (**Figure 7D**, log rank test for trend P = 0.0002). Low MTV

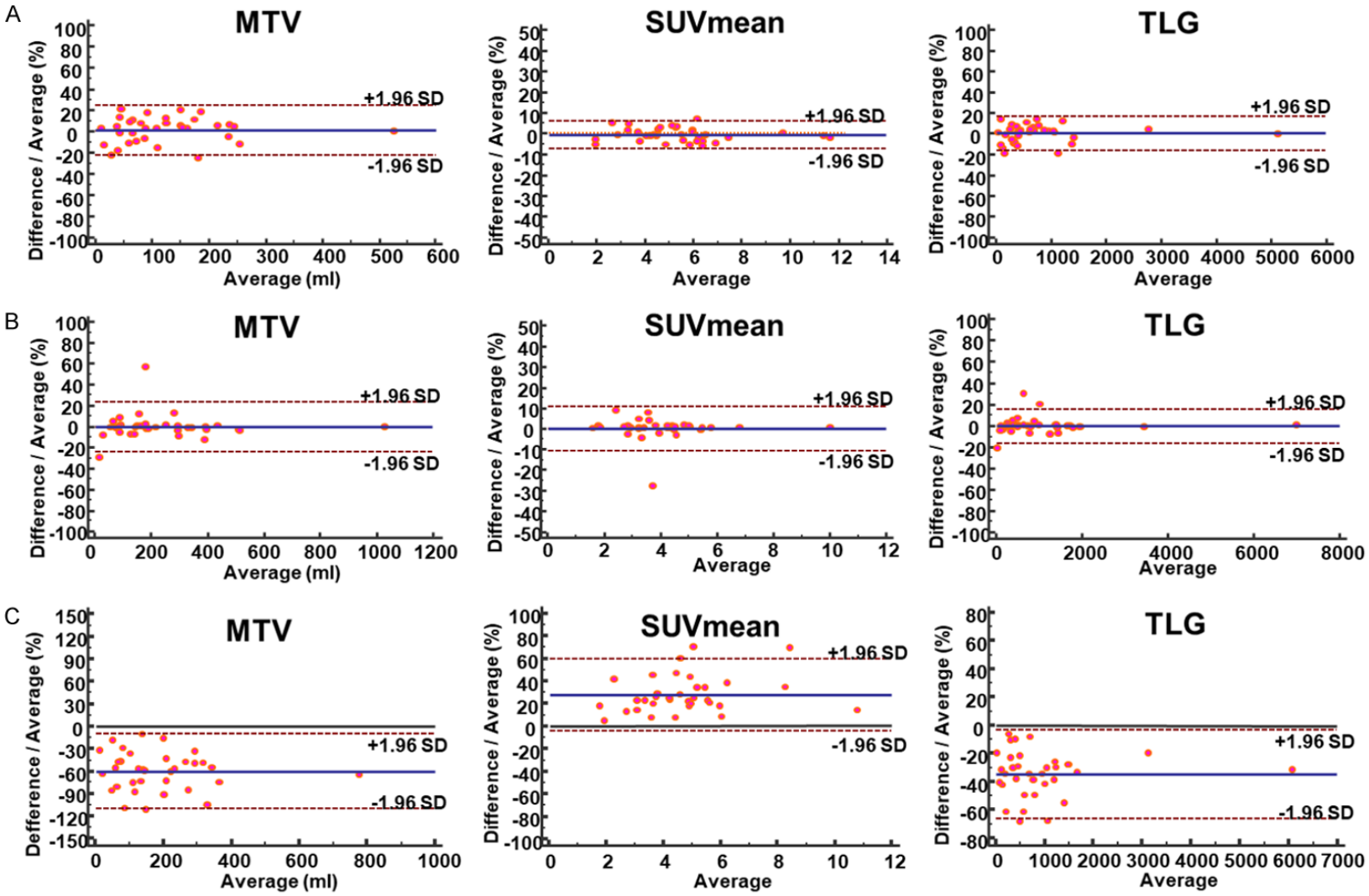

Figure 6. Bland-Altman plot for repeated measurement using MO-PET and PETedge. Repeated measurements of MTV, SUVmean, and TLG using (A) MO-PET and (B) PETedge showed excellent agreements. (C) However, measurements of MTV, SUVmean, and TLG between MO-PET and PETedge showed substantial variation.

Table 3. Event free survival time and hazard ratio

Classifier	Group	Mean survival time (Day) (95% CI)	HR* (95% CI)
MO-PET	Low MTV (n = 19)	1480 (1239-1722)	-
	High MTV (n = 15)	657 (355-959)	6.07 (2.14-17.20)
PETedge	Low MTV (n = 19)	1493 (1223-1763)	-
	High MTV (n = 15)	649 (366-932)	5.55 (1.99-15.50)
Combined MO-PET and PETedge	Low MTV (MO PET and PETedge) (n = 17)	1525 (1285-1764)	-
	High MTV (MO PET or PETedge) (n = 4)	1036 (271-1802)	3.78 (0.74-19.22)
	High MTV (MO PET and PETedge) (n = 13)	605 (325-885)	7.73 (2.55-23.46)

* = Hazard ratio compared to low MTV group, Low MTV (MO PET and PETedge) = patient group with low MTV by MTV (MO-PET) and MTV (PETedge), high MTV (MO-PET or PETedge) = patient group with high MTV by either MTV (MO-PET) or MTV (PETedge), high MTV (MO-PET and PETedge) = patient group with high MTV by both MTV (MO-PET) and MTV (PETedge).

(MO-PET and PETedge) group showed the longest EFS followed by high MTV (MO-PET or PETedge) group and high MTV (MO-PET and PETedge) group (**Table 3**).

Discussion

In our study, MO-PET demonstrated reliable lesion segmentation across the different lesion sizes, LBRs using NEMA IQ phantom. Reproducibility of MO-PET segmentation was further tested in the patients with osteosarcoma and showed an excellent agreement between repeated measurements. Furthermore, MTV (MO-PET) was predictive of EFS in patients with osteosarcoma.

Conventionally, fixed threshold methods are used for measuring MTV. Fixed absolute thresholds are commonly used for measuring MTV, and SUV of 2.0~5.0 has been reported as a potential absolute threshold. The fixed absolute threshold is easy to use but has several limitations. In particular, MTV cannot be measured in a certain tumor with lower uptake than the fixed absolute threshold. Also, a tumor with very high uptake can be overestimated by partial volume averaging effect. Fixed relative thresholds such as 40%, 42%, 50% of SUVmax of the tumor have also been widely used [11]. However, a tumor with heterogeneous uptake or very high SUVmax could be underestimated by a fixed relative threshold. Also, a small lesion with low LBR can be overestimated by fixed relative threshold method. In our NEMA phantom study, fixed threshold methods were not able to accurately segment the lesions across the different sizes and LBRs. Biehl et al. also reported that the optimal threshold increases as the size of the lesion becomes smaller [27]. Thus algorithm based tumor segmentation methods have been developed to overcome the shortcoming of fixed threshold methods and have undergone clinical evaluation [28].

Gradient-based methods are the most extensively validated algorithm based segmentation method in FDG PET images and define the tumor boundary by exploiting the image gradient that exists between the high SUV in tumor and the lower SUV in adjacent non tumor tissues [29]. Also, a gradient-based method such as PETedge has been shown to outperform conventional threshold methods in phantom studies [30] and for resected cancer specimens [31, 32]. The main benefit of gradient-based methods is that they are not dependent on absolute tumor uptake levels. However, gradient-based methods can be sensitive to the reconstruction parameters of the PET images. Another disadvantage of the method is the assumption of uniform contrast around tumor edges. Furthermore, PETedge only draws the outer boundary of the tumor and does not allow segmentation of inner hollow structures, which can be a limitation for measuring tumors with atypical shapes.

MO-PET has been developed using the multi-level Otsu method to segment tumor lesion in FDG PET images [21]. In the first validation of MO-PET using FDG PET image of patients with melanoma, twenty-five tumors of varying sizes and SUVs were segmented using MO-PET and compared to 7 different conventional segmentation methods. MO-PET showed more stable and consistent tumor delineation than the other conventional segmentation methods when CT based tumor volumes were used as a

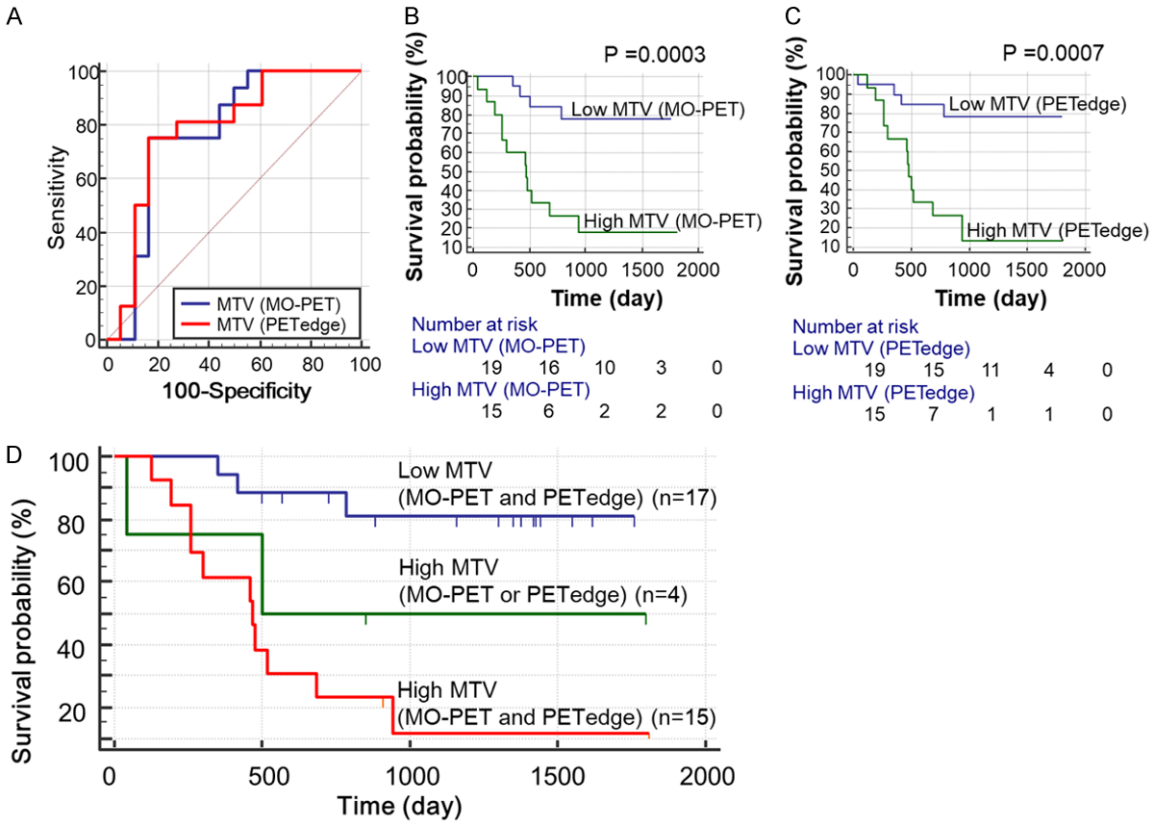

Figure 7. Prediction of event free survival (EFS) using MTV (MO-PET) and MTV (PETedge). (A) ROC curves of MTV (MO-PET) and MTV (PETedge) for prediction of the event at two years. MTV (MO-PET) and MTV (PETedge) showed the similar area under curve (AUC) of 0.771, 0.781 respectively for prediction of the event at two years. Kaplan-Meyer curves between low and high MTV groups classified by (B) MTV (MO-PET) and (C) MTV (PETedge). (D) EFS was longest in low MTV group followed by high MTV (MO-PET or PETedge) group and high MTV (MO-PET and PETedge) group (P for trend = 0.0002). Low MTV (MO PET and PETedge) = patient group with low MTV by MTV (MO-PET) and MTV (PETedge), high MTV (MO-PET or PETedge) = patient group with high MTV by either MTV (MO-PET) or MTV (PETedge), high MTV (MO-PET and PETedge) = patient group with high MTV by both MTV (MO-PET) and MTV (PETedge).

reference standard [23]. Also, the accuracy of MO-PET has been tested using FDG PET images of soft tissue sarcoma. The primary tumor was segmented using MO-PET and six different conventional threshold methods, and MO-PET showed the highest agreement with tumor volumes measured from magnetic resonance imaging (MRI) (ICC = 0.93) [22]. However, it is challenging to define a true reference standard tumor volume in clinical FDG PET images. Therefore, our present study utilized the NEMA phantom to evaluate the reliability of the MO-PET segmentation since we know the actual volumes of lesions in the phantom. In our phantom study, MO-PET showed excellent agreements (ICC over 0.9) with actual volumes of the lesions with a lesion to background ratio (LBR) of 1:8, 1:4, and 1:1.5. The segmentation accuracy of MO-PET was better than the other conventional fixed threshold methods and comparable to that of PETedge.

Although PETedge and MO-PET both showed accurate lesion segmentation in NEMA IQ phantoms, the two methods display a fundamental difference in the segmentation of tumor with high heterogeneity. Since PETedge only draws the outer contour of the tumor, internal structure with minimal FDG uptake is also included in the measured MTV. On the other hand, MO-PET segments the tumor with an optimized threshold using the multi-level Otsu method, internal tissue with minimal FDG uptake can be excluded (**Figure 5**). In primary osteosarcoma, tumors which have highly heterogenous FDG uptakes, we found that MTV (PETedge) was significantly larger than MTV (MO-PET) as expected.

Both MTV (MO-PET) and MTV (PETedge) were similarly predictive of EFS when the patients were classified into two groups (low MTV and high MTV groups) using optimized cut-offs (log rank P = 0.0003 and 0.0007, respectively). In our previous study using the same patient data, the majority of MTVs using conventional threshold methods at baseline scan were also similarly predictive of EFS with log rank *P* values range of 0.0003~0.145 and median of 0.005 [20]. However, given the fact that MTV using conventional threshold methods is not reliable across the different LBRs and lesion sizes, shown in our phantom study results, MTV (MO-PET) or MTV (PETedge) is superior to conventional threshold methods to make the prediction model for clinical outcome using FDG PET/CT scans. The predictive values of various FDG PET parameters including SUVmax, SUVpeak, MTV and TLG using fixed threshold methods for EFS were not presented in our present study since the results were published in our previous paper [20].

Using both MTV (MO-PET) and MTV (PETedge) could further stratify the prognosis in patients with osteosarcoma which indicates a potential additive prognostic value of using both methods. Further larger prospective studies are warranted to evaluate the clinical utility of using both MTVs measured by MO-PET and PETedge for the prediction of clinical outcome.

There are limitations in our study. The smallest three lesions in the NEMA phantom scan with LBR of 1.5:1 could not be evaluated because the lesions were not discernable in the images. The segmentation accuracy in the clinical PET scans could not be assessed because the gold standard tumor volume at baseline from the tumor specimen could not be obtained since the patients underwent the neoadjuvant chemotherapy before the tumor resection.

Conclusion

MO-PET demonstrated accurate, stable and consistent lesion segmentation across different lesion sizes and LBRs in NEMA PET phantom and excellent reproducibility in FDG PET/CT scans of patients with osteosarcoma. Furthermore, MTV (MO-PET) was predictive of EFS in patients with osteosarcoma. This method may be useful in providing additional prognostic information for patients with osteosarcoma.

Acknowledgements

Hyung-Jun Im was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A-1B03035556), and Ministry of Health and Welfare Korea (HI18C0886). Informed consent was obtained from all individual participants included in the study.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Steve Y Cho, Radiology, University of Wisconsin, Madison, WI, USA; University of Wisconsin Carbone Cancer Center, Madison, WI, USA. Tel: (608) 263-3172; Fax: (608) 265-7390; E-mail: scho@uwhealth.org

References

[1] Na F, Wang J, Li C, Deng L, Xue J and Lu Y. Primary tumor standardized uptake value measured on F18-Fluorodeoxyglucose positron emission tomography is of prediction value for survival and local control in non-small-cell lung cancer receiving radiotherapy: meta-analysis. J Thorac Oncol 2014; 9: 834-842.

[2] Sarker A, Im HJ, Cheon GJ, Chung HH, Kang KW, Chung JK, Kim EE and Lee DS. Prognostic implications of the SUVmax of primary tumors and metastatic lymph node measured by 18F-FDG PET in patients with uterine cervical cancer: a meta-analysis. Clin Nucl Med 2016; 41: 34-40.

[3] Im HJ, Kim TS, Park SY, Min HS, Kim JH, Kang HG, Park SE, Kwon MM, Yoon JH, Park HJ, Kim SK and Park BK. Prediction of tumour necrosis fractions using metabolic and volumetric 18F-FDG PET/CT indices, after one course and at the completion of neoadjuvant chemotherapy, in children and young adults with osteosarcoma. Eur J Nucl Med Mol Imaging 2012; 39: 39-49.

[4] Pak K, Cheon GJ, Nam HY, Kim SJ, Kang KW, Chung JK, Kim EE and Lee DS. Prognostic value of metabolic tumor volume and total lesion glycolysis in head and neck cancer: a systematic review and meta-analysis. J Nucl Med 2014; 55: 884-890.

[5] Fledelius J, Winther-Larsen A, Khalil AA, Hjorthaug K, Frokiaer J and Meldgaard P. Assessment of very early response evaluation with (18)F-FDG-PET/CT predicts survival in erlotinib treated NSCLC patients-A comparison

of methods. Am J Nucl Med Mol Imaging 2018; 8: 50-61.
[6] Vensby PH, Schmidt G, Kjaer A and Fischer BM. The value of FDG PET/CT for follow-up of patients with melanoma: a retrospective analysis. Am J Nucl Med Mol Imaging 2017; 7: 255-262.
[7] Suarez-Pinera M, Belda-Sanchis J, Taus A, Sanchez-Font A, Mestre-Fusco A, Jimenez M and Pijuan L. FDG PET-CT SUVmax and IASLC/ATS/ERS histologic classification: a new profile of lung adenocarcinoma with prognostic value. Am J Nucl Med Mol Imaging 2018; 8: 100-109.
[8] Im HJ, Kim YK, Kim YI, Lee JJ, Lee WW and Kim SE. Usefulness of combined metabolic-volumetric indices of (18)F-FDG PET/CT for the early prediction of neoadjuvant chemotherapy outcomes in breast cancer. Nucl Med Mol Imaging 2013; 47: 36-43.
[9] Lee JW, Kang CM, Choi HJ, Lee WJ, Song SY, Lee JH and Lee JD. Prognostic value of metabolic tumor volume and total lesion glycolysis on preoperative 18F-FDG PET/CT in patients with pancreatic cancer. J Nucl Med 2014; 55: 898-904.
[10] Cheebsumon P, van Velden FHP, Yaqub M, Frings V, de Langen AJ, Hoekstra OS, Lammertsma AA and Boellaard R. Effects of image characteristics on performance of tumor delineation methods: a test-retest assessment. J Nucl Med 2011; 52: 1550-1558.
[11] Moon SH, Hyun SH and Choi JY. Prognostic significance of volume-based PET parameters in cancer patients. Korean J Radiol 2013; 14: 1-12.
[12] Gorlick R, Janeway K, Lessnick S, Randall RL, Marina N; on behalf of the COGBTC. Children's Oncology Group's 2013 blueprint for research: Bone tumors. Pediatr Blood Cancer 2013; 60: 1009-1015.
[13] Bielack SS, Kempf-Bielack B, Delling G, Exner GU, Flege S, Helmke K, Kotz R, Salzer-Kuntschik M, Werner M, Winkelmann W, Zoubek A, Jurgens H and Winkler K. Prognostic factors in high-grade osteosarcoma of the extremities or trunk: an analysis of 1,702 patients treated on neoadjuvant cooperative osteosarcoma study group protocols. J Clin Oncol 2002; 20: 776-790.
[14] Davis AM, Bell RS and Goodwin PJ. Prognostic factors in osteosarcoma: a critical review. J Clin Oncol 1994; 12: 423-431.
[15] Costelloe CM, Macapinlac HA, Madewell JE, Fitzgerald NE, Mawlawi OR, Rohren EM, Raymond AK, Lewis VO, Anderson PM, Bassett RL Jr, Harrell RK and Marom EM. 18F-FDG PET/CT as an indicator of progression-free and overall survival in osteosarcoma. J Nucl Med 2009; 50: 340-347.
[16] Byun BH, Kong CB, Park J, Seo Y, Lim I, Choi CW, Cho WH, Jeon DG, Koh JS, Lee SY and Lim SM. Initial metabolic tumor volume measured by 18F-FDG PET/CT can predict the outcome of osteosarcoma of the extremities. J Nucl Med 2013; 54: 1725-1732.
[17] Franzius C, Bielack S, Flege S, Sciuk J, Jürgens H and Schober O. Prognostic significance of 18F-FDG and 99mTc-methylene diphosphonate uptake in primary osteosarcoma. J Nucl Med 2002; 43: 1012-1017.
[18] Sato J, Yanagawa T, Dobashi Y, Yamaji T, Takagishi K and Watanabe H. Prognostic significance of 18F-FDG uptake in primary osteosarcoma after but not before chemotherapy: a possible association with autocrine motility factor/phosphoglucose isomerase expression. Clin Exp Metastasis 2008; 25: 427-435.
[19] Im H-J, Wu H, Yi Z, Wu J, Shulkin B and Cho S. Baseline metabolic tumor volume measured by FDG PET/CT before neoadjuvant chemotherapy predicts survival in pediatric osteosarcoma. J Nucl Med 2016; 57: 429.
[20] Im HJ, Zhang Y, Wu H, Wu J, Daw NC, Navid F, Shulkin BL and Cho SY. Prognostic value of metabolic and volumetric parameters of FDG PET in pediatric osteosarcoma: a hypothesis-generating study. Radiology 2018; 287: 303-312.
[21] Cho S, Solaiyappan M and Huang E. Multi-level otsu for positron emission tomography (mopet). Google Patents 2016.
[22] Lee I, Im HJ, Solaiyappan M and Cho SY. Comparison of novel multi-level Otsu (MO-PET) and conventional PET segmentation methods for measuring FDG metabolic tumor volume in patients with soft tissue sarcoma. EJNMMI Phys 2017; 4: 22.
[23] Huang E, Solaiyappan M and Cho S. Improved stability and performance of 18F-FDG PET automated tumor segmentation using multi-level maximization of inter-class variance method. J Nucl Med 2015; 56: 452.
[24] Davis JC, Daw NC, Navid F, Billups CA, Wu J, Bahrami A, Jenkins JJ, Snyder SE, Reddick WE, Santana VM, McCarville MB, Guo J and Shulkin BL. FDG uptake during early adjuvant chemotherapy predicts histologic response in pediatric and young adult patients with osteosarcoma. J Nucl Med 2017; 59: 25-30.
[25] Navid F, Santana VM, Neel M, McCarville MB, Shulkin BL, Wu J, Billups CA, Mao S, Daryani VM, Stewart CF, Kunkel M, Smith W, Ward D, Pappo AS, Bahrami A, Loeb DM, Reikes Willert J, Rao BN and Daw NC. A phase II trial evaluating the feasibility of adding bevacizumab to standard osteosarcoma therapy. Int J Cancer 2017; 141: 1469-1477.

[26] Hurley C, McCarville MB, Shulkin BL, Mao S, Wu J, Navid F, Daw NC, Pappo AS and Bishop MW. Comparison of (18) F-FDG-PET-CT and bone scintigraphy for evaluation of osseous metastases in newly diagnosed and recurrent osteosarcoma. Pediatr Blood Cancer 2016; 63: 1381-1386.

[27] Biehl KJ, Kong FM, Dehdashti F, Jin JY, Mutic S, El Naqa I, Siegel BA and Bradley JD. 18F-FDG PET definition of gross tumor volume for radiotherapy of non-small cell lung cancer: is a single standardized uptake value threshold approach appropriate? J Nucl Med 2006; 47: 1808-1812.

[28] Im HJ, Bradshaw T, Solaiyappan M and Cho SY. Current methods to define metabolic tumor volume in positron emission tomography: Which one is better? Nucl Med Mol Imaging 2018; 52: 5-15.

[29] Graves EE, Quon A and Loo BW Jr. RT_image: an open-source tool for investigating PET in radiation oncology. Technol Cancer Res Treat 2007; 6: 111-121.

[30] Werner-Wasik M, Nelson AD, Choi W, Arai Y, Faulhaber PF, Kang P, Almeida FD, Xiao Y, Ohri N, Brockway KD, Piper JW and Nelson AS. What is the best way to contour lung tumors on PET scans? Multiobserver validation of a gradient-based method using a NSCLC digital PET phantom. Int J Radiat Oncol Biol Phys 2012; 82: 1164-1171.

[31] Geets X, Lee JA, Bol A, Lonneux M and Gregoire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging 2007; 34: 1427-1438.

[32] Sridhar P, Mercier G, Tan J, Truong MT, Daly B and Subramaniam RM. FDG PET metabolic tumor volume segmentation and pathologic volume of primary human solid tumors. AJR Am J Roentgenol 2014; 202: 1114-1119.