ORIGINAL ARTICLE

Intra-reader reliability of FDG PET volumetric tumor parameters: effects of primary tumor size and segmentation methods

B. Shah · N. Srivastava · A. E. Hirsch ·
G. Mercier · R. M. Subramaniam

Abstract

Objective To establish the effects of size and segmentation methods on intra-reader reliability of primary tumor metabolic tumor volume (MTV) and total glycolytic activity (TGA) in human solid tumors.

Methods This is a retrospective study of 121 patients who had a baseline FDG PET/CT scan for oncologic staging. Volumetric parameter readings were performed in random order on two separate occasions, 12 weeks apart, by the same reader. The MTV and TGA were segmented using gradient and fixed maximum standardized uptake value (SUV_{max}) threshold methods. Intra-reader reliability was established by the intraclass correlation coefficient (ICC) and Bland–Altman analysis.

Results The biases for MTV were 2.95, 14.76 and 11.13 % for gradient segmentation, 38 and 50 % SUV_{max} fixed threshold segmentations, respectively ($p < 0.0001$). For TGA, the corresponding biases were 0.76, 10.36 and 7.46 % ($p < 0.0001$). There were no statistically significant differences in the biases between the first and second reads for MTV segmented for small and large volume tumors by the gradient method ($p < 0.34$) or 50 % SUV_{max} threshold segmentation ($p < 0.08$). However, there were statistically significant differences in the corresponding biases for the 38 % SUV_{max} threshold segmentation ($p < 0.04$). There were no statistically significant differences in the biases between the first and second reads for TGA segmented for small and large volume tumors ($p < 0.98$).

Conclusion Intra-reader reliability for primary tumor FDG MTV and TGA is affected by the tumor size and segmentation methods. The segmentation bias was smaller for gradient method than percentage fixed threshold method for MTV. The segmentation biases were smaller for TGA than MTV.

Keywords Metabolic tumor volume ·
Total glycolytic activity · Reader reliability

B. Shah · N. Srivastava · G. Mercier · R. M. Subramaniam
Department of Radiology, Boston University School of
Medicine, Boston, MA, USA

A. E. Hirsch · R. M. Subramaniam
Department of Radiation Oncology, Boston University School
of Medicine, Boston, MA, USA

R. M. Subramaniam (✉)
Russel H. Morgan Department of Radiology and Radiological
Science, Johns Hopkins Medical Institutions, 601 N. Caroline
Street/JHOC 3235, Baltimore, MD 21287, USA
e-mail: rsubram4@jhmi.edu

Introduction

Imaging biomarkers must be objective, quantitative, accurate and reproducible to be useful in the era of molecular and personalized medicine [1]. The ^{18}F-fluoro-deoxyglucose positron emission tomography (FDG PET) maximum standardized uptake value (SUV_{max}) has been used as a prognostic, predictive and sometimes diagnostic imaging parameter in oncology [2–7]. However, SUV_{max}, representing single pixel metabolic information about the tumor, may not accurately represent the tumor volume or biology. Imaging parameters that measure FDG metabolic tumor volume (MTV) and total glycolytic activity (TGA) of the tumor are emerging as exploratory imaging biomarkers for prediction of outcome and therapy response in patients with solid tumors [8–14], adding value to current staging prognoses [15].

Studies have demonstrated that tumor volume segmentation depends on the tumor to background ratio, image

resolution, image noise, the size of the tumor and the segmentation method [16–18]. Manual contouring of tumor volume delineation depends on the experience of the reader and leads to high variability [19]. Semiautomatic segmentation methods have less variability than manual methods [11]. In addition, previous studies of intra-reader and inter-reader reliability of MTV using semiautomatic segmentation methods are mostly limited to phantoms [16]. The intra-reader and inter-reader reliability of TGA, to our knowledge, has not been studied.

The objective of this exploratory study is to establish the effects of primary tumor size and segmentation methods on intra-reader reliability of primary tumor MTV and TGA in head and neck, lung, and breast cancers.

Methods

Patients and study design

This study is a retrospective evaluation of positron emission tomography/computed tomography (PET/CT) images. Approval from our institutional review board was granted with a waiver of informed consent. All patients with lung, head and neck, and breast cancers, who had a baseline PET/CT at our institution in 2009 were included in the study. These cancer types were selected since the tumor and background tissue FDG uptakes provide a mixture of signal-to-noise characteristics representative of day-to-day clinical practice. Patients who had any local or systemic therapy were excluded. The study population included 121 patients (mean age of 60.9 ± 12.7 years).

PET/CT

All PET/CT studies were performed on a GE Discovery STE 16 (GE Healthcare, Milwaukee, WI, USA) PET/CT scanner according to the institutional standard clinical protocol. A dedicated head and neck protocol was instituted for all head and neck cancer patients. For patients with oral and oropharyngeal cancers, the scans were done from skull base to aortic arch with the arms down and clavicle to mid-thigh with the arms up. For patients with breast and lung cancers, the scans were done from skull base to mid-thigh with the arms up. The average patient blood glucose level was 102.5 mg/dl (SD 25.0). Patients were injected with a mean dose of 13.4 mCi (SD 3.1) of FDG and incubated for an average period of 92.9 min (SD 25.3).

When intravenous contrast was used ($n = 66$, 54.5 %), torso (mid-thigh to clavicle) and then neck (skull base to clavicle) were imaged regardless of the location of the primary tumor. About 100 ml of Optiray IV (Tyco

Healthcare/Mallinckrodt, Hazelwood, MO, USA) with a 30 ml saline chaser was injected using a power injector (GE Healthcare, Milwaukee, WI, USA) at 3 ml/s for torso and a second bolus of 60 ml of Optiray IV (Tyco Healthcare/Mallinckrodt, Hazelwood, MO, USA) was used for head and neck images.

Image analysis

All PET/CT studies were retrieved from the electronic archival system and reviewed on a MIMvista workstation (software version 4.1) (MIM Software Inc., Cleveland, OH, USA). To establish the intra-reader reliability, 1 author (RS), with board certification in radiology and nuclear medicine fellowship and 3 years of experience as faculty, was chosen. There was a 12-week interval between the first and second image analyses. The images were read in random order to minimize bias.

PET, CT, and fused PET/CT images were reviewed in axial, coronal, and sagittal planes. For the purposes of this study, the relevant imaging parameter measurements were the primary tumor MTV and TGA segmented from PET. MTV was defined as the tumor volume with FDG uptake segmented by a gradient-based (PET edge) method or a fixed threshold method at 38 and 50 % of SUV_{max}. The TGA was defined as $(MTV) \times (SUV_{mean})$. The commercially available MIMvista software analysis suite (MIM Software Inc., Cleveland, OH, USA) includes a contouring suite for radiation therapy planning and a PET/CT fusion suite. Once the primary tumor (target) was segmented, MTV and TGA were automatically calculated by the MIMvista software.

Segmentation methods

There are many methods of segmentation proposed for PET tumor volumetric measurements [20–23]. The gradient and threshold segmentation methods of volume measurement available in MIMvista software have been previously described [15, 16]. The gradient segmentation method in the MIMvista software relies on an operator-defined starting point near the center of the lesion. As the operator drags the cursor out from the center of the lesion, six axes extend out, providing visual feedback for the starting point of gradient segmentation. Spatial gradients are calculated along each axis interactively, and the length of an axis is restricted when a large gradient is detected along that axis. The six axes define an ellipsoid that is then used as an initial bounding region for gradient detection.

The fixed SUV_{max} threshold contouring method relies on including all voxels that are greater than a defined percent of the maximum voxel within an operator-defined sphere (in this study 38 and 50 %). We used 38 and 50 %

SUV$_{max}$ as it has been previously reported. Cross-sectional circles are displayed in all 3 projections (axial, sagittal, and coronal) to ensure three-dimensional coverage of the primary tumor. The edges of the primary tumor are automatically calculated and outlined in both segmentation methods.

Statistical methods

We present our summary statistics as the mean ± standard deviation for continuous variables, or frequency and percentage for categorical variables. Reproducibility of MTV and TGA was measured using the intraclass correlation coefficient (ICC) as generated by a two-way mixed effects model and reported as a point estimate with a 95 % confidence interval (95 % CI). The ICC ranges between 0 and 1.00, with values closer to 1.00 representing better intra-reader reliability. ICC is interpreted as follows: 0.00–0.20 indicates slight reproducibility, 0.21–0.40 indicates fair reproducibility, 0.41–0.60 indicates moderate reproducibility, 0.61–0.80 indicates substantial reproducibility, and >0.80 indicates almost perfect reproducibility [24]. We also investigated the reliability with Bland–Altman analyses [25].

To investigate the impact of partial volume effect on intra-reader reliability, we divided the cohort into small versus large volume tumors with a mean MTV cut point of 10 ml ($n = 69$, 10 ml or greater) (Figs. 1, 2). Thus, the cut point of 10 ml was chosen as the partial volume effect is greater for tumors <2 cm. Between-group analyses were performed using a t test and analysis of variance (ANOVA) with Friedman test and post hoc multiple pair comparisons, as appropriate. We used the Prism 5 (GraphPad Software Inc, San Diego, CA, USA) and SPSS 19 (SPSS Inc, Chicago, IL, USA) statistical packages for all analyses, and all hypothesis tests were two-sided with a significance level of 0.05.

Results

Patients

There were 67 women (55.4 %) and 54 men (44.6 %) in the study. The American Joint Committee on Cancer (AJCC 6th edition) stage of the cancers included 15 stage I patients (12.4 %), 24 stage II patients (19.8 %), 40 stage III patients (33.1 %), and 42 stage IV patients (34.7 %). Patient demographics are summarized in Table 1.

Intra-reader MTV reliability

For MTV, the ICCs were 0.97 for the gradient method, 0.98 for the 38 % threshold, and 0.99 for the 50 % SUV$_{max}$

threshold method (p value <0.001 for all point estimates). The Bland–Altman analysis showed a bias of 2.95 % (SD 32.5 %) for gradient, 14.76 % (SD 34.5 %) for the 38 % SUV$_{max}$, and 11.13 % (SD 31 %) for the 50 % SUV$_{max}$ threshold MTV segmentations ($p < 0.0001$).

For small lesions (<10 ml of mean MTV), the biases were −1.53 % (SD 40 %) for gradient, 22.12 % (SD 46.5 %) for the 38 %, and 16.4 % (SD 42.2 %) for the 50 % SUV$_{max}$ threshold MTV segmentations ($p < 0.0006$).

For large lesions (>10 ml of mean MTV), biases were 3.9 % (SD 21.9 %) for gradient, 9.32 % (SD 20.4 %) for the 38 %, and 6.55 % (SD 17.9 %) for the 50 % SUV$_{max}$ threshold MTV segmentations ($p < 0.01$) (Table 2; Fig. 3). There were no statistically significant differences in the biases between the first and second readings for MTV segmented into small and large volume tumors by the gradient method ($p < 0.34$) or 50 % SUV$_{max}$ method ($p < 0.08$). However, there were statistically significant differences in the corresponding biases for the 38 % SUV$_{max}$ threshold segmentation ($p < 0.04$) (Fig. 4).

Intra-reader TGA reliability

The ICC for gradient, 38 %, and 50 % SUV$_{max}$ threshold TGA was 0.99 for all methods ($p < 0.001$ for all point estimates). The Bland–Altman analysis showed a bias of 0.76 % (SD 30.3 %) for gradient, 10.36 % (SD 37.9 %) for the 38 % SUV$_{max}$, and 7.46 % (SD 34.6) for the 50 % SUV$_{max}$ threshold TGA segmentations ($p < 0.03$).

For lesions <10 ml of mean MTV, the biases were −4.26 % (SD 39.8 %) for gradient, 16 % (SD 46.4 %) for the 38 % SUV$_{max}$, and 8 % (SD 49.6 %) for the 50 % SUV$_{max}$ threshold segmentations ($p < 0.0005$).

For lesions >10 ml of mean MTV, the biases were 2.42 % (SD 17.8 %) for gradient, 6 % (SD 29.8 %) for the 38 % SUV$_{max}$, and 6.32 % (SD 17 %) for the 50 % SUV$_{max}$ threshold TGA segmentations ($p < 0.06$). There were no statistically significant differences in the biases between the first and second reads for TGA segmented for small and large volume tumors by the gradient method ($p < 0.22$), the 38 % SUV$_{max}$ ($p < 0.18$) or the 50 % SUV$_{max}$ threshold segmentation ($p < 0.98$).

Discussion

In this study, we sought to assess the intra-reader variability of FDG metabolic parameters (primary tumor metabolic tumor volume and total glycolytic activity) using two different semi-automatic segmentation methods. Our results show that overall bias was generally minimal for gradient threshold MTV and TGA, followed by the 50 % SUV$_{max}$ and then the 38 % SUV$_{max}$ threshold
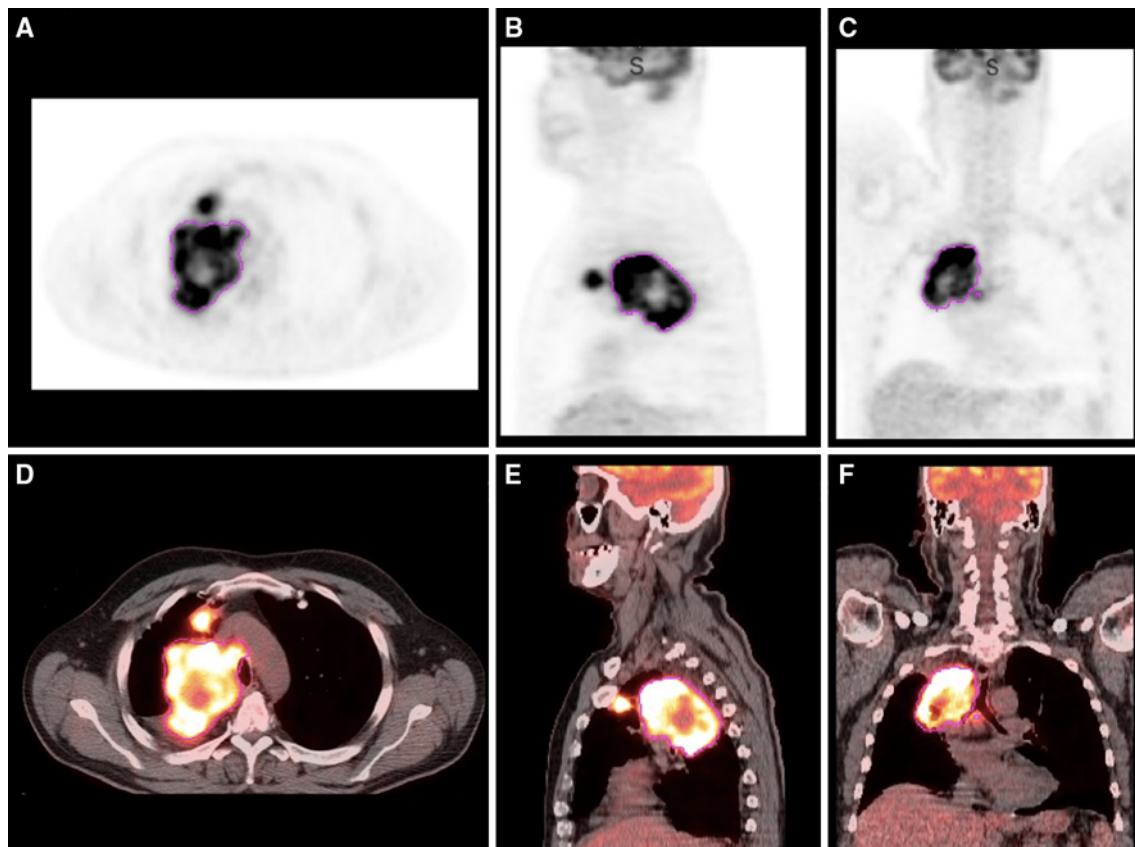
Fig. 1 Large tumor volumetric segmentation and reader reliability: a 60-year-old male with a large stage IV tumor in the upper lobe of right lung. **a** Axial, **b** sagittal, **c** coronal PET and corresponding **d**, **e** and **f** fused PET/CT images. The metabolic volumetric contours demonstrated in this figure were drawn using gradient segmentation. The first and second read volumes were 328 and 311 ml. For 38 % SUV_{max} segmentation, the corresponding first and second read volumes were 259 and 247 ml and for 50 % SUV_{max} segmentation, volumes were 168 and 155 ml

segmentations. We further demonstrate that the biases for MTV for smaller primary tumors are significantly greater for the 38 and 50 % fixed SUV_{max} threshold segmentation methods than for the gradient segmentation method. No significant difference in bias was noted for segmentation of TGA between small and large volume tumors using the gradient and fixed threshold methods, though the bias for gradient segmentation is less than for threshold segmentation.

New imaging parameters need to progress as quantitative biomarkers for therapy response assessment and for patient outcome. The reproducibility, reader reliability and variability must be established for emerging parameters to become clinically useful biomarkers. The small intra-reader variability in this study demonstrated for the gradient method is likely due to the semiautomated feature of the software algorithm. For the gradient method segmentation, the reader needs to point in the center of the tumor and drag towards the periphery as the software program creates six radii to identify the edge of the tumor. This reader intervention is likely the cause of

variability between the reads. There was no statistically significant difference in the reader bias for TGA between the small and large tumors while there was difference in MTV. Since TGA is calculated by multiplication of MTV and SUV_{mean}, bias effects may be reduced in TGA. It is also important for the results to be generalizable as the FDG uptake in the background tissue can vary and may affect the segmentation. Our study included head and neck, lung and breast tumors as the background uptake in normal lung, breast tissue and neck tissue can be variable.

The gradient method is a function of local relative change in image count levels at the tumor-normal tissue boundary. It is not sensitive to varying background counts as it depends on the location rather than the magnitude of the count change [16]. Previous phantom studies have demonstrated, for segmentation of MTV, that gradient segmentation is more accurate than threshold segmentation (using SUV_{max} of 45 %) for spheres <20 mm ($p < 0.015$) [16]. These authors further demonstrated that inter-observer variability is significantly reduced when gradient
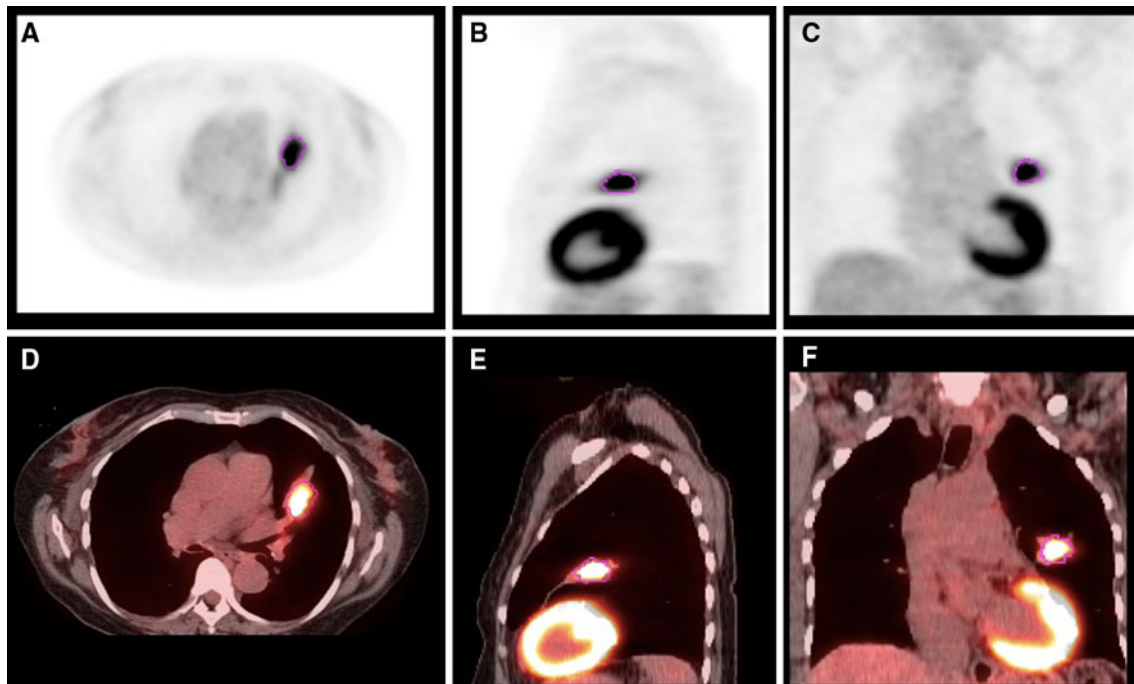
Fig. 2 Small tumor volumetric segmentation and reader reliability: a 63-year-old female with a small, stage IB tumor in the upper lobe of left lung. **a** Axial, **b** sagittal, **c** coronal PET and corresponding **d**, **e** and **f** PET/CT. The metabolic volumetric contours demonstrated in this figure were drawn using gradient segmentation. The first and second read volumes were 6.7 and 13 ml. For 38 % SUV$_{max}$ segmentation, the corresponding first and second read volumes were 7.9 and 8.3 ml and for 50 % SUV$_{max}$ segmentation, volumes were 6.2 and 4.7 ml

Table 1 Characteristics of the patients included in the study

Characteristics	$n = 121$
Age (mean ± SD), years	
Men	57.8 ± 13.4
Women	64.7 ± 10.8
All patients	60.9 ± 12.7
Sex	
Women	54 (44.6 %)
Men	67 (55.4 %)
Primary tumor site	
Lung	39 (32.2 %)
Breast	37 (30.6 %)
Head and neck	45 (37.2 %)
AJCC stage (6th edition)	
I	15 (12.4 %)
II	24 (19.8 %)
III	40 (33.1 %)
IV	42 (34.7 %)

segmentation is used compared to manual or a 25 % SUV$_{max}$ threshold segmentation.

Tumor size has been shown to have an effect on the reliability of metabolic volume segmentation [16, 26]. The fixed SUV$_{max}$ percentage threshold segmentation depends on the SUV$_{max}$ and the set threshold. There is no best single threshold for all volumes. As smaller volumes require a larger threshold because of partial volume effects, the optimum threshold is a function of volume [16]. Furthermore, fixed segmentation suffers from its binary deterministic nature and lack of robustness for varying tumor to background contrast [18]. Our study confirms that the lower intra-reader reliability seen with smaller volumes in phantom studies is also seen in patients, especially for segmenting MTV.

Our study results need to be interpreted in the context of the study design. As an exploratory study, we used a single reader with 3 years of faculty experience reading PET/CT and 18 months of experience in volumetric segmentations. We have previously demonstrated that reader training and experience have an impact on reader reliability [27]. Further studies are needed to assess the applicability of our results to different malignancies using multiple readers, with varied experience, working with semiautomated segmentation programs, and to establish the intra-and inter-reader reliability of MTV and TGA. The impact of intravenous contrast was not studied as the segmentation methods were based on the PET images and IV contrast has non-significant effect on PET images.

In conclusion, the intra-reader reliability of MTV and TGA depends on the segmentation methods and size of the

Table 2 MTV and TGA—Bland–Altman analysis: effect of metabolic volume of primary tumor and segmentation methods

	Bias (%)	SD (%)	95 % lower bound (%)	95 % upper bound (%)
Gradient MTV				
All tumors	2.95	32.48	−60.71	66.61
Tumors <10 ml	−1.53	40.03	−80.0	76.95
Tumors 10 ml or greater	3.9	21.9	−39.05	46.83
38 % SUV$_{max}$ MTV				
All tumors	14.76	34.5	−52.87	82.39
Tumors <10 ml	22.12	46.47	−69.0	113.2
Tumors 10 ml or greater	9.32	20.42	−30.71	49.35
50 % SUV$_{max}$ MTV				
All tumors	11.13	31.0	−49.6	71.9
Tumors <10 ml	16.4	42.18	−66.25	99.1
Tumors 10 ml or greater	6.55	17.91	−28.57	41.67
Gradient TGA				
All tumors	0.76	30.3	−58.64	60.15
Tumors <10 ml	−4.26	39.8	−82.35	73.82
Tumors 10 ml or greater	2.42	17.8	−32.46	37.31
38 % SUV$_{max}$ TGA				
All tumors	10.36	37.86	−63.86	84.57
Tumors <10 ml	16.03	46.44	−75	107.06
Tumors 10 ml or greater	6.05	29.82	−52.40	64.50
50 % SUV$_{max}$ TGA				
All tumors	7.46	34.57	−60.29	75.21
Tumors <10 ml	8.0	49.58	−89.21	105.15
Tumors 10 ml or greater	6.32	17.0	−27.11	39.76
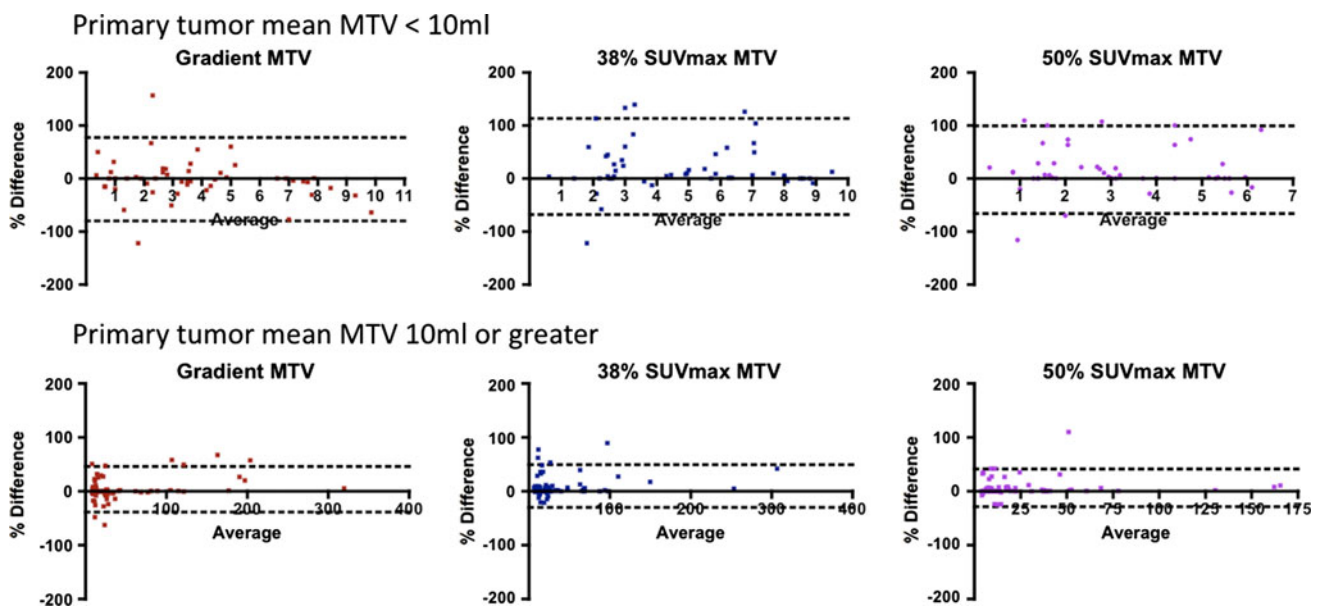
Fig. 3 Intra-reader reliability of segmentation of MTV: patients were divided into those with a mean MTV < 10 ml and 10 ml or greater. The biases were significantly smaller for larger tumors (3.9–9.32 %) than for smaller tumors (−1.53 to 22.12 %)

tumor. Gradient segmentation has less bias than fixed threshold segmentation for MTV, especially for smaller volumetric tumors. Intra-reader variability for TGA is lower than for MTV for both segmentation methods in small and large volume tumors. Further studies are needed to validate these exploratory findings and conclusions.
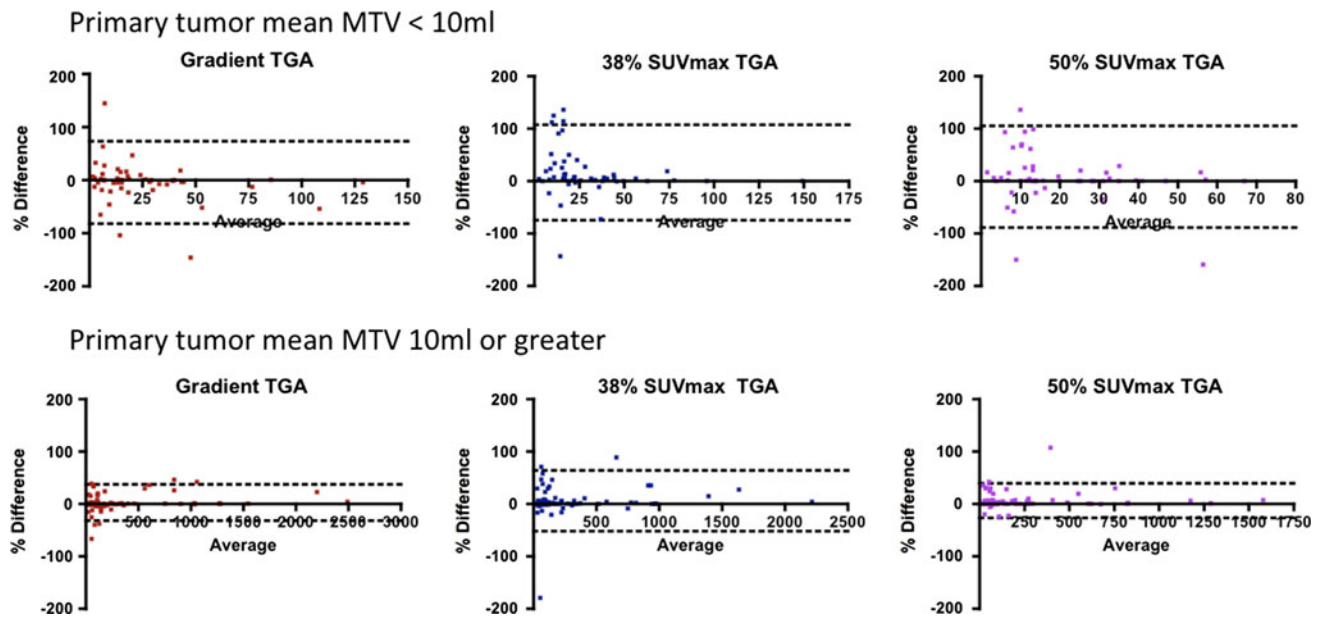
Fig. 4 Intra-reader reliability of segmentations of TGA: patients were divided into those with a mean MTV < 10 ml and 10 ml or greater. The bias was smaller for larger tumors (2.42–6.32 %) than smaller tumors (−4.26 to 16 %), though not statistically significant

References

1. Sullivan DC, Gatsonis C. Response to treatment series: part 1 and introduction, measuring tumor response–challenges in the era of molecular medicine. AJR Am J Roentgenol. 2011;197:15–7.
2. Subramaniam RM, Truong M, Peller P, Sakai O, Mercier G. Fluorodeoxyglucose-positron-emission tomography imaging of head and neck squamous cell cancer. AJNR Am J Neuroradiol. 2010;31:598–604.
3. Davison JM, Ozonoff A, Imsande HM, Grillone GA, Subramaniam RM. Squamous cell carcinoma of the palatine tonsils: FDG standardized uptake value ratio as a biomarker to differentiate tonsillar carcinoma from physiologic uptake. Radiology. 2010;255:578–85.
4. Karantanis D, Bogsrud TV, Wiseman GA, Mullan BP, Subramaniam RM, Nathan MA, et al. Clinical significance of diffusely increased 18F-FDG uptake in the thyroid gland. J Nucl Med. 2007;48:896–901.
5. Karantanis D, Subramaniam RM, Witte RJ, Mullan BP, Nathan MA, et al. 18F-FDG PET/CT in primary central nervous system lymphoma in HIV-negative patients. Nucl Med Commun. 2007;28:834–41.
6. Karantanis D, Subramaniam RM, Peller PJ, Lowe VJ, Durski JM, Collins DA, et al. The value of [(18)F]fluorodeoxyglucose positron emission tomography/computed tomography in extranodal natural killer/T-cell lymphoma. Clin Lymphoma Myeloma. 2008;8:94–9.
7. Imsande HM, Davison JM, Truong MT, Devaiah AK, Mercier G, Ozonoff Al, et al. Use of 18F-FDG PET/CT as a predictive biomarker of outcome in patients with head-and-neck non-squamous cell carcinoma. AJR Am J Roentgenol. 2011;197(4):976–80.
8. Seol YM, Kwon BR, Song MK, Choi YJ, Shin HJ, Chung JS, et al. Measurement of tumor volume by PET to evaluate prognosis in patients with head and neck cancer treated by chemoradiation therapy. Acta Oncol. 2010;49:201–8.
9. Lee HY, Hyun SH, Lee KS, Kim BT, Kim J, Shim YM, et al. Volume-based parameter of 18F-FDG PET/CT in malignant pleural mesothelioma: prediction of therapeutic response and prognostic implications. Ann Surg Oncol. 2010;17:2787–94.
10. Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. Eur J Nucl Med Mol Imaging. 2011;38:663–72.
11. Hatt M, Cheze-Le Rest C, Aboagye EO, Kenny LM, Rosso L, Turkheimer FE, et al. Reproducibility of 18F-FDG and 3′-deoxy-3′-18F-fluorothymidine PET tumor volume measurements. J Nucl Med. 2010;51:1368–76.
12. Hyun SH, Choi JY, Shim YM, Kim K, Lee SJ, Cho YS, et al. Prognostic value of metabolic tumor volume measured by 18F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. Ann Surg Oncol. 2010;17(1):115–22.
13. La TH, Filion EJ, Turnbull BB, Chu JN, Lee P, Nguyen K, et al. Metabolic tumor volume predicts for recurrence and death in head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2009;74:1335–41.
14. Hadiprodjo D, Ryan T, Truong M, Mercier G, Subramaniam R. Parotid gland tumors: preliminary data for the value of FDG PET/CT diagnostic parameters. AJR (in press). 2011.
15. Dibble E, Lara Alvarez A, Truong M, Mercier G, Cook E, Subramaniam RM. FDG metabolic tumor volume and total glycolytic activity: prognostic imaging biomarkers of oral and oropharyngeal squamous cell cancers. J Nucl Med. 2012;53:709–15.
16. Werner-Wasik M, Nelson AD, Choi W, Arai Y, Faulhaber PF, Kang P, et al. 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;82:1164–71.
17. Cheebsumon P, van Velden FH, Yaqub M, Frings V, de Langen AJ, Hoekstra OS, et al. Effects of image characteristics on

performance of tumor delineation methods: a test–retest assessment. J Nucl Med. 2011;52:1550–8.

18. Hatt M, Visvikis D, Le Rest CC. Autocontouring versus manual contouring. J Nucl Med. 2011;52:658.

19. MacManus M, Nestle U, Rosenzweig KE, Carrio I, Messa C, Belohlavek O, et al. Use of PET and PET/CT for radiation therapy planning: IAEA expert report 2006–2007. Radiother Oncol. 2009;91:85–94.

20. Geets X, Lee JA, Bol A, Lonneux M, Gregoire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging. 2007;34:1427–38.

21. Wanet M, Lee JA, Weynand B, De Bast M, Poncelet A, Lacroix V, et al. Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with threshold-based approaches. CT and surgical specimens. Radiother Oncol. 2011;98:117–25.

22. Murphy JD, Chisholm KM, Daly ME, Wiegner EA, Truong D, Iagaru A, et al. Correlation between metabolic tumor volume and pathologic tumor volume in squamous cell carcinoma of the oral cavity. Radiother Oncol. 2011;101:356–61.

23. Dewalle-Vignion AS, Yeni N, Petyt G, Verscheure L, Huglo D, Beron A, et al. Evaluation of PET volume segmentation methods: comparisons with expert manual delineations. Nucl Med Commun. 2012;33:34–42.

24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.

25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10.

26. Hatt M, Cheze-le Rest C, van Baardwijk A, Lambin P, Pradier O, Visvikis D. Impact of tumor size and tracer uptake heterogeneity in 18F-FDG PET and CT non-small cell lung cancer tumor delineation. J Nucl Med. 2011;52:1690–7.

27. Jackson T, Chung M, Ozonoff A, Mercier G, Subramaniam RM. FDG PET/CT inter-observer agreement in head and neck cancer: FDG and CT measurements of the primary tumor site. Nucl Med Commun. 2012;33:305–12.