Nucl Med Mol Imaging (2018) 52:5–15
DOI 10.1007/s13139-017-0493-6

REVIEW

ISSN (print) 1869-3482
ISSN (online) 1869-3474

Current Methods to Define Metabolic Tumor Volume in Positron Emission Tomography: Which One is Better?

Hyung-Jun Im[1,2] · Tyler Bradshaw[1] · Meiyappan Solaiyappan[3] · Steve Y. Cho[1,3,4]

Received: 15 July 2017 / Revised: 17 August 2017 / Accepted: 28 August 2017 / Published online: 19 September 2017

Abstract Numerous methods to segment tumors using ^{18}F-fluorodeoxyglucose positron emission tomography (FDG PET) have been introduced. Metabolic tumor volume (MTV) refers to the metabolically active volume of the tumor segmented using FDG PET, and has been shown to be useful in predicting patient outcome and in assessing treatment response. Also, tumor segmentation using FDG PET has useful applications in radiotherapy treatment planning. Despite extensive research on MTV showing promising results, MTV is not used in standard clinical practice yet, mainly because there is no consensus on the optimal method to segment tumors in FDG PET images. In this review, we discuss currently available methods to measure MTV using FDG PET, and assess the advantages and disadvantages of the methods.

Keywords Metabolic tumor volume · ^{18}F-fluorodeoxyglucose · Positron emission tomography · Tumor · Segmentation

✉ Steve Y. Cho
scho@uwhealth.org

[1] Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

[2] Department of Transdisciplinary Studies, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea

[3] Russell H. Morgan Department of Radiology and Radiological Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[4] University of Wisconsin Carbone Cancer Center, Madison, WI, USA

Introduction

^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography/computed tomography (PET/CT) has been used in the staging, restaging, and monitoring of treatment response in multiple types of malignancies. The high metabolic activity of a tumor in a pretreatment PET scan is associated with worse prognosis, and changes in metabolic activity from pretreatment to follow-up can be used in predicting response to treatment. The most commonly used parameter for the quantification of metabolic activity is the standardized uptake value (SUV), which is a ratio of tissue radioactivity concentration and the injected dose normalized by body weight (or lean body weight). The maximum SUV (SUVmax) is the maximum voxel value of SUV in the tumor. Since measuring SUVmax is simple and observer independent, SUVmax is the most commonly used parameter in clinical practice. However, SUVmax does not represent the whole tumor metabolic burden because the value is from only one voxel. Also, for the same reason, SUVmax is sensitive to image noise, and is therefore impacted by various patient characteristics and imaging parameters. Peak SUV (SUVpeak), which is the average value within a small, fixed-size region of interest (ROI) in the tumor, can be a more robust alternative to SUVmax. However, SUVpeak is sensitive to the size and the shape of the region of interest (ROI), and standards for measuring SUVpeak have not been established yet [1]. Metabolic tumor volume (MTV) is a measurement of the tumor volume with a high metabolism, while total lesion glycolysis (TLG) is defined as the product of the mean SUV and the MTV. In 1999, Larson et al. introduced the concept of TLG [2]. Since then, MTV and TLG have been extensively evaluated and have demonstrated efficacy in multiple types of malignancies. Moreover, MTV and TLG are considered to be more comprehensive parameters that better reflect metabolic tumor burden than SUVmax [3–5]. Also, tumor volume measurement using FDG PET has advantages over using anatomic imaging methods such as magnetic resonance imaging (MRI) or computed tomography (CT). Firstly, measuring MTV is easier and faster than tumor volume measurement from anatomic imaging. Also anatomic imaging methods may not reflect the shrinkage of the viable tumor portion after chemo- or radiotherapy. For example, it has been reported that MTV is more useful than tumor volume measured from MR for prediction of histologic response in osteosarcoma [3, 6]. However, the volumetric parameters (MTV and TLG) of FDG PET/CT have not been incorporated into standard clinical practice yet. This is because volumetric measurements of FDG PET/CT require an accurate segmentation of the tumor, unlike SUVmax. The optimal segmentation method to measure these values has not been established, and these values are significantly affected by segmentation methods [7, 8].

In this review, we briefly summarize the clinical utility of MTV and TLG, describe multiple segmentation methods to measure MTV, summarize the results of studies comparing multiple segmentation methods, and lastly, discuss what may be the most suitable method for measuring MTV.

Clinical Utility of MTV and TLG of FDG PET/CT

MTV and TLG have shown prognostic value in a variety of malignancies [9–12]. Multiple studies have shown that baseline MTV and TLG have prognostic value [13–15]. In particular, in non-small cell lung cancer, high MTV and TLG predicted worse prognosis in patients with low TNM stage who were treated with curative surgery [16], and also in patients with advanced stages treated with chemotherapy [17]. Also, in osteosarcoma, baseline MTV was an independent prognostic factor for metastasis-free survival [18]. A recent meta-analysis revealed that MTV and TLG are prognostic factors in non-small cell lung cancer and head and neck cancer [19, 20]. In the meta-analysis of non-small cell lung cancer, patients with high MTV had a worse prognosis with a hazard ratio of 2.71 for adverse events and a hazard ratio of 2.31 for death [19].

Changes in MTV and TLG during chemotherapy have been shown to be associated with overall tumor response. Larson et al. used a change of TLG after chemotherapy to measure treatment response [2]. Since then, there have been multiple studies that show the predictive role of MTV and TLG for treatment response [21, 22]. For example, MTV and TLG are reported to correlate better with histopathological response in NSCLC compared to SUVmax [23]. Also, changes in MTV after only 1 or 2 cycles of neoadjuvant chemotherapy were associated with tumor necrosis fraction in osteosarcoma and breast cancer [3, 4].

Tumor delineation using FDG PET has also been utilized for radiotherapy treatment planning in multiple types of malignancies [24–27]. CT is the standard imaging modality for defining gross target volume (GTV) for radiotherapy in lung cancer. In multiple recent reports, however, MTV from FDG PET/CT showed better characteristics for tumor delineation in lung cancer than CT and the results are summarized in a previous review article [28]. For example, Ashamalla et al. reported that FDG PET-based MTV had lower inter-observer variability than CT-based gross tumor volume (GTV). Also, MTV resulted in clinically significant modification of GTV in 52% of the enrolled patients [29]. Mah et al. reported that PET lowers physician variation in GTV delineation and alters patient management [30].

As briefly described above, multiple studies have shown the usefulness of MTV and TLG for prediction of treatment response or patient outcome, and tumor delineation for radiotherapy planning. However, there is no consensus on the optimal way to measure MTV using FDG PET/CT.

Current Methods to Segment Tumor for Measuring MTV

The definition of MTV is the volume inside a user- or algorithm-defined ROI that segments the metabolically active tumor. To determine the boundaries of the ROI, threshold-based or algorithm-based methods have been proposed and evaluated.

Numerous PET segmentation algorithms have been developed and applied to FDG PET images. It is challenging to group the numerous segmentation methods into distinct classes because of the vast variety of methods that have been developed. More advanced algorithms often integrate techniques from a variety of methods. Nonetheless, PET segmentation algorithms can generally be classified into threshold-based methods and algorithm-based methods. Both classes can be further broken down into subclasses. In this section, segmentation methods to measure MTV are described.

Threshold-Based Methods

In threshold-based methods, the image is partitioned into tumor and background using a distinct threshold value—all voxels with SUV above the threshold are assigned to the tumor, and all SUV below the threshold belong to the background.

Fixed Absolute Threshold

Absolute SUV thresholds are commonly used for measuring MTV. SUVs of 2.0, 2.5, 3.0, 4.5, 5.0 have all been reported as potential absolute thresholds. Among them, SUV 2.5 is the most widely accepted. SUV 2.5 was selected based on the results of early studies which reported SUV 2.5 as the optimal cut off between malignant and benign pulmonary nodule [31].

However, SUV 2.5 has been widely used in other types of malignancy as well. MTV and TLG measured using a threshold of SUV 2.5 have shown consistently good predictive value for prognosis in most studies, and have been associated with patient outcome in a meta-analysis [19]. Also, MTV measured using fixed absolute threshold was useful for assessing treatment response as well [3, 4]. However, absolute thresholds have clear limitations. Certain tumors with lower uptake can completely fall outside an absolute threshold, precluding the measurement of MTV in such tumors. Also, if a tumor has intense FDG uptake such as over SUV of 15, tumor volume can easily be overestimated by spillover effect (Fig. 1, Table 1).

Fixed Relative Threshold

Relative thresholds, which are defined as a certain percentage of SUVmax of a tumor, are also commonly used to measure MTV. In an early study, Erdi et al. found that relative thresholds of 36-44% produced similar volumes to the volumes measured from CT in lung cancer lesions bigger than 4 ml. Consequently, 40% or 42% have been the most widely used relative thresholds to measure MTV [32]. A limitation of thresholds, in general, is that a tumor volume with heterogeneous uptake (e.g., necrotic cores) could be underestimated by a relative threshold (Fig. 2). Also, as in the study by Erdi et al., a small lesion with low signal to background ratio can be overestimated by fixed relative threshold method. This can be problematic for treatment response assessment: a tumor with decreasing SUVmax may appear to grow in volume when in fact its boundaries remain the same (Table 1). Thus, using one fixed relative threshold for segmenting cancer lesions with a variety of sizes and signal-to-backgrounds can produce misleading results. Biehl et al. reported that there is no single threshold which provides accurate tumor delineation [33].

Background Threshold

To find a more accurate threshold which is both patient- and scan-specific, a background threshold has been proposed. In background thresholding, an ROI is placed in the liver or mediastinal blood pool to measure background SUV. Generally, SUVmean plus 1 or 2 standard deviation (SD) of the background is then used for the threshold [21, 34]. The shortcoming of the method is that the method is relatively more time-consuming than the other thresholding methods because background uptake needs to be measured separately. Also, even with the added efforts, the thresholds typically end up being consistently around SUV 3-4 with liver based thresholds, or around SUV 2 with mediastinal blood pool based thresholds. Instead of using the liver or blood pool as the background reference region, the background uptake immediately surrounding the tumor can be used to estimate the background uptake [23]. In this method, the mode of the ROI is used to describe the background uptake instead of the mean value. Based on the assumption that background activity has a Gaussian distribution, the Gaussian distribution can be subtracted from original ROI to get tumor segmentation. While fixed relative threshold can exclude a large part in the case of heterogeneous tumor, back ground subtracted volume (BSV) can include most of the heterogenous tumor (Fig. 3, from [23]). Burger et al. reported that BSV outperformed a 42% of SUVmax threshold in predicting histologic response in patients with lung cancer [23]. Further comparison study with other segmentation methods such as SUV 2.5 or algorithm based methods are needed to reinforce the utility of the

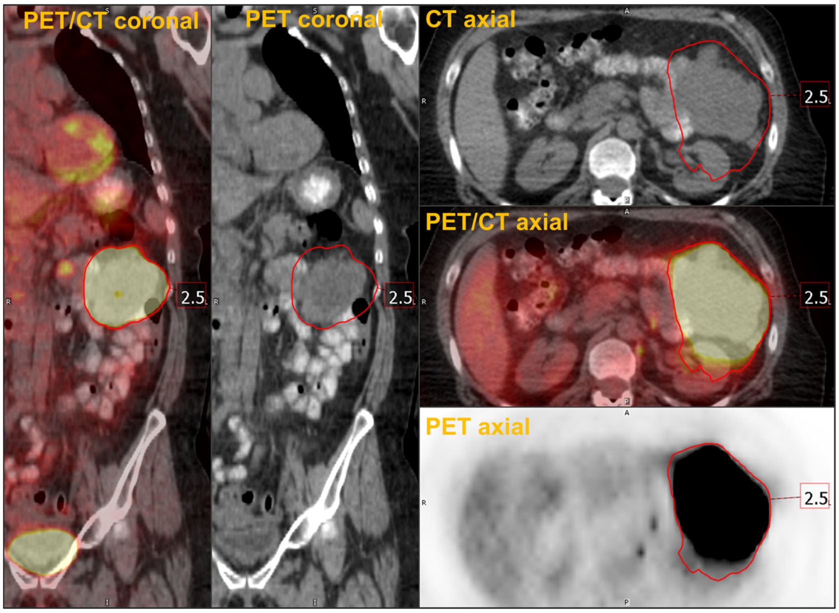

Fig. 1 Overestimation of MTV by a fixed absolute threshold. A patient with melanoma had a metabolically active soft tissue metastasis in left peritoneal space with SUVmax of 80.8. Tumor segmentation using a threshold of SUV 2.5 resulted in overestimation of the tumor volume because of spillover effect of FDG PET image. Red volume of interest indicates MTV using threshold of SUV 2.5

Table 1 Threshold based methods for measuring MTV

Methods	Examples	Advantage	Disadvantage
Fixed absolute	• SUV 2.0~ 5.0	• Simple and easy • High reproducibility • Observer independent	• Arbitrary • Overestimation in tumor with intense FDG uptake
Fixed relative	• 30~60% of tumor SUVmax	• Simple and easy • Observer independent	• Underestimation in heterogeneous tumor • Overestimation in low signal to noise lesion (low tumor uptake and/or high background) • Limitation in assessment of treatment response
Background	• Liver +1SD, 2SD • Mediastinal blood pool +1SD, 2SD • Tumor background (BSV)	• Patient and scan adjusted threshold	• More time consuming • Relatively low reproducibility
Adaptive	• (0.15 x SUVmean of tumor) + SUVmean of local background • Signal to background ratio	• Patient and scan adjusted threshold	• Relatively low reproducibility • Many different methods but no comparison study yet

SUV standardized uptake value, *SD* standard deviation, *BSV* background subtracted volume

BSV. Also, the sensitivity of the method to ROI size has not been reported (Table 1).

Adaptive Threshold

Adaptive thresholds do not use fixed relative thresholds or absolute thresholds, but rather adjust the threshold on a case-by-case basis according to different measurable properties of the image. Adaptive thresholds are commonly calculated as functions of the tumor volume, tumor uptake, background uptake, and/or contrast. For example, Erdi et al. describe an adaptive threshold that decreases as an exponential function of tumor volume [32]. Nestle et al. reported an adaptive method based on tumor and background uptake, and showed that the method is more suitable for heterogeneous tumor than relative thresholds [35]. Since multiple factors are considered for an adaptive threshold, there can be multiple ways to calculate an adaptive threshold, and there is no consensus yet on the optimal method (Table 1).

Algorithm-Based Methods

Threshold-based segmentation methods are easy to implement and widely used, but they do have several well-known shortcomings. Thresholding may exclude cold regions inside heterogeneous tumors, or erroneously include regions of elevated background. Also, thresholding assumes that a tumor volume has the same intensity value at every point along its boundary, an assumption that is often violated. Consequently, more advanced algorithms have been developed to address some of these issues.

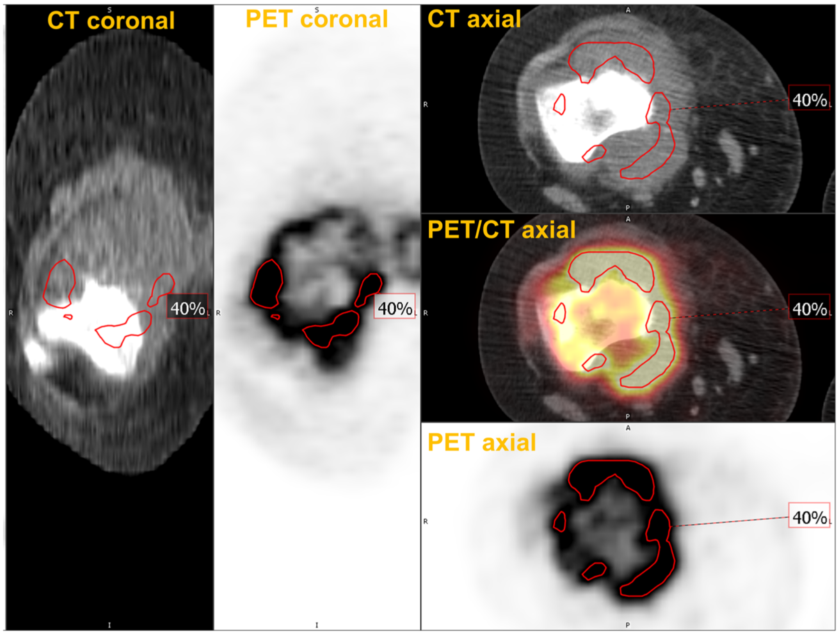

Fig. 2 Underestimation of MTV by a fixed relative threshold. A patient with osteosarcoma had the primary lesion in his right distal femur. The tumor had heterogeneous uptake with SUVmax of 13.6. Tumor segmentation using a relative threshold of 40% of SUVmax resulted in an underestimation of the tumor volume

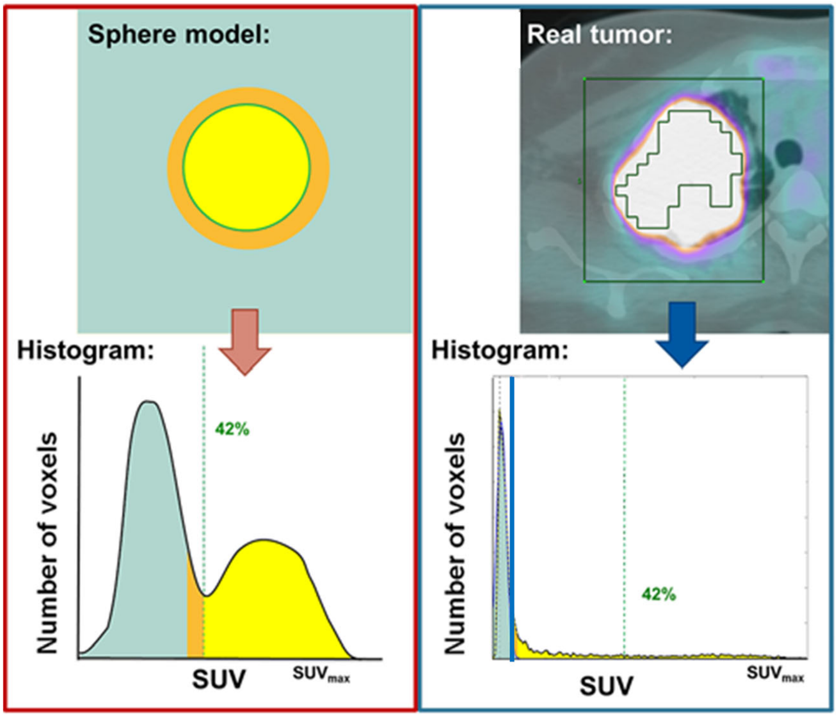

Fig. 3 Comparison of fixed relative threshold and background based threshold methods. (Left) Relative threshold of SUVmax of 42% has shown to delineate tumor volume optimally in a spherical tumor with homogenous uptake. (Right) In contrast, heterogeneous real tumors are often underestimated using the 42% threshold (green ROI). Using background based threshold (blue line), Gaussian normal distribution representing the background can be subtracted and all tumor voxels can be included in MTV. (Reproduced with permission [23])

Gradient-Based Methods

Gradient-based methods define the tumor boundary by exploiting the image gradient that exists between the high SUV in tumor cells and the lower SUV in adjacent non tumor tissues [36]. Gradient methods have been shown to outperform threshold methods in phantom studies [37] and for resected cancer specimens [38, 39]. An advantage of gradient methods is that they are not necessarily dependent on tumor uptake levels. However, gradient methods can be sensitive to the reconstruction parameters of the PET images. Also, the method has the disadvantage of assuming uniform contrast around tumor edges (Table 2). There are different implementations of gradient-based segmentation because there are different ways to extract boundaries from gradient images. For example, Geets et al. used a watershed algorithm together with clustering to determine tumor edges [38]. More recently, the PET Edge tool was developed and implemented by MIM Software (Cleveland, OH) and has been used to evaluate MTV in lung cancer [40]. As with all algorithms, segmentation results should always be checked manually, although one report suggested that manual adjustment may not be necessary for one such gradient method [41].

Image gradients are often used as inputs to more advanced segmentation algorithms. One such class of algorithms is active contours, or, as they are commonly called, snakes. The basic concept of active contours is that a contour is placed around the tumor as an initial guess. The contour is then actively deformed to better fit the edges of the tumor—like a snake wrapping around its prey. The deformation is controlled by a function which considers different aspects of the ROI, such as the image gradient and possibly other features, depending on its implementation [42]. Active contours have been used in many image processing applications, including segmentation of lung tumors in PET images [43].

Classifier-Based Method

Classification is the general term used in statistics and computer science for partitioning data into groups with similar characteristics. For this review, we will consider two types of classifiers used in segmentation: clustering and supervised learning algorithms.

In clustering, voxels with similar features are grouped into clusters by the algorithm. These clusters of voxels then define the ROIs (e.g., tumor and background clusters). Each voxel can have multiple features that are considered by the algorithm, such as the voxel's SUV and its gradient. The number and type of features used as inputs vary depending on the implementation. A popular clustering algorithm is the fuzzy c-means (FCM) algorithm. FCM is considered 'fuzzy' because, for each voxel, it reports the probability of the voxel's belonging to the tumor cluster and to the background cluster. FCM has shown good reproducibility in ^{18}F-FDG and ^{18}F-FLT PET images [44]. Many variations of FCM have been developed over the years, with a recent implementation demonstrating high accuracy and robustness in phantoms, simulated tumors, and in patients with non-small cell lung cancer [45]. The method is simple, fast and adaptable, but not highly stable with heterogeneous tumors (Table 2).

Table 2 Algorithm based methods for measuring MTV

Methods	Description	Advantages	Disadvantages
Gradient method	• Image gradients (spatial derivative) used to find tumor edges	• Not dependent on tumor uptake level	• Errors due to reconstruction steps • Assumes uniform contrast around tumor edges
Fuzzy C-means (FCM)	• Clustering method	• Simple, fast, and adaptable	• Struggles with heterogeneous tumors or for certain geometries
Artificial neural networks	• Machine learning method	• Can learn to handle various conditions	• Requires a large quantity of high quality training data
Fuzzy locally adaptive Bayesian (FLAB)	• Statistical modeling method	• Accurate and reproducible across a variety of conditions	• Challenging to implement
Multi-Otsu method	• Statistical method	• Simple and fast • Stable and consistent	• Clinical significance has not been evaluated yet

Supervised learning is a form of machine learning where the user trains the algorithm by giving it training data together with known results. In segmentation, this means the algorithm is trained by feeding it PET images together with the best ROIs for those images. The algorithm then learns how to predict the best ROIs for future PET images. Different supervised learning algorithms exist. Artificial neural networks (ANNs) are one of the oldest machine learning techniques, and continue to be one of the best. ANNs have been used in multiple different image segmentation tasks, including segmenting lung tumors in PET images [46]. The method has a potential to be a highly accurate method which can handle various conditions after training with large data sets. However, it is hard to get a large quantity of high quality training data with accurate true volume data (Table 2).

Statistical Methods

Statistical-based segmentation algorithms are a broad class of methods that attempt to describe the image in terms of statistical distributions of voxel intensity values. The goal is to describe the PET image as a mixture of the intensity distribution belonging to the tumor class together with the intensity distribution belonging to the background class. Those voxels belonging to the tumor's intensity distribution are then segmented accordingly.

The fuzzy locally adaptive Bayesian (FLAB) algorithm uses a Bayesian framework for determining whether a voxel belongs to the tumor or background. It combines various statistical models and includes fuzzy classes (in addition to tumor and background classes) which helps to simultaneously address issues of both noise and blur resulting from partial-volume effects in PET images. FLAB is also able to deal with highly heterogeneous tumors when three classes are used in the algorithm [47]. Hatt et al. introduced FLAB method for MTV measurement and showed robustness of the method in a phantom study [48]. Robustness of FLAB also has been reported in lung cancer [47] and breast cancer [49]. The method is reproducible across a variety of conditions but implementation of the method is complicated (Table 2).

The Otsu method for thresholding is one of the earliest statistical image segmentation techniques [50]. The method classifies pixels in an image into two classes by exhaustively searching for a threshold value that minimizes the intra-class variances, 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). Several different improvements and variations of the Otsu method have been developed over the years, including a recently developed multi-Otsu method [51]. Using the algorithm, tumor with high FDG uptake can be segmented with very minimal user interaction. The method has demonstrated stable and consistent delineation across a range of tumor sizes and SUV values in phantom study and metastatic melanoma lesions [51]. Also multi-Otsu method can segment the heterogenous tumor more reliably than fixed relative threshold (Fig. 4). Also, multiple lesions can be segmented in one process. However, clinical relevance of the measured MTV using the method has not been evaluated (Table 2).

Comparison of the FDG PET Tumor Segmentation Methods

Lung cancer is one of the most extensively evaluated malignancies using MTV and TLG. MTV and TLG have demonstrated clinical utilities in lung cancer in a number of applications including risk stratification, response evaluation and radiotherapy planning. A systematic search of PUBMED and MEDLINE was performed using the keywords "FDG", "lung cancer" and "MTV or TLG or GTV". A total of 132 studies were identified, and 32 studies were excluded which were reviews, meta-analysis, or included only CT based volume. Among 100 studies, 70 studies (70%) used the single

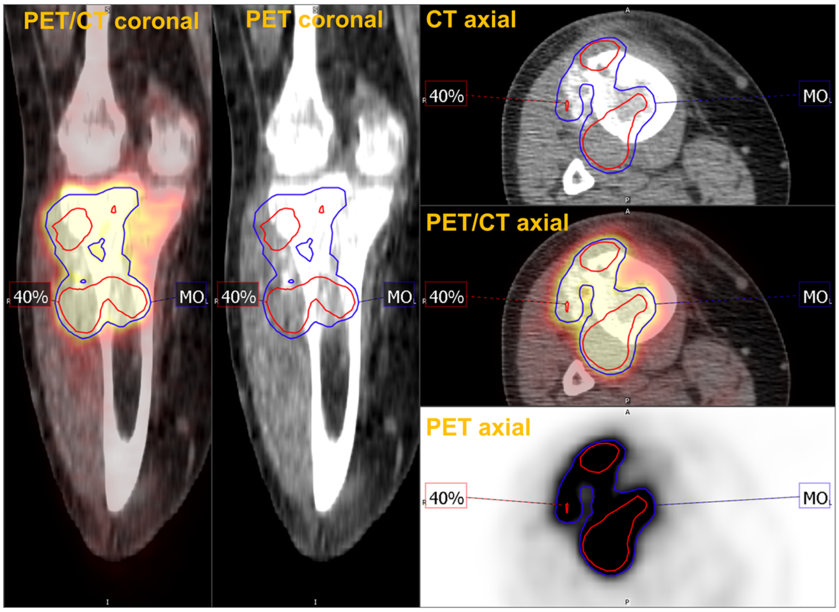

Fig. 4 Multi-Otsu method (MO, blue) showed better tumor segmentation than 40% threshold (red) in a case of osteosarcoma with heterogenous uptake

threshold based method. The threshold based methods consist of fixed absolute, fixed relative, background, and adaptive methods. In 14 studies (14%) with algorithm based methods, 13 studies used gradient methods and one study used the FLAB method (Fig. 5). The other 16 studies used multiple segmentation methods and the results of the studies were summarized in Table 3. Among the 16 studies, 11 studies evaluated the predictive value of MTV [23, 52–61]. Seven studies showed comparable predictive value between the different segmentation methods. Park et al. reported that MTVs using fixed thresholds of SUV 1.5, 2.0, 2.5, and 3.0 showed comparable abilities for predicting occult LN metastasis [61]. In a study by Yoo et al., predictive values of MTVs using fixed thresholds of SUV 2.5, and 25%, 50%, 75% of SUVmax had comparable performance in predicting patient outcome. However, MTV using liver based threshold was not predictive of survival [54]. Lin et al. reported that MTV using an absolute threshold of SUV 2.5 was predictive of survival but MTVs using relative thresholds of 40% and 50% of SUVmax were not predictive of survival [55]. Also, Abelson et al. reported that MTVs using thresholds of SUV 7 and SUV 10 were predictive of survival, but MTVs using thresholds of SUV 2, SUV 4, and 50% of SUVmax were not [56]. Among the studies that employed multiple tumor segmentation methods, only one study included an algorithm based method. Harris et al. reported that MTV using 50% of SUVmax as a threshold and a gradient based method were comparable in predicting prognosis [58].

This review of the previous studies suggests that using fixed absolute thresholds to measure MTV may be better at extracting clinically-relevant MTV information than using a relative fixed threshold to measure MTV, although we caution that this conclusion would need to be tested in a large study designed to measure such an endpoint. As we discussed above, a tumor with intense FDG uptake is overestimated by fixed absolute threshold, and underestimated by fixed relative threshold. On the other hand, a tumor with faint FDG uptake is underestimated by fixed absolute threshold, and overestimated by fixed relative threshold. In consequence, the difference in metabolism between the tumors would be increased by fixed absolute threshold and reduced by fixed relative threshold. This may be the reason that fixed absolute threshold is more predictive of prognosis than relative threshold.

Four studies compared the accuracy of the tumor delineation using different thresholds [33, 62–64]. Burger et al. reported that BSV has higher correlation with true tumor volume than MTV using a threshold of SUV 2.5 or 42% of

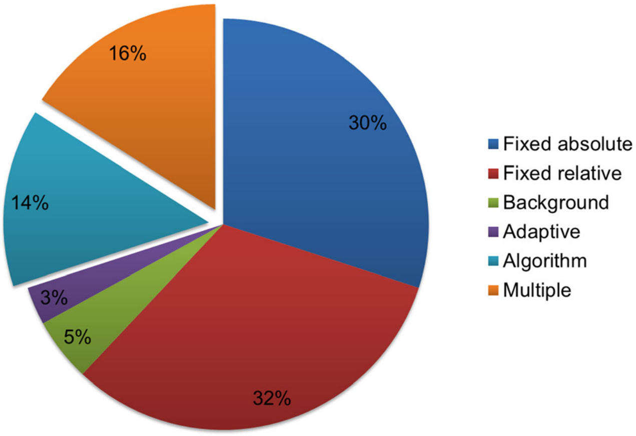

Fig. 5 Segmentation methods for measuring MTV used in lung cancer studies. Of the studies 70% used single threshold based methods, and while 15% used single algorithm based methods; 15% of the studies used multiple methods to measure MTV

Table 3 Lung cancer studies including multiple methods to measure MTV

First author (ref)	Design	Purpose	Pt no.	Segmentation methods	Findings
Mehta et al. [52]	Retrospective	Predict outcome	288	40%, 50%	Comparable (predictive)
Arslan et al. [53]	Retrospective	Predict outcome	25	SUV 2.5 / 50%	Comparable (predictive)
Yoo Ie et al. [54]	Retrospective	Predict outcome	58	SUV 2.5 / 25%, 50%, 75% / liver based	Liver based threshold was inferior. The others were comparable.
Lin et al. [55]	Retrospective	Predict outcome	60	SUV 2.5 / 40%, 50%	SUV 2.5 was better than 40%, 50%.
Abelson et al. [56]	Retrospective	Predict outcome	54	SUV 2, 4, 7, 10 / 50%	SUV 7, 10 were better than the others.
Kim et al. [57]	Retrospective	Predict outcome	91	SUV 2.5, 3.0, 3.5, 4.0	Comparable (predictive)
Harris et al. [58]	Retrospective	Predict outcome	29	50% / Gradient	Comparable (predictive)
Carvalho et al. [59]	Retrospective	Predict outcome	220	2.5, 3, 4 / 40%, 50%	Comparable (not predictive)
Lee et al. [60]	Retrospective	Predict outcome	57	40%, 50%	Comparable (not predictive)
Park et al. [61]	Retrospective	Predict occult LN metastasis	39	SUV 1.5, 2.0, 2.5, 3.0	Comparable, SUV 2.0 selected
Burger et al. [23]	Retrospective	Predict treatment response	44	42% / BSV	BSV had higher correlation with response.
Burger et al. [62]	Retrospective	Compare accuracy of the tumor delineation	50	2.5 / 42% / BSV	BSV had higher correlation with reference volume.
Chen et al. [63]	Retrospective	Compare accuracy of the tumor delineation	37	SUV 2.5 / 40%, 50% / Adaptive	Adaptive method had higher correlation with CT volume.
Yu et al. [64]	Prospective	Compare accuracy of the tumor delineation	15	SUV 1.5~5.5 / 15~60%	Optimal relative and absolute thresholds were 31% ± 11% and 3.0 ± 1.6.
Biehl et al. [33]	Retrospective	Compare accuracy of the tumor delineation	20	10%, 20%, 30%, 40%, 50%	The optimal threshold is different according to CT volume.
Laffon et al. [65]	Retrospective	Assess variability of TLG measurement	13	40%, 50%, 60%, 70%, 80%	Variability was the lowest in 40%.

BSV background subtracted volume, *SUV* standardized uptake value, __% relative fixed threshold using __% of SUVmax of the tumor, *TLG* total lesion glycolysis, *CT* computed tomography

SUVmax [62]. Chen et al. reported that the adaptive method was better than SUV 2.5 or 40% or 50% of SUVmax [63]. By comparing multiple fixed relative thresholds, Biehl et al. reported that the optimal threshold is different according to CT volume [33]. In summary, a single fixed threshold may be not well suited for assessing actual tumor volume. Thus, tumor background based or an adaptive threshold or algorithm based method would be more appropriate for accurate tumor delineation and assessment of the actual tumor volume [51].

Several algorithm-based methods including PET Edge do not allow hollow inside the tumor contour. Meanwhile, threshold-based methods allow hollow inside the contour. Thus, the tumor part with lower FDG uptake can be excluded using threshold-based methods, which can be problematic in some cases. Especially, since a tumor region with lower FDG uptake may represent the heterogeneous nature of the tumor, threshold-based methods may lose information regarding tumor heterogeneity. Therefore, when a researcher tries to estimate the heterogeneity of the tumor, it might be better to use algorithm-based MTV. However, a systematic study is warranted to confirm which method is better for evaluating tumor heterogeneity.

Fixed relative threshold methods are not suitable to use for tumor volume measurement during or after treatment because the methods have high variability according to SUVmax of the tumor. In particular, if SUVmax of the tumor declines during or after treatment, MTV using fixed relative threshold will overestimate the residual tumor volume because of the declined SUVmax. Also, residual tumor volume can be underestimated by fixed absolute threshold. Thus, algorithm-based methods may be more suitable for estimating MTV during or after treatment since the methods are compatible with tumors with various ranges of FDG uptake values.

Conclusions

The optimal tumor segmentation method may be different according to the purpose of the study. To predict patient outcome, MTV measured using fixed absolute thresholds consistently performs well and has better prognostic value than even MR-defined volumes in several studies [18–20]. Thus, fixed absolute thresholds may be a suitable choice to evaluate the prognostic value of MTV, since the method is simple, fast, and maximizes the difference of metabolic burden between different tumors. Meanwhile, fixed absolute and relative thresholds have shown clear limitations in tumor segmentation tasks, while adaptive or algorithm based methods can segment the tumor more accurately in tumors with wide ranges of uptake and size. Thus, for tumor response prediction or accurate tumor delineation (e.g., for radiotherapy applications), algorithm based methods seem to be better than fixed threshold methods. However, since the numerous algorithm-based segmentation methods have not been systematically tested for accuracy, robustness, and repeatability on the same datasets, it is hard to select the best algorithm based method for now. Therefore, unbiased phantom data acquired under various conditions and publically-available patient images with ground truth (e.g., consensus segmentations) to compare multiple algorithm-based methods are warranted.

Compliance with Ethical Standards

Conflict of Interest Hyung-Jun Im was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (490-20,170,035). Tyler Bradshaw, Meiyappan Solaiyappan, and Steve Y. Cho declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent The institutional review board of our institute approved retrospective studies which were used in this review article, and the requirement to obtain informed consent was waived.

References

1. 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.
2. Larson SM, Erdi Y, Akhurst T, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. Clin Positron Imaging. 1999;2:159–71.
3. Im HJ, Kim TS, Park SY, et al. 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. Im HJ, Kim YK, Kim YI, Lee JJ, Lee WW, 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.
5. Lee JW, Kang CM, Choi HJ, et al. 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.
6. Byun BH, Kong CB, Lim I, et al. Early response monitoring to neoadjuvant chemotherapy in osteosarcoma using sequential (1)(8)F-FDG PET/CT and MRI. Eur J Nucl Med Mol Imaging. 2014;41:1553–62.
7. Cheebsumon P, van Velden FHP, Yaqub M, et al. Effects of image characteristics on performance of tumor delineation methods: a test–retest assessment. J Nucl Med. 2011;52:1550–8.
8. Moon SH, Hyun SH, Choi JY. Prognostic significance of volume-based PET parameters in cancer patients. Korean J Radiol. 2013;14: 1–12.
9. JH O, Choi WH, Han EJ, et al. The prognostic value of (18)F-FDG PET/CT for early recurrence in operable breast cancer: comparison with TNM stage. Nucl Med Mol Imaging. 2013;47:263–7.
10. Costelloe CM, Macapinlac HA, Madewell JE, et al. 18F-FDG PET/CT as an indicator of progression-free and overall survival in osteosarcoma. J Nucl Med. 2009;50:340–7.
11. Hyun SH, Choi JY, Shim YM, et al. Prognostic value of metabolic tumor volume measured by 18F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. Ann Surg Oncol. 2009;17:115–22.
12. Yoo J, Choi JY, Moon SH, et al. Prognostic significance of volume-based metabolic parameters in uterine cervical cancer determined using 18F-fluorodeoxyglucose positron emission tomography. Int J Gynecol Cancer. 2012;22:1226–33.
13. Chen HH, Chiu NT, WC S, Guo HR, Lee BF. Prognostic value of whole-body total lesion glycolysis at pretreatment FDG PET/CT in non-small cell lung cancer. Radiology. 2012;264:559–66.
14. Hyun SH, Ahn HK, Ahn MJ, et al. Volume-based assessment with 18F-FDG PET/CT improves outcome prediction for patients with stage IIIA-N2 non-small cell lung cancer. AJR Am J Roentgenol. 2015;205:623–8.
15. Hyun SH, Ahn HK, Kim H, et al. Volume-based assessment by (18)F-FDG PET/CT predicts survival in patients with stage III non-small-cell lung cancer. Eur J Nucl Med Mol Imaging. 2014;41:50–8.
16. Kim DH, Son SH, Kim CY, et al. Prediction for recurrence using F-18 FDG PET/CT in pathologic N0 lung adenocarcinoma after curative surgery. Ann Surg Oncol. 2014;21:589–96.
17. Zaizen Y, Azuma K, Kurata S, et al. Prognostic significance of total lesion glycolysis in patients with advanced non-small cell lung cancer receiving chemotherapy. Eur J Radiol. 2012;81:4179–84.
18. Byun BH, Kong C-B, Park J, et al. Initial metabolic tumor volume measured by 18F-FDG PET/CT can predict the outcome of Osteosarcoma of the extremities. J Nucl Med. 2013;54:1725–32.
19. Im HJ, Pak K, Cheon GJ, et al. Prognostic value of volumetric parameters of (18)F-FDG PET in non-small-cell lung cancer: a meta-analysis. Eur J Nucl Med Mol Imaging. 2015;42:241–51.
20. Pak K, Cheon GJ, Nam HY, et al. 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;55:884–90.
21. Han EJ, Yang YJ, Park JC, Park SY, Choi WH, Kim SH. Prognostic value of early response assessment using 18F-FDG PET/CT in chemotherapy-treated patients with non-small-cell lung cancer. Nucl Med Commun. 2015;36:1187–94.
22. Huang W, Fan M, Liu B, et al. Value of metabolic tumor volume on repeated 18F-FDG PET/CT for early prediction of survival in locally advanced non-small cell lung cancer treated with concurrent chemoradiotherapy. J Nucl Med. 2014;55:1584–90.
23. Burger IA, Casanova R, Steiger S, et al. FDG-PET/CT of non-small cell lung carcinoma under neo-adjuvant chemotherapy: background based adaptive volume metrics outperform TLG and MTV in predicting histopathological response. J Nucl Med. 2016;57:849–54.
24. Braendengen M, Hansson K, Radu C, Siegbahn A, Jacobsson H, Glimelius B. Delineation of gross tumor volume (GTV) for radiation treatment planning of locally advanced rectal cancer using information from MRI or FDG-PET/CT: a prospective study. Int J Radiat Oncol Biol Phys. 2011;81:e439–45.
25. Spratt DE, Diaz R, McElmurray J, et al. Impact of FDG PET/CT on delineation of the gross tumor volume for radiation planning in non-small-cell lung cancer. Clin Nucl Med. 2010;35:237–43.

26. Heron DE, Andrade RS, Flickinger J, et al. Hybrid PET-CT simulation for radiation treatment planning in head-and-neck cancers: a brief technical report. Int J Radiat Oncol Biol Phys. 2004;60:1419–24.
27. Terezakis SA, Hunt MA, Kowalski A, et al. [(1)(8)F]FDG-positron emission tomography coregistration with computed tomography scans for radiation treatment planning of lymphoma and hematologic malignancies. Int J Radiat Oncol Biol Phys. 2011;81:615–22.
28. Nestle U, Kremp S, Grosu AL. Practical integration of [18F]-FDG-PET and PET-CT in the planning of radiotherapy for non-small cell lung cancer (NSCLC): the technical basis, ICRU-target volumes, problems, perspectives. Radiother Oncol. 2006;81:209–25.
29. Ashamalla H, Rafla S, Parikh K, et al. The contribution of integrated PET/CT to the evolving definition of treatment volumes in radiation treatment planning in lung cancer. Int J Radiat Oncol Biol Phys. 2005;63:1016–23.
30. Mah K, Caldwell CB, Ung YC, et al. The impact of (18)FDG-PET on target and critical organs in CT-based treatment planning of patients with poorly defined non-small-cell lung carcinoma: a prospective study. Int J Radiat Oncol Biol Phys. 2002;52:339–50.
31. Paulino AC, Johnstone PA. FDG-PET in radiotherapy treatment planning: Pandora's box? Int J Radiat Oncol Biol Phys. 2004;59: 4–5.
32. Erdi YE, Mawlawi O, Larson SM, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. Cancer. 1997;80:2505–9.
33. Biehl KJ, Kong FM, Dehdashti F, et al. 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;47:1808–12.
34. Hyun SH, Choi JY, Kim K, et al. Volume-based parameters of (18)F-fluorodeoxyglucose positron emission tomography/computed tomography improve outcome prediction in early-stage non-small cell lung cancer after surgical resection. Ann Surg. 2013;257:364–70.
35. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of 18F-FDG PET–positive tissue for target volume definition in radiotherapy of patients with non–small cell lung cancer. J Nucl Med. 2005;46:1342–8.
36. Graves EE, Quon A, Loo BW, Jr. RT_Image: an open-source tool for investigating PET in radiation oncology. Technol Cancer Res Treat 2007;6:111-121.
37. Werner-Wasik M, Nelson AD, Choi W, 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.
38. Geets X, Lee JA, Bol A, Lonneux M, Gregoire VA. Gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging. 2007;34:1427–38.
39. Sridhar P, Mercier G, Tan J, Truong MT, Daly B, Subramaniam RM FDG-PET. Metabolic tumor volume segmentation and pathologic volume of primary human solid tumors. AJR Am J Roentgenol. 2014;202:1114–9.
40. Liao S, Penney BC, Zhang H, Suzuki K, Prognostic PY. Value of the quantitative metabolic volumetric measurement on 18F-FDG PET/CT in stage IV nonsurgical small-cell lung cancer. Acad Radiol. 2012;19:69–77.
41. Obara P, Liu H, Wroblewski K, et al. Quantification of metabolic tumor activity and burden in patients with non-small-cell lung cancer: is manual adjustment of semiautomatic gradient-based measurements necessary? Nucl Med Commun. 2015;36:782–9.
42. Xu C, Prince JL. Snakes, shapes, and gradient vector flow. IEEE Trans Image Process. 1998;7:359–69.
43. Abdoli M, Dierckx RA, Zaidi H. Contourlet-based active contour model for PET image segmentation. Med Phys. 2013;40:082507.
44. Hatt M, Cheze-Le Rest C, Aboagye EO, et al. Reproducibility of 18F-FDG and 3′-deoxy-3′-18F-fluorothymidine PET tumor volume measurements. J Nucl Med. 2010;51:1368–76.
45. Lapuyade-Lahorgue J, Visvikis D, Pradier O, Cheze Le Rest C, Hatt M. SPEQTACLE: an automated generalized fuzzy C-means algorithm for tumor delineation in PET. Med Phys. 2015;42:5720–34.
46. Sharif MS, Abbod M, Amira A, Zaidi H. Artificial neural network-based system for PET volume segmentation. Int J Biomed Imaging. 2010;2010:105610.
47. Hatt M, Cheze le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys. 2010;77:301–8.
48. 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.
49. Hatt M, Groheux D, Martineau A, et al. Comparison between 18F-FDG PET image-derived indices for early prediction of response to neoadjuvant chemotherapy in breast cancer. J Nucl Med. 2013;54: 341–9.
50. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9:62–6.
51. Huang E, Solaiyappan M, 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.
52. Mehta G, Chander A, Huang C, Kelly M, Fielding P. Feasibility study of FDG PET/CT-derived primary tumour glycolysis as a prognostic indicator of survival in patients with non-small-cell lung cancer. Clin Radiol. 2014;69:268–74.
53. Arslan N, Tuncel M, Kuzhan O, et al. Evaluation of outcome prediction and disease extension by quantitative 2-deoxy-2-[18F] fluoro-D-glucose with positron emission tomography in patients with small cell lung cancer. Ann Nucl Med. 2011;25:406–13.
54. Yoo Ie R, Chung SK, Park HL, et al. Prognostic value of SUVmax and metabolic tumor volume on 18F-FDG PET/CT in early stage non-small cell lung cancer patients without LN metastasis. Biomed Mater Eng. 2014;24:3091–103.
55. Lin Y, Lin WY, Kao CH, Yen KY, Chen SW, Yeh JJ. Prognostic value of preoperative metabolic tumor volumes on PET-CT in predicting disease-free survival of patients with stage I non-small cell lung cancer. Anticancer Res. 2012;32:5087–91.
56. Abelson JA, Murphy JD, Trakul N, et al. Metabolic imaging metrics correlate with survival in early stage lung cancer treated with stereotactic ablative radiotherapy. Lung Cancer. 2012;78:219–24.
57. Kim K, Kim SJ, Kim IJ, Kim YS, Pak K, Kim H. Prognostic value of volumetric parameters measured by F-18 FDG PET/CT in surgically resected non-small-cell lung cancer. Nucl Med Commun. 2012;33:613–20.
58. Harris JP, Chang-Halpenny CN, Maxim PG, et al. Outcomes of modestly Hypofractionated radiation for lung tumors: pre- and mid-treatment positron emission tomography-computed tomography metrics as prognostic factors. Clin Lung Cancer. 2015;16:475–85.
59. Carvalho S, Leijenaar RT, Velazquez ER, et al. Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. Acta Oncol. 2013;52: 1398–404.
60. Lee VH, Chan WW, Lee EY, et al. Prognostic significance of standardized uptake value of lymph nodes on survival for stage III non-small cell lung cancer treated with definitive concurrent chemoradiotherapy. Am J Clin Oncol. 2014;39:355–62.
61. Park SY, Yoon JK, Park KJ, Lee SJ. Prediction of occult lymph node metastasis using volume-based PET parameters in small-sized peripheral non-small cell lung cancer. Cancer Imaging. 2015;15:21.

62. Burger IA, Vargas HA, Apte A, et al. PET quantification with a histogram derived total activity metric: superior quantitative consistency compared to total lesion glycolysis with absolute or relative SUV thresholds in phantoms and lung cancer patients. Nucl Med Biol. 2014;41:410–8.

63. Chen GH, Yao ZF, Fan XW, et al. Variation in background intensity affects PET-based gross tumor volume delineation in non-small-cell lung cancer: the need for individualized information. Radiother Oncol. 2013;109:71–6.

64. Yu J, Li X, Xing L, et al. Comparison of tumor volumes as determined by pathologic examination and FDG-PET/CT images of non-small-cell lung cancer: a pilot study. Int J Radiat Oncol Biol Phys. 2009;75:1468–74.

65. Laffon E, de Clermont H, Lamare F, Marthan R. Variability of total lesion glycolysis by 18F-FDG-positive tissue thresholding in lung cancer. J Nucl Med Technol. 2013;41:186–91.