PLOS ONE

RESEARCH ARTICLE

A phantom study to assess the reproducibility, robustness and accuracy of PET image segmentation methods against statistical fluctuations

Mahbubunnabi Tamal *

Department of Biomedical Engineering, College of Engineering, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

* mtamal@yahoo.com, mtamal@iau.edu.sa

OPEN ACCESS

Citation: Tamal M (2019) A phantom study to assess the reproducibility, robustness and accuracy of PET image segmentation methods against statistical fluctuations. PLoS ONE 14(7): e0219127. https://doi.org/10.1371/journal.pone.0219127

Editor: Huafeng Liu, Zhejiang University, CHINA

Received: October 31, 2018

Accepted: June 17, 2019

Published: July 8, 2019

Data Availability Statement: Data is available and uploaded to Open Science Framework (OSF) at the following link: https://osf.io/kt2eh/.

Funding: This study was funded by the Deanship of Scientific Research (DSR), Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia (grant number 2016-212-Eng).

Competing interests: The author declares that he has no conflict of interest.

Abstract

Background

Automatic and semi-automatic segmentation methods for PET serve as alternatives to manual delineation and eliminate observer variability. The robustness of these segmentation methods against statistical fluctuations arising from variable size, contrast and noise are vital for providing reliable clinical outcomes for diagnosis and treatment response assessment. In this study, the performances of several segmentation methods have been investigated using the torso NEMA phantom against statistical fluctuations.

Methods

The six hot spheres (0.5-27ml) and the background of the phantom were filled with different activities of ^{18}F to yield 2:1 and 4:1 contrast ratios. The phantom was scanned on a TrueV PET-CT scanner for 120 minutes. The images were reconstructed using OSEM (4iterations-21subsets) for different durations (15, 20, 34 and 67 minutes) to represent different noise levels and smoothed with a 4-mm Gaussian filter. Each sphere with different settings was delineated using a fixed 40% threshold (40T), fuzzy clustering mean (FCM), adaptive threshold and region based variational (C-V) segmentation methods and compared with the gold standard volume, which was estimated from the known diameter and position of each sphere.

Results

The smallest three spheres at the 2:1 contrast level are not evaluable for the 40T method. For the other spheres, the 40T method grossly overestimates the volumes and the segmented volumes are highly dependent on the statistical variations. These volumes are the least reproducible (80%) with a mean Dice Similarity Coefficient (DSC) of 0.67 and 90% classification error (CE). The other three methods reduce the dependency on noise and contrast in a similar manner by providing low bias (<10%) and CE (<25%) as well as a high DSC (0.88) and reproducibility (30%) for objects >17mm in diameter. However, for the smallest

three spheres at a 2:1 contrast level, the performances of all three methods were significantly low, with the adaptive method being superior to the FCM and C-V (mean bias 168% and 350%, mean DSC 0.65 and 0.50, mean CE 227% and 454% for the adaptive and other two methods (approximately similar for FCM and C-V), respectively).

Conclusions

The segmentation accuracy of the fixed threshold-based method depends on size, contrast and noise. The intensity thresholds determined by the adaptive threshold methods are less sensitive to noise and therefore, the segmented volumes are more reproducible across different acquisition durations. A similar performance can be achieved with the FCM and C-V methods. Though, for small lesions (< 2cm diameter) with low counts and contrast, the adaptive threshold-based method outperforms the FCM and C-V methods, and the performance of neither of these methods is optimal for volumes <2cm in diameter. These three methods can only reliably be used to delineate tumours for diagnostic and monitoring purposes provided that the contrast between the tumour and background is not below a 2:1 ratio and the size of the tumour does not fall not below 2cm in diameter in response to treatment. They can also be used for different radiotracers with variable uptake. However, the FCM and C-V methods have the advantage of not requiring calibrations for different scanners and settings.

Introduction

Positron emission tomography (PET), a functional imaging technique, provides 3D images of the whole body. The availability of a wide range of physiologically relevant imaging contrast agents also makes PET a flexible imaging modality. Functional or metabolic volumes along with standardized uptake values (SUVs) extracted from 3D whole-body PET images are becoming a vital component in early disease detection and staging [1, 2], treatment planning [3, 4] and assessing response to therapy [5–7] in oncology.

Functional volume segmentation of clinical images typically relies on the manual delineation of regions of interest (ROIs) by expert radiologist either directly on PET images or using co-registered anatomical images (CT or MRI) [8, 9]. The accuracy of the manual ROI delineation on PET images is very much dependent on the intensity window chosen for visualization [10]. Additionally, the precision and accuracy of the co-registration procedures are limited [11] and co-registration does not adequately account for movement [12, 13]. Furthermore, anatomical images do not always necessarily relate to the underlying physiological process. Irrespective of the method employed, manual delineation of ROIs is always labourious, highly operator dependent, requires significant knowledge of the local anatomy and may be expected to produce significant intra and interobserver variability [7, 14].

To overcome these limitations of the manual segmentation method, several automatic and semi-automatic segmentation methods have been proposed over the years [15–18]. The validation of the precision and accuracy of these algorithms poses different sorts of challenges [19]. For validation purposes, manual delineation is still considered as the gold standard for clinical images. However, the gold standard manual delineation method has its own limitations, as mentioned earlier. Moreover, since the anatomical boundary of a lesion provided by anatomical CT or MRI images does not necessarily overlap with the functional boundary provided by

PET, the use of anatomical CT or MRI images to validate a PET segmentation algorithm is also restricted. Validating PET image segmentation techniques with macroscopic surgical specimens for clinical studies [19–21] is another alternative, provided that shrinkage of the specimen after surgical excision is appropriately considered [22], and can only be considered as a surrogate of the actual imaging data [23]. However, this validation procedure is not suitable across multiple settings. Experimental or simulated phantom studies can overcome these limitations as the phantom can either be scanned or simulated with different settings [24], and thus phantoms are now widely used to validate PET segmentation algorithms [25–28]. A recent study proposes a reconstruction frame-work for simultaneous estimation of the activity distribution, parametric images and segmentation [29].

The performances of different automatic and semi-automatic segmentation methods have been evaluated using different parameters for different scanners, scanning protocols and reconstruction algorithms by several groups and hence, the selection of a single common parameter for validation is challenging [30]. Moreover, the size, contrast and signal to noise ratio (SNR–representing counts or scan duration) of different lesions are subject to change due to intra and intertumour uptake variability within and between patients both before and after treatment [8, 31]. The changes in contrast and SNR can also result from the use of different radiotracers [32, 33].

The aim of this study is to investigate the robustness of four most commonly used PET image segmentation algorithms with different parameters against variations in size, contrast and SNR using a torso NEMA phantom. The methods were chosen based on the PET segmentation literature. The reference volume was estimated using the calculated boundaries based on the known diameter and position of each sphere of the phantom, which serves as an alternative to the ground truth.

Materials and methods

Phantom data acquisition

The torso NEMA phantom contains six spheres with diameters of 10, 13, 17, 22, 28 and 37 mm, which correspond to volumes of 0.52, 1.15, 2.57, 5.58, 11.49 and 26.52 cm^3, respectively, that were filled with ^{18}F solutions to yield two different contrast ratios between the homogeneous hot spheres and the cold uniform background (2:1 and 4:1). In this article, the diameter and volume will be used interchangeably to indicate a particular sphere. The activity ratio between the spheres and background are shown in Table 1.

The phantom data were acquired in 3D mode on the TrueV PET-CT scanner (Siemens, USA) for 120 minutes, which provides 109 image planes or slices covering a 21.6 cm axial FOV (field of view). The images were reconstructed into a 256×256×109 matrix with voxel dimensions of 2.67×2.67×2.00 mm using an OSEM reconstruction algorithm with 4 iterations and 21 subsets for five different scan durations (900, 1200, 2000 and 4000 seconds corresponding to 15, 20, 33.3 and 66.6 minutes) to represent different levels of noise. The starting time of each static frame was shifted to reconstruct five different non-overlapping and overlapping realizations for all durations (Fig 1). All the reconstructed images were smoothed with a 4-mm

Table 1. Activity concentration of the spheres and background for different contrasts.

	2:1 (kBq/ml)	4:1 (kBq/ml)
Sphere	1668.52	2775.43
Background	838.59	697.24
Measured Ratio	1.99:1	3.98:1

https://doi.org/10.1371/journal.pone.0219127.t001

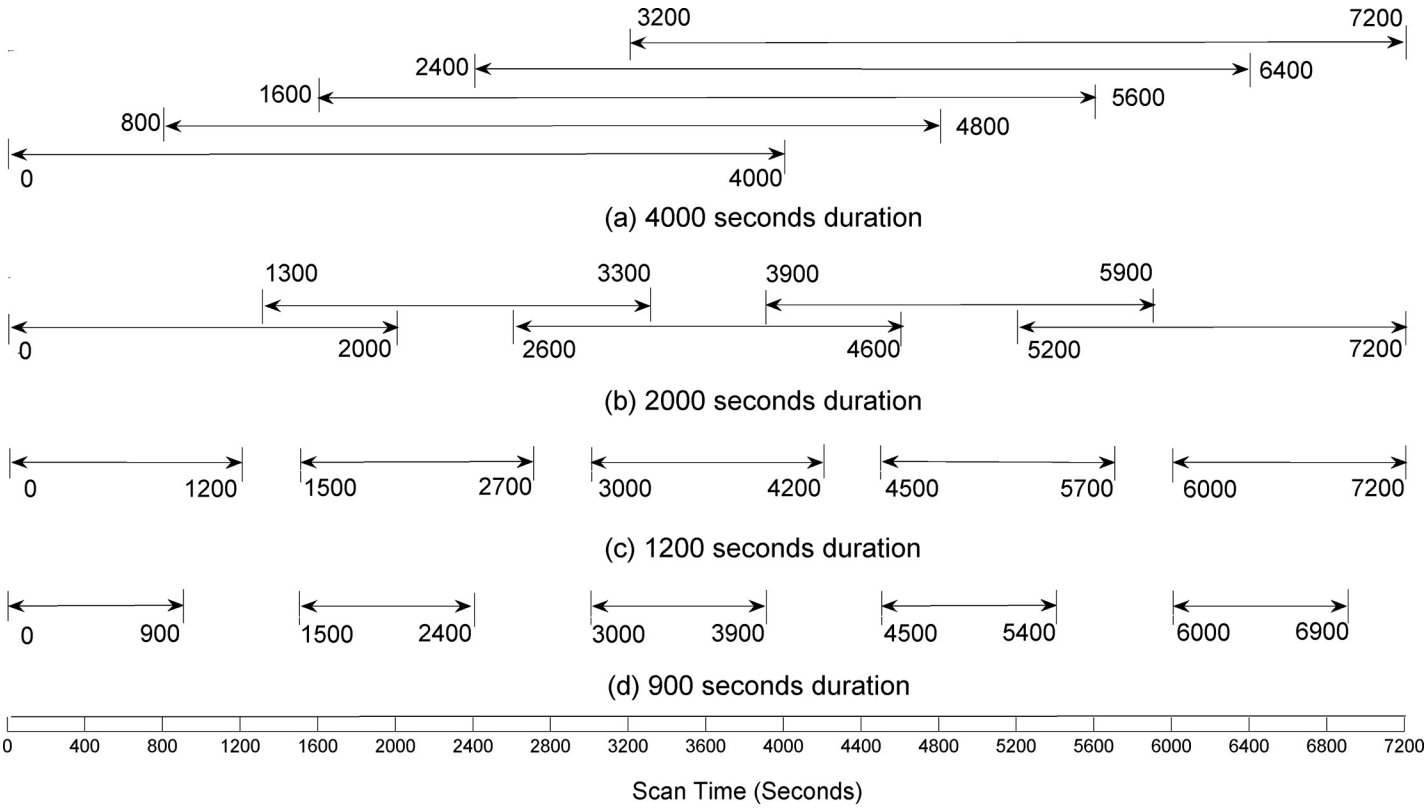

Fig 1. Start and end times to generate five realizations for 900, 1200, 2000, 4000 second reconstruction durations. For the 2000 and 4000 second reconstruction durations, the start and end times of the static frames had to be overlapped to generate five realizations. In total, twenty 3D PET images were reconstructed from 7200 seconds of list-mode data.

https://doi.org/10.1371/journal.pone.0219127.g001

FWHM (full width at half maximum) Gaussian filter after correcting for decay. The true volume of interest (VOI_{True}) was estimated using the boundaries calculated from the known diameter and position of each sphere.

Segmentation methods

All the spheres were delineated using four different segmentation methods. Each segmentation method was applied separately on each roughly delineated volume of interest (VOI) that contained only one sphere to generate the corresponding VOIs. The first delineation method was with a fixed threshold set to 40% (I_{40T}) of the maximum intensity (I_{Max}) within the sphere with the delineated VOI, noted as VOI_{40T} [34, 35]. The second method was a fuzzy c-mean (FCM) clustering method with two clusters to provide a VOI_{FCM} [36]. The FCM method defines the varying degrees of membership of each voxel in multiple clusters and is based on the minimization of the following objective function:

$$D_m = \sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}^{m}\|x_i - c_j\|^2 \quad (1)$$

Where I is the number of data points, J is the number of clusters, x_i is the i^{th} voxel, c_j is the centre of the j^{th} cluster, and μ_{ij}^{m} is the degree of membership of x_i in the j^{th} cluster, with m being a weighting exponent to control for the fuzzy aspect of the image and is usually set to 2. The sphere and background are defined as the two clusters for the purpose of phantom image segmentation.

The third method of estimating the VOI was an adaptive threshold-based method [37], where the adaptive threshold intensity ($I_{Adaptive}$) is given by the following equation:

$$I_{Adaptive} = (\alpha \times I_{70T}) + (\beta \times I_{BG}) \quad (2)$$

where I_{70T} is the mean intensity in a contour containing all voxels with a value greater than 70% of the I_{Max} in the sphere, and I_{BG} is the mean background intensity within a sphere that is 26.52 cm^3 in size and is located away from all the other spheres to avoid partial volume effect (PVE). The α and β parameters for the adaptive threshold were calculated for each acquisition duration using the mean values of the optimal cutoff intensities ($I_{optimal}$) of the five realizations of both contrast ratios. The $I_{Optimal}$ of each hot sphere was derived using the optimal threshold ($T_{Optimal}$) and I_{Max}.

$$I_{Optimal} = T_{Optimal} \times I_{Max} \quad (3)$$

where $T_{Optimal}$ is the percentage threshold value of I_{max} that provides the best matched thresholded volume for the VOI_{True} of each sphere. The optimized values of α and β parameters for all the acquisition durations are shown in Table 2.

The threshold value ($T_{Adaptive}$) is then defined by the percentage ratio of $I_{Adaptive}$ to I_{Max}.

$$T_{Adaptive} = 100 \times \frac{I_{Adaptive}}{I_{Max}} \quad (4)$$

Three different ways to calculate the α and β parameters were considered and compared. First, the adaptive threshold-based volume of interest ($VOI_{Adaptive}$) was delineated using the $I_{Adaptive}$ value that was derived using the α and β parameters of each individual acquisition duration. Two other different volumes of interests, $VOI_{A\text{-}900}$ and $VOI_{A\text{-}4000}$, which correspond to two different $T_{Adaptive}$ values derived from $I_{A\text{-}900}$ and $I_{A\text{-}4000}$, were generated for all spheres and acquisition durations using only the α and β values obtained from the 900 second (A-900) and 4000 second (A-4000) acquisition duration data, respectively. The purpose of generating three different segmented volumes of interest using the adaptive threshold method was to investigate the effects of noise on α and β.

The final segmentation method considered was region based variational method proposed by Chan and Vese (C-V) [38] to provide the segmented volume $VOI_{C\text{-}V}$. This method was considered over an edge-based method [39] because the boundaries of the lesions in the PET images cannot necessarily be defined by a gradient. The method works by minimizing an energy function $E(c_1,c_2,C)$ related to a particular segment of the image $\Omega(x,y,z)$. The variable curve, C segregates the images Ω in two regions c_1 and c_2. The energy function $E(c_1,c_2,C)$ is given by the following equation:

$$\begin{aligned} E(c_1, c_2, C) = \mu.\mathrm{Length}(C) + \vartheta.\mathrm{area}(\mathrm{inside}(C)) + \lambda_1 \int_{\mathrm{inside}(C)} |\Omega(x, y, z) - c_1|^2 dxdydz \\ + \lambda_2 \int_{\mathrm{outside}(C)} |\Omega(x, y, z) - c_2|^2 dxdydz \end{aligned} \quad (5)$$

Table 2. Values of and for adaptive thresholds of the different acquisition durations.

Duration	α	β
900 Seconds	0.40	0.59
1200 Seconds	0.41	0.57
2000 Seconds	0.42	0.59
4000 Seconds	0.44	0.52

https://doi.org/10.1371/journal.pone.0219127.t002

where $\mu \geq 0, \vartheta \geq 0, \lambda_1, \lambda_2 > 0$ are fixed parameters and are typically fixed to $\lambda_1 = \lambda_1 = 1$ and $\vartheta = 0$. The method will be referred to as the C-V method.

Performance analysis metrics

Along with the conventional measurements of change in the delineated VOIs due to changes in sphere size, noise and contrast, the percent bias of segmented volume, Dice similarity coefficient (DSC) and classification error (CE) were also analysed for each segmentation method. The percent bias is calculated with the following equation:

$$\%\text{Bias} = 100 \times \frac{(\text{VOI}_{\text{Mean}} - \text{VOI}_{\text{True}})}{\text{VOI}_{\text{True}}} \tag{6}$$

where is the mean of the segmented volumes of the five realizations, and DSC provides quantitative measures of the spatial overlap index with VOI_{True}. DSC can be used to evaluate the segmentation accuracy, and is given by the following equation:

$$\text{DSC} = \frac{2(\text{VOI}_{\text{True}} \cap \text{VOI}_{\text{Seg}})}{\text{VOI}_{\text{True}} + \text{VOI}_{\text{Seg}}} \tag{7}$$

where $\cap$ is the intersection, and is the segmented volume. A DSC value of 0 indicates complete non-overlap, and a value of 1 indicates a complete match or overlap between the two volumes.

Classification error (CE) is defined as the following:

$$\text{CE} = 100 \times \frac{(\text{PCE} + \text{NCE})}{\text{VOI}_{\text{True}}} \tag{8}$$

where PCE (positive classification error) refers to the background voxels that are classified as voxels belonging to the sphere. In contrast, NCE (negative classification error) refers to the voxels within the sphere belonging to the background. A high CE value is indicative of poor segmentation accuracy.

Results

For the threshold-based segmentation, different threshold intensities (I_{Optimal}, I_{40T} and I_{Adaptive}) are estimated as a fraction of I_{Max} within the lesion; thus, it is important to understand the relationship of these intensities with I_{Max} and their dependencies on lesion size and varying data conditions, e.g., contrast and noise. The mean of I_{Max}, I_{Optimal}, I_{40T} and I_{Adaptive} of the five realizations as a function of the log of the segmented volume of each method is shown in Fig 2. I_{Max} is the lowest for the 10 mm sphere and highest for the 37 mm sphere for all acquisition durations (55% to 70% difference between these two spheres based on acquisition durations). However, for any given size, I_{Max} is the lowest for 4000 second acquisitions (6.34 for 37 mm sphere for a contrast of 2:1) followed by 2000, 1200 and 900 second acquisitions (6.71, 7.16 and 7.47 respectively, approximately 18% difference between the 4000 and 900 second acquisition durations). The differences between I_{Max} for the different acquisition durations decrease as the size of sphere decreases (e.g., 3.91, 4, 4.14 and 4.31 are the intensity values for 4000, 2000, 1200 and 900 second acquisition durations for the smallest sphere for contrast 2:1 yielding a 10% difference between the 4000 and 900 second durations). The difference is even smaller for the smallest sphere for contrast 4:1 (6.60, 6.67, 6.64 and 6.38 for 4000, 2000, 1200 and 900 second acquisition durations, respectively, a 3% difference).

Since I_{Max} is higher for longer acquisition durations (i.e., low noise), the 40T threshold value is always higher for these acquisition durations that for the short durations for lesions

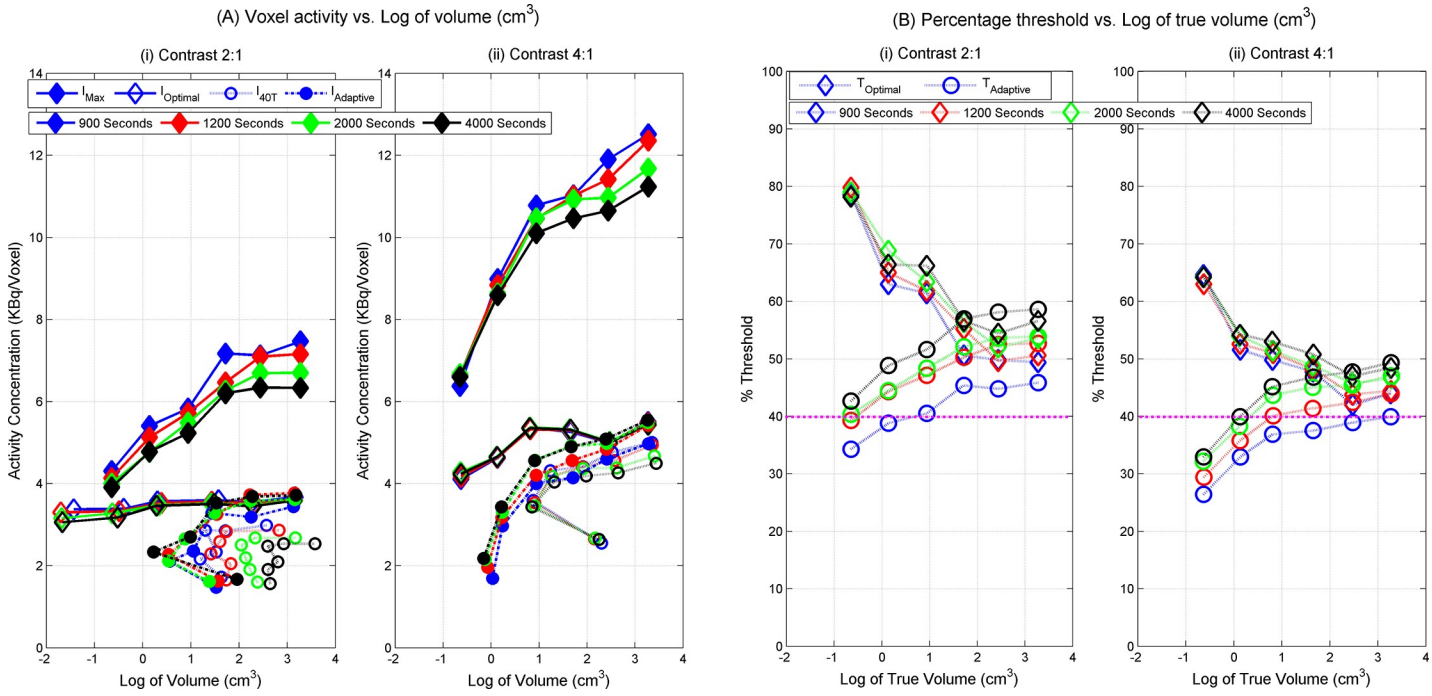

Fig 2. Voxel activity and percent threshold against the log of the volumes. (A) Voxel activity against the log of the volume for both contrasts for all acquisition durations. The log of the volume was used to highlight the separation between the small spheres. $I_{Optimal}$, I_{40T} and $I_{Adaptive}$ are the intensities derived by applying the optimal, 40% and adaptive thresholds, respectively. (B) The optimal ($T_{Optimal}$) and adaptive ($T_{Adaptive}$) percentage thresholds against the log of the volume for both contrasts and acquisition durations.

https://doi.org/10.1371/journal.pone.0219127.g002

larger than 17 mm in diameter for both contrasts. The VOI_{40T} for the 4000 second acquisition duration is approximately 80% bigger than that of 900 second acquisition duration (Figs 3 and 4). For the same given activity, I_{Max} is always lower for smaller volumes due to PVEs [40]. I_{Max} increases as the volume increases and remains the same after a certain volume, especially for low noise cases.

$I_{Optimal}$ shows less noise dependency, and the values are similar for both contrasts irrespective of the acquisition duration (ranging from 3.06 for 4000 seconds and 10 mm spheres to 3.68 for 900 seconds and 37 mm spheres for a contrast of 2:1 with a maximum difference of 25%, whereas a maximum difference of 75% is observed for I_{Max}). The difference amongst acquisition duration is also less noticeable, especially for contrast 4:1. Since $I_{Optimal}$ is related to I_{Max} via $T_{Optimal}$ (Eq 3), $T_{Optimal}$ increases with acquisition duration, decreases with size and is inversely related to I_{Max} to compensate for the noise for shorter acquisition durations. I_{40T} is the highest for the 900 second duration (ranging from 1.72 to 2.99, depending on the size, for a contrast of 2:1 and 2.55 to 5 for a contrast of 4:1) and lowest for the 4000 second duration (1.57 to 2.54 for a contrast of 2:1 and 2.64 to 4.50 for a contrast of 4:1); the values become stable for spheres larger than 17 mm in diameter. The values of I_{40T} are always lower than those for $I_{Optimal}$. The differences between I_{40T} and $I_{Optimal}$ decrease the most for the 900 second acquisition durations, followed by the 1200, 2000 and 4000 second durations for the biggest three spheres. The differences also decrease as the size of the spheres increases for a contrast of 4:1.

$I_{Adaptive}$, estimated using Eq 2, has the highest value for the 4000 second duration (ranging from 1.67 to 3.72 for a contrast of 2:1 and 2.28 to 5.55 for a contrast of 4:1, depending on the size of the sphere) and the lowest value for the 900 second duration (1.48 to 3.44 for a contrast of 2:1 and 1.70 to 4.98 for a contrast of 4:1); the values are in reverse order in terms of

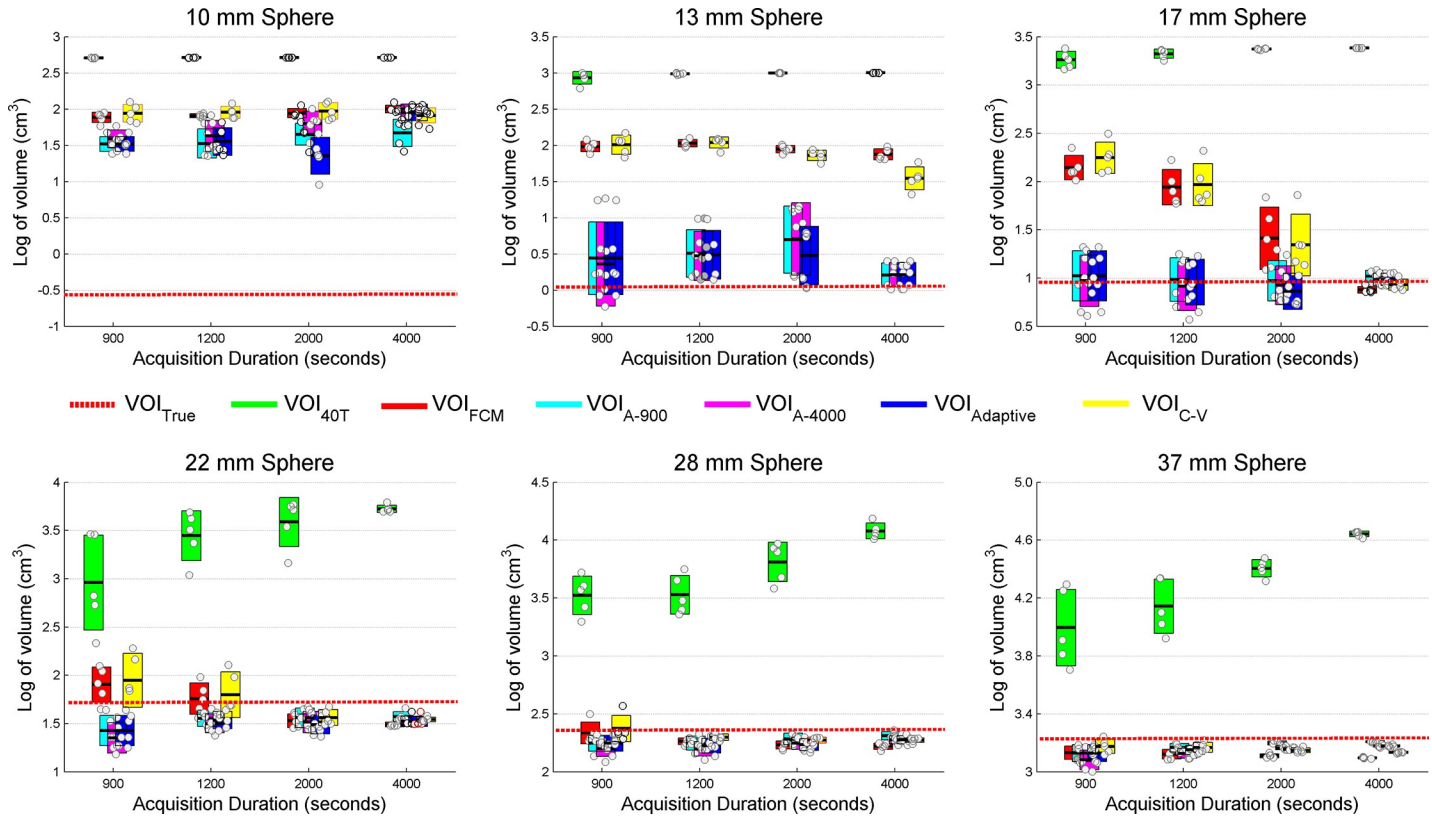

Fig 3. Log of all the segmented volumes for contrast 2:1. Each circle represents the log of the segmented volume of each realization. The thick black line within each coloured rectangle represents the mean of the five realizations of each segmentation method. The maximum and minimum limit of each rectangle is represented by the meanSD (standard deviation). The dotted red line is the log of the true volume.

https://doi.org/10.1371/journal.pone.0219127.g003

acquisition durations compared to those of I_{Max} and I_{40T}. The values of $I_{Adaptive}$ do not change significantly for spheres larger than 5.58 cm^3 (equivalent log volume of 1.72 cm^3) for a contrast of 2:1. Since $T_{Adaptive}$ is directly proportional to $I_{Adaptive}$ (Eq 4) and inversely proportional to I_{Max}, the values of $T_{Adaptive}$ increase with the acquisition duration. $T_{Adaptive}$ has the highest value for the 4000 second duration (ranging from 47.45 to 61.09 for a contrast of 2:1 and 34.03 to 48.02 for a contrast of 4:1, depending on the size) and the lowest value for the 900 second duration (38.83 to 48.43 for a contrast of 2:1 and 28.03 to 39.23 for a contrast of 4:1). The order of $T_{Optimal}$ and $T_{Adaptive}$ values are opposite to each other with respect to acquisition duration.

The log of the volumes of all five realizations segmented by different methods along with the mean and standard deviations are shown in Figs 3 and 4 for contrasts of 2:1 and 4:1, respectively. The representative segmentation results of the three methods (40T, FCM, adaptive and C-V) along with the true volume of the 28 mm sphere for both contrasts are shown in Fig 5.

The 40T method always overestimates the volumes. The overestimation is consistent irrespective of the acquisition duration for 10 mm, 13 mm and 17 mm spheres for both contrasts, and the segmented volume is several times larger than the true volume and matches the size of the roughly delineated VOI. For large spheres, the overestimation of the volumes by 40T increases as the acquisition duration increases for contrast 2:1 (+110% to 291% bias based on acquisition durations for the 37 mm sphere). The differences in overestimation between acquisition durations decreases with the increase in contrast (+8 to 17% bias for a contrast of 4:1 for the same sphere).

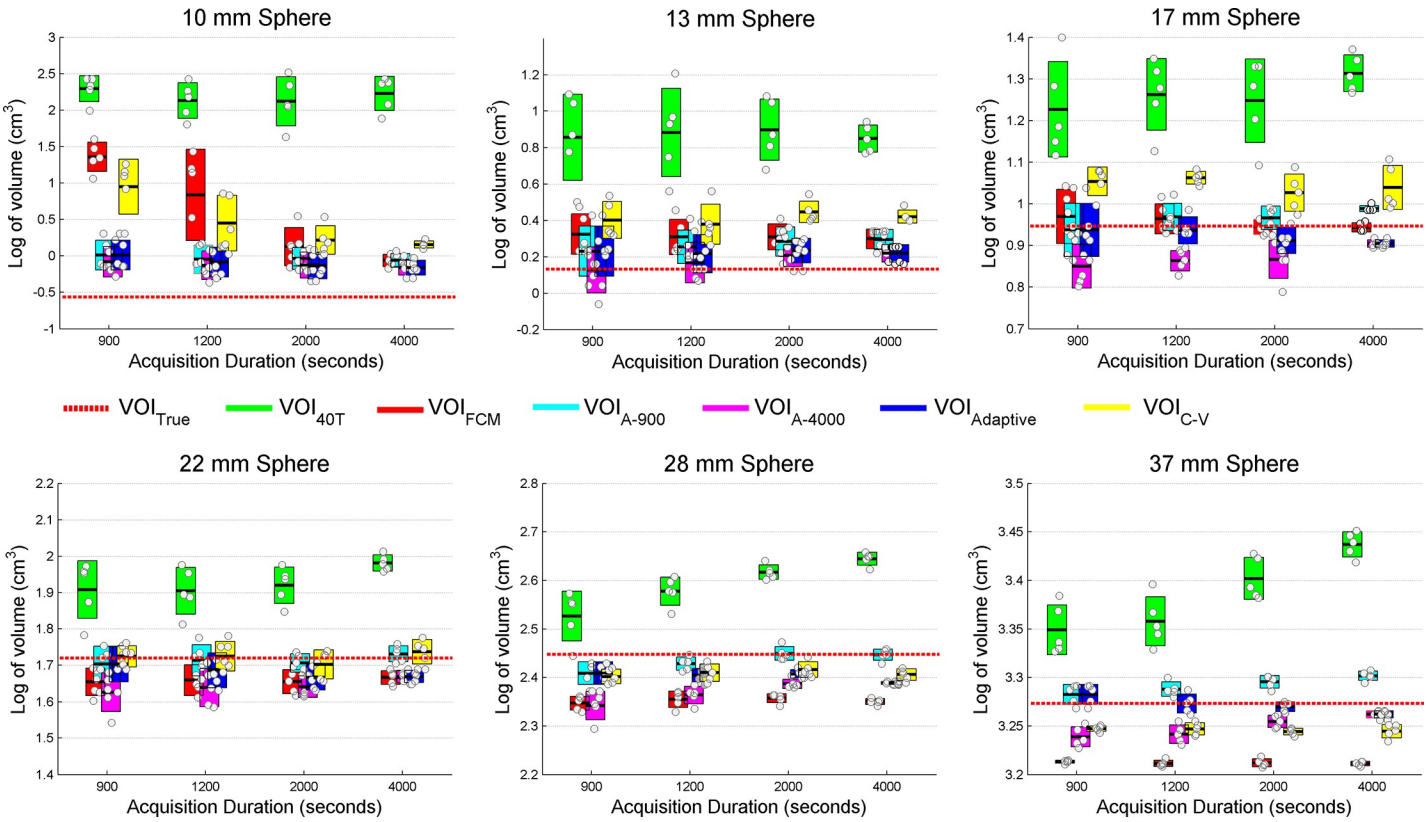

Fig 4. Log of all the segmented volumes for contrast 4:1. Each circle represents the log of the segmented volume of each realization. The thick black line within each coloured rectangle represents the mean of the five realizations of each segmentation method. The maximum and minimum limit of each rectangle is represented as the meanSD (standard deviation). The dotted red line is the log of the true volume.

https://doi.org/10.1371/journal.pone.0219127.g004

For a contrast of 2:1 (Fig 3), both the FCM and C-V methods significantly overestimate the volumes for 10 mm and 13 mm spheres. The rate of volume overestimation decreases as the contrast increases (from 13% to 5% for the 37 mm sphere, as shown in Figs 3 and 4). For the 17 mm sphere, the FCM and C-V methods overestimate the volumes for 900, 1200 and 2000 second acquisition durations for a contrast of 2:1 (+233% to 66% bias based on the acquisition durations). At the same contrast level, both the FCM and C-V methods match the true volume more closely for spheres larger than 17 mm in diameter than for small spheres, with a bias of 15%. However, the mismatch of volumes segmented by the FCM and C-V methods is further reduced for a contrast of 4:1 for large spheres. For large spheres, both methods are less dependent on noise. Similar to the FCM method, all the three adaptive methods show less dependency on the acquisition durations, and the volumes estimated by the different methods closely match with the true volume, except for the smallest two spheres (approximately 7% across acquisition durations). The differences in volume estimation between the three different adaptive threshold methods with different α and β parameters are not noticeable.

Reproducibility, as represented by standard deviation (SD), of the five realizations is shown in Figs 3 and 4 for contrasts of 2:1 and 4:1, respectively. Fig 3 shows that for a 2:1 contrast, the segments from the 40T method are roughly the same delineated areas for the three smallest spheres, hence the reproducibility of these spheres are not evaluable. For the three biggest spheres, the SD for VOI_{40T} decreases as the acquisition duration increases (14.93, 11.95, 4.86 and 1.97 for 900, 1200, 2000 and 4000 second acquisition duration, respectively, for the 37 mm

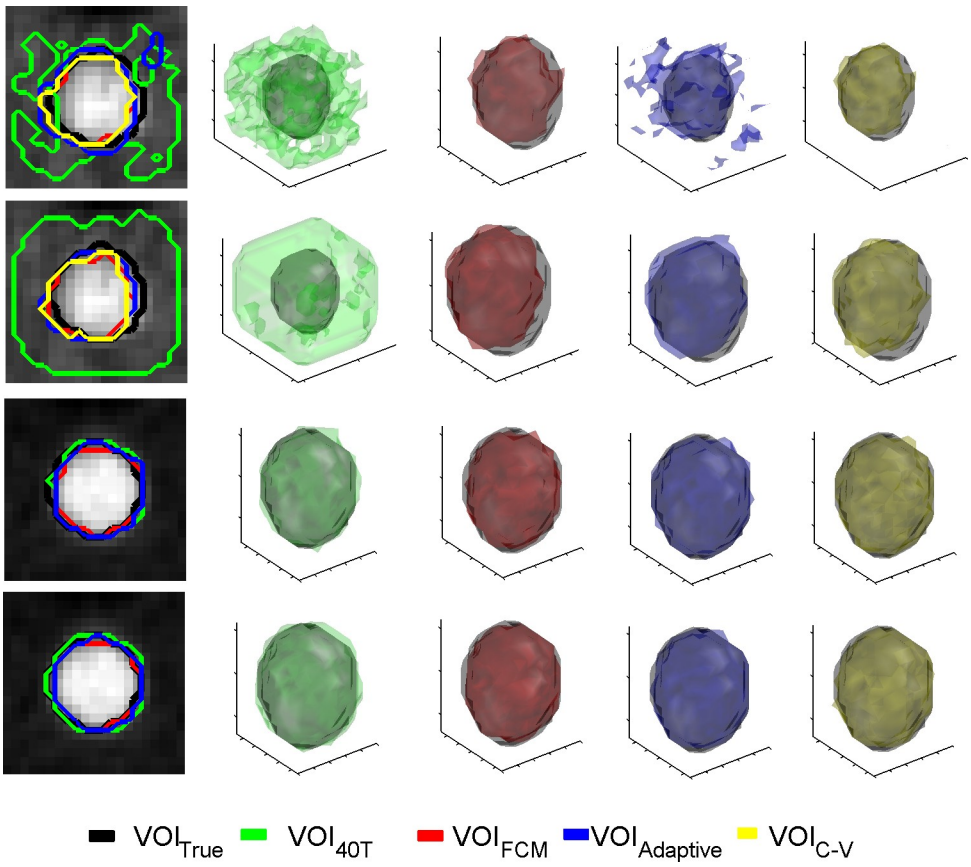

Fig 5. Representative contours and VOIs of the four segmentation methods along with the true volume for a 28 mm sphere. First row: contrast 2:1 and 900 seconds; second row: contrast 2:1 and 4000 seconds; third row: contrast 4:1 and 900 seconds; and fourth row: contrast 4:1 and 4000 seconds.

https://doi.org/10.1371/journal.pone.0219127.g005

sphere). The reproducibility improved between 2 between the different acquisition durations compared to that of the 900 second acquisition. With an increase in contrast, the performance of the 40T method improves for the smallest three spheres. The performance of the 40T method for the other three spheres are also significantly improved (SD 0.74, 0.73, 0.65 and 0.40 for 900, 1200, 2000 and 4000 second acquisition duration, respectively, for the 37 mm sphere). In this case, the reproducibility improved between 1 between the different acquisition durations compared to that of the 900 second acquisition. The reproducibility between the contrasts of 2:1 and 4:1 for the biggest three spheres increases between 80 to 95%, with the biggest improvement being observed for the shortest acquisition duration of 900 seconds. The reproducibility for the FCM, C-V and adaptive threshold methods significantly improves for these spheres compared to that of the 40T method. For the 2:1 contrast, all three methods, Adaptive, C-V and FCM, provide similar SDs ranging between 0.52 to 0.84 (an improvement of 60–95% based on the acquisition duration compared to that of the 40T method). At a contrast of 4:1, the SD range for these methods is 0.07 to 0.19—a reduction of 75 to 87% compared to that at a contrast of 2:1, and the biggest improvement is observed for the shortest acquisition duration of 900 seconds. These results indicate that with the increase in contrast, the reproducibility of the volume segmentation increases, and the improvement is more remarkable for high noise conditions.

The mean DSCs of the five realizations for all methods at both contrasts are shown in Fig 6. The DSC increases as the size of the sphere and contrast increase for all methods. The DSCs for the FCM, A-900, A-4000, adaptive and C-V methods increase as the acquisition duration increases. However, the DSC for the 40T method decreases as the acquisition duration increases for spheres larger than 17 mm, especially for a contrast of 2:1. For 22 mm, 28 mm and 37 mm spheres, the differences in DSCs between the FCM, A-900, A-4000, adaptive and C-V methods are insignificant at both contrasts. At a contrast level of 2:1, the DSC for the FCM method is smaller than that for the adaptive threshold-based methods for the 10 mm, 13 mm and 17 mm spheres. A similar trend is observed for the C-V method. The 40T method always provides a lower DSC than all other methods, except with the 28 mm and 37 mm spheres at a contrast level of 4:1.

The mean percentage CEs, along with the standard deviations as error, are shown as bar graphs in Fig 7. The CE decreases as the size of the sphere and contrast increase for the FCM, A-900, A-4000, adaptive and C-V methods. In contrast, the CE increases as the acquisition duration increases but decreases with high contrast for the 40T method. The CE for the 40T method is always higher than that of the other four segmentation methods, except for a 900 second acquisition duration with the 28 mm sphere at a contrast of 4:1. The performance of both the FCM and C-V methods is inferior compared to that of the different adaptive methods for the 13 mm and 17 mm spheres at a contrast of 2:1 and for the 10 mm sphere at a contrast of 4:1. For the 13 mm, 17 mm and 22 mm spheres at a contrast of 4:1, the FCM method performs better than the C-V method, and the performance is comparable to that of all the adaptive methods in terms of percentage CE.

Discussion

PET is currently being used for different clinical purposes ranging from diagnosis [1] and treatment planning [3] to early response assessments [8]. To exploit the full potential of PET for reliable clinical outcomes, robust and accurate delineations of lesions from PET images are vital. Fully automatic and semi-automatic segmentation methods are being developed to remove the influence of intra- and interobserver variability in PET lesion segmentation. The accuracy and robustness of these segmentation methods have generally been assessed against certain criteria [19]. However, these criteria are not fixed for different tracers and can change for different clinical settings. The objective of this study is to quantitatively assess the performance of four different PET lesion segmentation methods for different statistical settings to determine the most suitable segmentation method against statistical fluctuations. Before segmenting images with the four segmentation methods used in this study for comparison, all images were smoothed with a 4 mm Gaussian filter to reduce the effects of noise.

This investigation confirms that though the widely used fixed threshold-based automatic segmentation method is straightforward to implement, it is highly dependent on the maximum intensity within the lesion (I_{Max}). I_{Max} is also dependent on the size of lesion, contrast and acquisition duration [24] and thus, the segmentation results can differ with the variations of noise in the image. For small spheres, the PVE has more significant influence than acquisition duration and therefore, the segmented volume using a fixed threshold of 40% does not depend on the noise, and roughly, the segmented volume is generally the same delineated area. This study also confirms that the magnitude of overestimation, and hence voxel classification errors, using the 40T method result from the combined effects of the size of the lesion and contrast [34] as well as the acquisition duration [41]. Because of this reason, for small objects where PVE plays a main role, the 40% fixed threshold may not provide bias-free estimates of the volumes. Though the performance of the 40% threshold method may improve with high

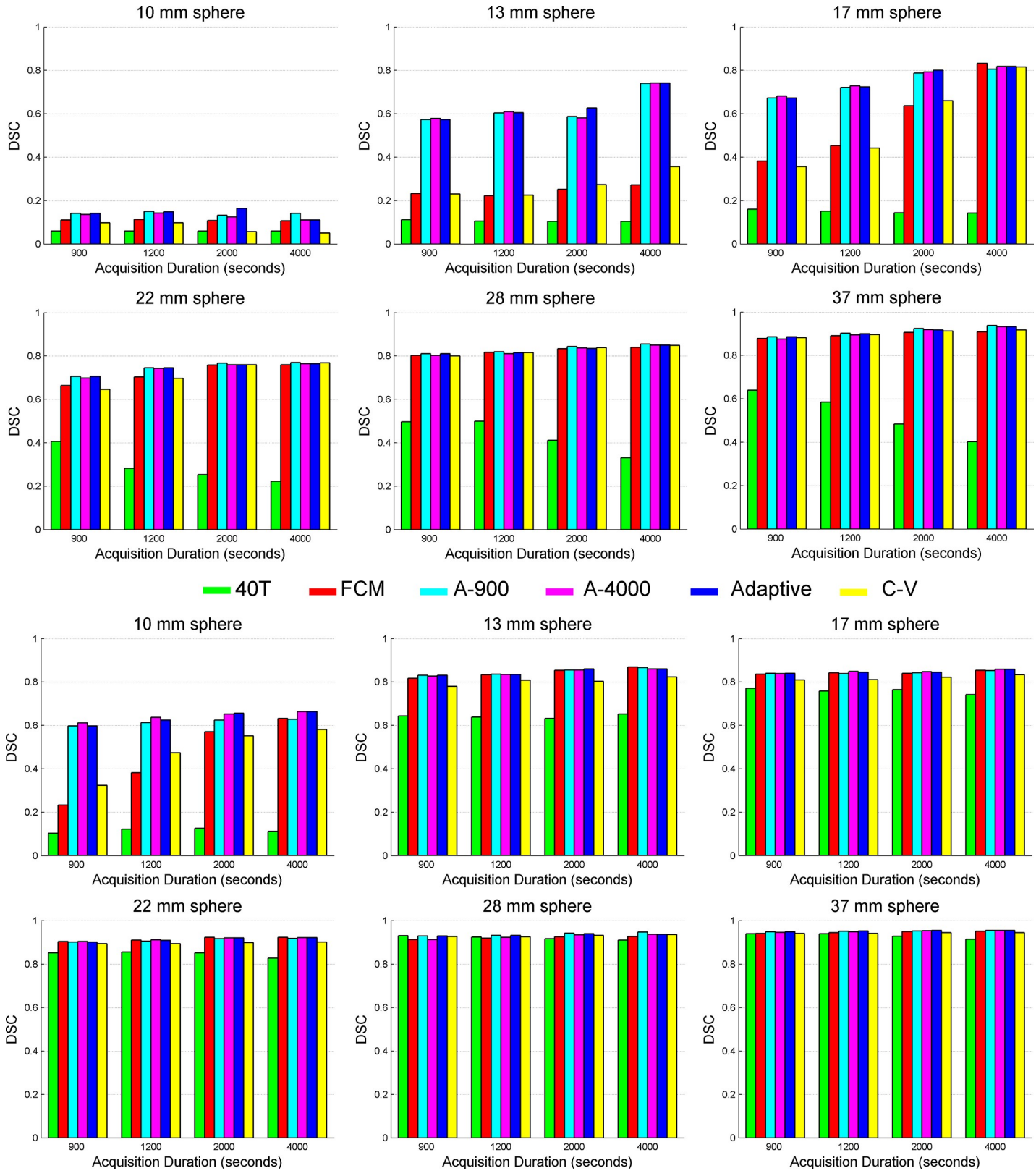

Fig 6. Mean DSCs of the five realizations for all six spheres at a contrast of 2:1 (top two rows) and a contrast of 4:1 (bottom two rows) for all segmentation methods considered (40T, FCM, A-900, A-4000, adaptive and C-V).

https://doi.org/10.1371/journal.pone.0219127.g006

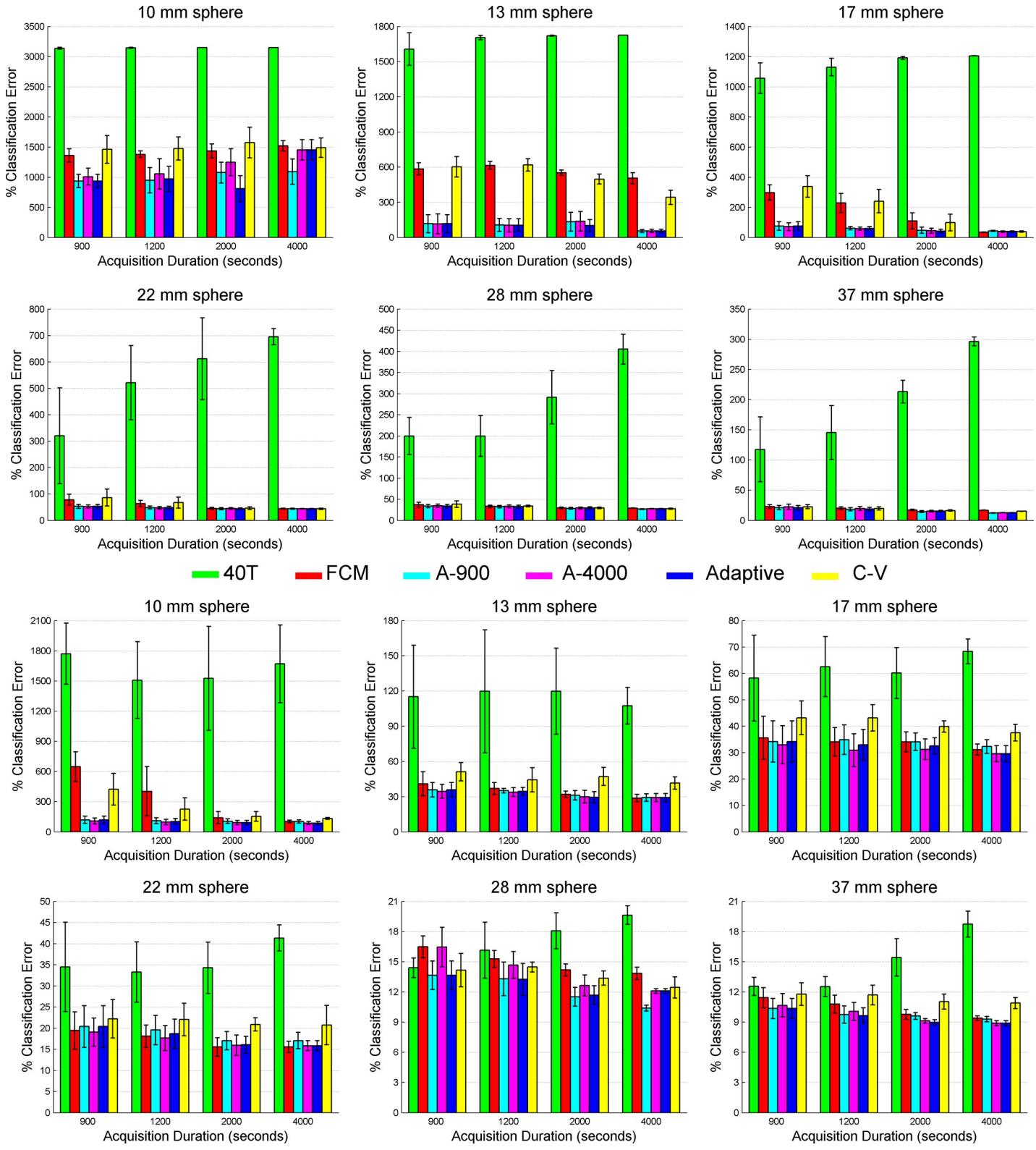

Fig 7. Mean percentage classification error of the five realizations for all six spheres at a contrast of 2:1 (top two rows) and a contrast of 4:1 (bottom two rows) for all segmentation methods (40T, FCM, A-900, A-4000, adaptive and C-V). The error bars indicate the standard deviations of the means.

https://doi.org/10.1371/journal.pone.0219127.g007

contrast and large spheres, the problem of volume overestimation will still remain. Another finding of this study is that the fixed threshold-based segmentation method is a suboptimal automatic segmentation method that does not provide the same VOI in response to only a reduction in contrast. The fixed threshold-based segmentation method also provides suboptimal segmented volumes even for two different lesions of same size in the same patient with varying contrasts.

The standard deviations of the five realizations representing the reproducibility of the segmented volumes indicate that reproducibility increases as the acquisition duration increases for all the methods. However, the reproducibility of the volumes with the fixed threshold method is the lowest, and the FCM, C-V and adaptive methods have similar performances for the largest three spheres (>17 mm).

The adaptive threshold method removes some of the aforementioned limitations by allowing the threshold value to be determined by the initial rough estimate of the activity within the lesion (I_{70T}) and background (I_{BG}). To determine the threshold value, the method also requires the calculations of two parameters (and) using the optimal threshold ($I_{Optimal}$), where $I_{Optimal}$ is defined as the percentage of I_{Max} that provides the volume closest to the true volume. Since the true volume for the tumour is not known, a phantom with different sizes of spheres with known volumes are generally used to determine $I_{Optimal}$. In this study, the effects of noise on the parameters and that are used to determine the adaptive threshold intensity have been investigated. The values of decrease and increase as noise increases. In contrast, I_{70T} increases with more noise. Since $I_{Adaptive}$ is related to I_{70T} via, the influence of noise is lower for the adaptive threshold method as decreases with increasing noise, which minimizes the effect increasing I_{70T} values due to noise. Nonetheless, $I_{Adaptive}$ still shows noise dependency for high noise and low contrast conditions as $I_{Adaptive}$ is related to I_{Max} via I_{70} (Eq 2).

The optimal results are obtained when the and values correspond to the same noise level. The segmentation results using and values derived from the longest acquisition duration closely match with the optimal results. Since it is cumbersome to derive and values for each noise level, the and values estimated using the longest acquisition duration can be used for the other noise levels. Though an adaptive threshold can minimize the effects of noise on the segmented volumes for spheres larger than 13 mm diameter, one of its major drawbacks is that the and parameters need to be calibrated for different PET scanners and data acquisition protocols.

The segmentation volumes estimated using the FCM and C-V methods are less dependent on noise. However, the dependency of the FCM and C-V methods is higher for low contrast and smaller spheres (less than 2 cm diameter) compared to that of the adaptive method. A similar observation has previously been reported for the FCM method [42]. The FCM, C-V and adaptive threshold methods overestimate volumes for small objects. However, the overestimation is small compared to that of the fixed threshold method. As the sphere size increases, both methods provide bias-free volume estimations, which is in accordance with previous findings.

For all segmentation methods, the reproducibility, bias DSC and CE are at their lowest and the dependency on volume and contrast are at their highest for a typical clinical scan duration of 15 minutes, which is equal to a 900 second scan duration, compared to those of longer acquisition durations for the same object size and contrast level.

Conclusion

In this study, the performance of four different PET volume segmentation methods against statistical fluctuations were compared using a torso NEMA phantom. The study demonstrates that the differences in performance between all the methods decreases as the object size,

contrast and acquisition duration increases. The fixed threshold method always overestimates the segmented volume, and its overall performance is inferior compared to that of all the other methods. The fixed threshold method is also the most sensitive to statistical fluctuations and hence should not be used when statistical fluctuations are expected (e.g., for monitoring response). The FCM, C-V and all adaptive threshold-based methods provide similar improved performance compared to the fixed threshold-based method. The adaptive threshold method with individually calculated parameters for each acquisition duration outperforms all other methods and is the most robust method against statistical fluctuations in the PET data. The drawbacks of this method are that its performance is not optimal for small volumes (less than 17 mm in diameter) with low contrast and the method also requires a calibration for every PET scanner, acquisition protocol and acquisition duration or counts. In contrast, the adaptive threshold method with parameters derived from the longest acquisition duration has a similar performance to the other methods and only needs to be optimized once for each PET scanner and acquisition protocol. The performances of the FCM and C-V methods are similar to those of the adaptive methods for spheres larger than 2 cm in diameter, and these methods do not require calibrations. Therefore, the FCM and C-V methods are the most suitable in cases where the tracer uptake is expected to vary across tumours, patients or tracers and the tumour is not expected to shrink in size in response to treatment. Furthermore, both these methods are only inferior to the adaptive method in cases where both the size of the lesion and contrast are very low. This study also highlights the importance of assessing the robustness of automatic PET segmentation methods against statistical fluctuations (e.g., volume, contrast, acquisition durations, etc.), especially if the method is going to be used for delineating tumours for different radiotracers with variable uptake as well as for assessing treatment response. The study also suggests that the reproducibility, accuracy and robustness of the automatic PET segmentation methods are still not reliable for low contrast levels and small lesions with diameters less than 22 mm.

Author Contributions

Conceptualization: Mahbubunnabi Tamal.

Data curation: Mahbubunnabi Tamal.

Formal analysis: Mahbubunnabi Tamal.

Funding acquisition: Mahbubunnabi Tamal.

Investigation: Mahbubunnabi Tamal.

Methodology: Mahbubunnabi Tamal.

Validation: Mahbubunnabi Tamal.

Writing – original draft: Mahbubunnabi Tamal.

References

1. Rohren EM, Turkington TG, Coleman RE. Clinical applications of PET in oncology. Radiology. 2004; 231(2):305–32. https://doi.org/10.1148/radiol.2312021185 PMID: 15044750.
2. Valk PE, BD L., TD W., MM N. Positron Emission Tomography: Basic Science and Clinical Practice: Springer; 2003.
3. Jarritt PH, Carson KJ, Hounsell AR, Visvikis D. The role of PET/CT scanning in radiotherapy planning. Br J Radiol. 2006; 79 Spec No 1:S27–35. https://doi.org/10.1259/bjr/35628509 PMID: 16980683.
4. Gregoire V, Haustermans K, Geets X, Roels S, Lonneux M. PET-based treatment planning in radiotherapy: a new standard? J Nucl Med. 2007; 48 Suppl 1:68S–77S. PMID: 17204722.

5. Weber WA, Petersen V, Schmidt B, Tyndale-Hines L, Link T, Peschel C, et al. Positron emission tomography in non-small-cell lung cancer: prediction of response to chemotherapy by quantitative assessment of glucose use. J Clin Oncol. 2003; 21(14):2651–7. https://doi.org/10.1200/JCO.2003.12.004 PMID: 12860940.
6. Kinahan PE, Fletcher JW. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. Semin Ultrasound CT MR. 2010; 31 (6):496–505. https://doi.org/10.1053/j.sult.2010.10.001 PMID: 21147377; PubMed Central PMCID: PMC3026294.
7. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. Eur J Nucl Med Mol Imaging. 2005; 32(3):294–301. https://doi.org/10.1007/s00259-004-1566-1 PMID: 15791438.
8. Trigonis I, Koh PK, Taylor B, Tamal M, Ryder D, Earl M, et al. Early reduction in tumour [18F]fluorothymidine (FLT) uptake in patients with non-small cell lung cancer (NSCLC) treated with radiotherapy alone. Eur J Nucl Med Mol Imaging. 2014; 41(4):682–93. https://doi.org/10.1007/s00259-013-2632-3 PMID: 24504503; PubMed Central PMCID: PMC3955141.
9. Hogenauer M, Brendel M, Delker A, Darr S, Weiss M, Bartenstein P, et al. Impact of MRI-based Segmentation Artifacts on Amyloid- and FDG-PET Quantitation. Curr Alzheimer Res. 2016; 13(5):597–607. PMID: 27025775.
10. Chen K, Bandy D, Reiman E, Huang SC, Lawson M, Feng D, et al. Noninvasive quantification of the cerebral metabolic rate for glucose using positron emission tomography, 18F-fluoro-2-deoxyglucose, the Patlak method, and an image-derived input function. J Cereb Blood Flow Metab. 1998; 18(7):716–23. https://doi.org/10.1097/00004647-199807000-00002 PMID: 9663501.
11. Kiebel SJ, Ashburner J, Poline JB, Friston KJ. MRI and PET coregistration—a cross validation of statistical parametric mapping and automated image registration. Neuroimage. 1997; 5(4 Pt 1):271–9. https://doi.org/10.1006/nimg.1997.0265 PMID: 9345556.
12. Riegel AC, Bucci MK, Mawlawi OR, Johnson V, Ahmad M, Sun X, et al. Target definition of moving lung tumors in positron emission tomography: correlation of optimal activity concentration thresholds with object size, motion extent, and source-to-background ratio. Medical physics. 2010; 37(4):1742–52. https://doi.org/10.1118/1.3315369 PMID: 20443495; PubMed Central PMCID: PMC3820629.
13. 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(2):209–25. https://doi.org/10.1016/j.radonc.2006.09.011 PMID: 17064802.
14. Maroy R, Boisgard R, Comtat C, Frouin V, Cathier P, Duchesnay E, et al. Segmentation of rodent whole-body dynamic PET images: an unsupervised method based on voxel dynamics. IEEE Trans Med Imaging. 2008; 27(3):342–54. https://doi.org/10.1109/TMI.2007.905106 PMID: 18334430.
15. Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. Comput Biol Med. 2014; 50:76–96. https://doi.org/10.1016/j.compbiomed.2014.04.014 PMID: 24845019; PubMed Central PMCID: PMC4060809.
16. Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging. 2010; 37(11):2165–87. https://doi.org/10.1007/s00259-010-1423-3 PMID: 20336455.
17. Day E, Betler J, Parda D, Reitz B, Kirichenko A, Mohammadi S, et al. A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients. Medical physics. 2009; 36(10):4349–58. https://doi.org/10.1118/1.3213099 PMID: 19928065.
18. Li H, Thorstad WL, Biehl KJ, Laforest R, Su Y, Shoghi KI, et al. A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours. Medical physics. 2008; 35(8):3711–21. https://doi.org/10.1118/1.2956713 PMID: 18777930; PubMed Central PMCID: PMC3304493.
19. Zaidi H, Abdoli M, Fuentes CL, El Naqa IM. Comparative methods for PET image segmentation in pharyngolaryngeal squamous cell carcinoma. Eur J Nucl Med Mol Imaging. 2012; 39(5):881–91. https://doi.org/10.1007/s00259-011-2053-0 PMID: 22289958; PubMed Central PMCID: PMC3326239.
20. Daisne JF, Duprez T, Weynand B, Lonneux M, Hamoir M, Reychler H, et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology. 2004; 233(1):93–100. https://doi.org/10.1148/radiol.2331030660 PMID: 15317953.
21. Stroom J, Blaauwgeers H, van Baardwijk A, Boersma L, Lebesque J, Theuws J, et al. Feasibility of pathology-correlated lung imaging for accurate target definition of lung tumors. Int J Radiat Oncol Biol Phys. 2007; 69(1):267–75. https://doi.org/10.1016/j.ijrobp.2007.04.065 PMID: 17707281.

22. Dahele M, Hwang D, Peressotti C, Sun L, Kusano M, Okhai S, et al. Developing a methodology for three-dimensional correlation of PET-CT images and whole-mount histopathology in non-small-cell lung cancer. Curr Oncol. 2008; 15(5):62–9. https://doi.org/10.3747/co.v15i5.349 PMID: 19008992; PubMed Central PMCID: PMC2582510.

23. Berthon B, Spezi E, Galavis P, Shepherd T, Apte A, Hatt M, et al. Toward a standard for the evaluation of PET-Auto-Segmentation methods following the recommendations of AAPM task group No. 211: Requirements and implementation. Medical physics. 2017; 44(8):4098–111. https://doi.org/10.1002/mp.12312 PMID: 28474819; PubMed Central PMCID: PMC5575543.

24. Akamatsu G, Ikari Y, Nishida H, Nishio T, Ohnishi A, Maebatake A, et al. Influence of Statistical Fluctuation on Reproducibility and Accuracy of SUVmax and SUVpeak: A Phantom Study. J Nucl Med Technol. 2015; 43(3):222–6. https://doi.org/10.2967/jnmt.115.161745 PMID: 26271802.

25. Berthon B, Marshall C, Holmes R, Spezi E. A novel phantom technique for evaluating the performance of PET auto-segmentation methods in delineating heterogeneous and irregular lesions. EJNMMI Phys. 2015; 2(1):13. https://doi.org/10.1186/s40658-015-0116-1 PMID: 26501814; PubMed Central PMCID: PMC4538718.

26. Berthon B, Marshall C, Evans M, Spezi E. Evaluation of advanced automatic PET segmentation methods using nonspherical thin-wall inserts. Medical physics. 2014; 41(2):022502. https://doi.org/10.1118/1.4863480 PMID: 24506646.

27. Brockway K D., Nelson A. PET TUMOR SEGMENTATION: VALIDATION OF A GRADIENT-BASED METHOD USING A NSCLC PET PHANTOM2009.

28. Tan S, Li L, Choi W, Kang MK, D'Souza WD, Lu W. Adaptive region-growing with maximum curvature strategy for tumor segmentation in 18F-FDG PET. Phys Med Biol. 2017; 62(13):5383–402. https://doi.org/10.1088/1361-6560/aa6e20 PMID: 28604372; PubMed Central PMCID: PMC5497763.

29. Cui J, Yu H, Chen S, Chen Y, Liu H. Simultaneous estimation and segmentation from projection data in dynamic PET. Medical physics. 2019; 46(3):1245–59. https://doi.org/10.1002/mp.13364 PMID: 30593666.

30. Hatt M, Lee JA, Schmidtlein CR, Naqa IE, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. Medical physics. 2017; 44(6):e1–e42. https://doi.org/10.1002/mp.12124 PMID: 28120467.

31. Bollineni VR, Kramer GM, Jansma EP, Liu Y, Oyen WJ. A systematic review on [(18)F]FLT-PET uptake as a measure of treatment response in cancer patients. Eur J Cancer. 2016; 55:81–97. https://doi.org/10.1016/j.ejca.2015.11.018 PMID: 26820682.

32. Buck AK, Halter G, Schirrmeister H, Kotzerke J, Wurziger I, Glatting G, et al. Imaging proliferation in lung tumors with PET: 18F-FLT versus 18F-FDG. J Nucl Med. 2003; 44(9):1426–31. PMID: 12960187.

33. van Westreenen HL, Cobben DC, Jager PL, van Dullemen HM, Wesseling J, Elsinga PH, et al. Comparison of 18F-FLT PET and 18F-FDG PET in esophageal cancer. J Nucl Med. 2005; 46(3):400–4. PMID: 15750150.

34. Erdi YE, Mawlawi O, Larson SM, Imbriaco M, Yeung H, Finn R, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. Cancer. 1997; 80(12 Suppl):2505–9. https://doi.org/10.1002/(sici)1097-0142(19971215)80:12+<2505::aid-cncr24>3.3.co;2-b PMID: 9406703.

35. Hong R, Halama J, Bova D, Sethi A, Emami B. Correlation of PET standard uptake value and CT window-level thresholds for target delineation in CT-based radiation treatment planning. Int J Radiat Oncol Biol Phys. 2007; 67(3):720–6. https://doi.org/10.1016/j.ijrobp.2006.09.039 PMID: 17293230.

36. Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. Medical physics. 2010; 37(3):1309–24. https://doi.org/10.1118/1.3301610 PMID: 20384268.

37. Schaefer A, Kremp S, Hellwig D, Rube C, Kirsch CM, Nestle U. A contrast-oriented algorithm for FDG-PET-based delineation of tumour volumes for the radiotherapy of lung cancer: derivation from phantom measurements and validation in patient data. Eur J Nucl Med Mol Imaging. 2008; 35(11):1989–99. https://doi.org/10.1007/s00259-008-0875-1 PMID: 18661128.

38. Chan T, Vese L. An active contour model without edges. Lect Notes Comput Sc. 1999; 1682:141–51. WOS:000170515400013.

39. Kass M, Witkin A, Terzopoulos D. Snakes—Active Contour Models. Int J Comput Vision. 1987; 1 (4):321–31. WOS:A1987M205300003.

40. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. J Nucl Med. 2007; 48 (6):932–45. https://doi.org/10.2967/jnumed.106.035774 PMID: 17504879.

41. 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(4):663–72. https://doi.org/10.1007/s00259-010-1688-6 PMID: 21225425.

42. Hatt M, Rest CCl, Turzo A, Roux C, Visvikis D. A Fuzzy Locally Adaptive Bayesian Segmentation Approach for Volume Determination in PET. IEEE transactions on medical imaging. 2009; 28(6):881–93. https://doi.org/10.1109/TMI.2008.2012036 PMID: 19150782