PET functional volume delineation: a robustness and repeatability study.

Mathieu Hatt, Catherine Cheze Le Rest, Nidal Albarghach, Olivier Pradier, Dimitris Visvikis

PET functional volume delineation: a robustness and repeatability study

Mathieu Hatt[1], PhD, Catherine Cheze Le Rest[1,2], PhD, MD, Nidal Albarghach[1,3], MD, Olivier Pradier[1,3], PhD, MD, Dimitris Visvikis[1], PhD

1. INSERM, U650, LaTIM, CHU Morvan, Brest F-29200, France

2. Academic Department of Nuclear Medicine, CHU Brest F-29200, France

3. Institute of Oncology, CHU Brest F-29200, France

Short running title: Robustness of functional volume determination in PET

Keywords: PET uptake volume determination, robustness, repeatability, FLAB, thresholding

Conflict of interest: none

Corresponding author:

Mathieu HATT

LaTIM, INSERM U650

CHU MORVAN, 5 avenue Foch, 29609 Brest, France

e-mail: hatt@univ-brest.fr

Tel.:+33298018111

Fax: +33298018124

Word count: 5317

Abstract

Purpose: Current state of the art algorithms for functional uptake volume segmentation in PET imaging consist of threshold-based approaches, whose parameters often require specific optimization for a given scanner and associated reconstruction algorithms. Different advanced image segmentation approaches previously proposed and extensively validated, such as among others the fuzzy C-means (FCM) clustering, or the Fuzzy Locally Adaptive Bayesian (FLAB) have the potential to improve robustness of functional uptake volume measurements. The objective of this study was to investigate robustness and repeatability with respect to various scanner models, reconstruction algorithms and acquisition conditions.

Methods and materials: Robustness was evaluated using a series of IEC phantom acquisitions carried out on different PET/CT scanners (Philips Gemini and Gemini Time-of-Flight, Siemens Biograph and GE Discovery LS) with their associated reconstruction algorithms (RAMLA, TF MLEM, OSEM). A range of acquisition (contrast, duration) and reconstruction (voxel size) parameters were considered for each scanner model. On the other hand, the repeatability of each method was evaluated on simulated and clinical tumors and compared to manual delineation.

Results: For all the scanner models, acquisition parameters and reconstruction algorithms considered, FLAB demonstrated higher robustness in delineation of the spheres with low mean errors (10%) and variability (5%), with respect to threshold-based methodologies and FCM. The repeatability provided by all segmentation algorithms considered was very high with a negligible variability of <5% in comparison to that associated with manual delineation (5-35%).

Conclusion: The use of advanced image segmentation algorithms may allow not only high accuracy as previously demonstrated, but also provide a robust and repeatable tool to aid physicians as an initial guess in determining functional volumes in PET.

1. Introduction

Accurate, robust, reproducible and fast delineation of functional tumor uptake volumes in three dimensions using positron emission tomography (PET) has been identified as a pressing challenge for an increasing number of oncology applications, such as image-guided radiotherapy [1-3], diagnosis, prognosis and therapy response assessment [4,5]. On the one hand, manual delineation of functional uptake volumes using PET images is tedious and associated with very low repeatability due to high inter- and intra-observer variability [4], principally arising from the poor quality of PET images. On the other hand, current state-of-the-art algorithms for functional uptake volume segmentation using PET images consist of fixed [6] or adaptive thresholding approaches [7,8]. Regarding the use of fixed threshold, numerous studies have shown the need for variable threshold, depending on numerous factors, such as among them, lesion contrast, lesion size, and image noise [9]. As a solution, in the case of adaptive thresholding, the applied threshold depends on the measured contrast between the object to delineate and its surrounding background, as well as parameters requiring optimisation on phantom acquisitions. This optimisation has to be performed for each scanner model and associated reconstruction and correction algorithms, making these approaches system-dependent. In addition, recent studies show that even considering the same scanner model, a significant variation of the "ideal" threshold may exist due to differences in clinical acquisition and reconstruction protocols [10] underlying the possibility that such deterministic approaches may not be sufficiently robust and reproducible for functional uptake volume determination.

Recently several advanced image segmentation algorithms have been proposed in the literature for PET volume delineation [11-16]. The physical accuracy of these algorithms in differentiating the uptake signal from its surrounding

background has, in most cases, already been assessed with respect to ground-truth, provided by a combination of realistic simulated or acquired phantom images as well as, in some cases, clinical tumors with associated histopathology measurements.

However, apart from physical accuracy, different characteristics can be equally important in terms of assessing the performance of such advanced image segmentation algorithms, which in principle have the potential of being more robust and repeatable than “threshold-based” approaches. A robust and repeatable performance may facilitate their use with images acquired on different scanner models without any previous optimization to individual image quality, providing a less hardware dependent solution to the problem of 3D functional uptake segmentation. However none of these methodologies have been shown to be system independent, considering the potential variability that can be observed in PET image characteristics, depending on the scanner or associated reconstruction and correction algorithms used. Such an evaluation is essential for the efficient application of these approaches to the different clinical applications targeted, not simply within a given institution but also concerning their use within a multi-center trial context. Finally, such a robustness analysis could provide some insight regarding the potential behavior of a given segmentation algorithm considering the use of different tracers. On the one hand, the PET scanner properties in terms of spatial resolution will be similar for acquisitions performed with the same radioisotope, therefore resulting in similar magnitude partial volume effects. On the other hand, acquisitions performed using different radiotracers lead to different properties of uptake intensity and therefore subsequent different contrast and noise level characteristics for a given tumor uptake. For instance, 18F-FLT and 18F-FMISO images are usually characterized by higher noise levels and reduced tumor uptake contrast with respect

to what is usually observed in 18F-FDG images [17,18]. Therefore, by studying the behavior of automated algorithms dedicated to the delineation of elevated activity in 18F-FDG images, considering variable contrast and noise levels, one could gain an insight on the potential behavior of such algorithms when applied to other 18F-labeled PET tracers.

The objectives of this study were to (i). provide a robustness and repeatability evaluation framework, and (ii). assess within this framework the performance of different advanced and threshold-based segmentation algorithms in delineating elevated activity distributions in a PET image.

2. Materials and methods

2.1 Segmentation algorithms

Threshold-based and more advanced approaches were considered in this work. Two different fixed thresholds were considered, at 42% (T42) and 50% (T50) of the maximum tumor value, using a region growing algorithm with the maximum intensity voxel as seed [4]. An adaptive thresholding approach (TSBR) [7] was also included:

$$I_{threshold} = a + b\frac{1}{SBR} \quad (1)$$

SBR is the tumor-to-background ratio determined by ROI analysis, and the couple of parameters (a,b) is optimized for each scanner using phantom acquisitions of spheres.

In terms of more advanced image segmentation approaches, the Fuzzy C-Means (FCM) [16] clustering, previously used for functional volume segmentation

tasks in both brain and oncology applications [14,15,19,20], was considered. This algorithm iteratively estimates clusters' "centroids" (centers of mass) in the image, computing a voxel's membership between 0 and 1 to a given cluster depending on the distance between the voxel's value and the clusters' centroids. However, FCM lacks explicit noise and spatial correlation modeling. The second advanced algorithm considered was an unsupervised Bayesian segmentation, known as Fuzzy Locally Adaptive Bayesian (FLAB) [14,15]. It computes, for each voxel, a probability of belonging to a given "class" (for instance, tumor, background or a given uptake level within a tumor). This probability takes into account the voxel intensity, spatial correlation with surrounding voxels (the assumption being that voxels of similar intensities and close to each other have higher probability belonging to the same class) as well as the overall statistical distributions in the regions of the image by estimating the mean and variance for each class. FLAB automatically estimates the parameters of interest (number of classes, classes' mean and variance, spatial correlation of each voxel) within a Stochastic Expectation Maximization (SEM) framework [21]. In order to deal with the inherent blurry properties of PET images due to the limited spatial resolution of the scanners, the algorithm considers that each voxel may contain a mixture of classes by modeling both spatial correlation and statistical distributions with a combination of Dirac "hard" and Lebesgue "fuzzy" measures. This enables a classification of voxels as belonging to what we denote as "hard classes" or "fuzzy transitions", the first referring to fairly homogeneous regions, the second to blurred areas occurring at the frontier between two homogeneous regions. FLAB is therefore able to accurately differentiate if necessary both the overall tumor spatial extent from its surrounding background as well as tumor sub-volumes with different uptakes. The accuracy of FLAB has been previously

extensively investigated for both homogeneous [14] and heterogeneous non spherical tumors [15] and demonstrated satisfactory accuracy even for small (<2 cm in diameter) volumes of interest (both overall tumors or tumor sub-volumes), short acquisition durations (associated with higher noise levels) or low (<4:1) contrast (both for overall tumors with respect to their surrounding background or between a tumor and its smaller sub-volumes).

2.2 Accuracy, robustness, repeatability: definitions

For a given segmentation algorithm we define accuracy as the precision of retrieving the true 3D object spatial extent, shape and volume based on the reconstructed activity distribution in a PET image, irrespectively of the correlation between this distribution and the underlying physiological process. Thus an image segmentation algorithm is not expected to differentiate specific from non-specific tracer uptake (for example inflammation and tumor in the case of FDG) if they are of the same intensity. The defined accuracy of each of the methodologies considered, has been determined as in previous studies [14,15] by calculating the classification errors (CE, see section 2.4).

We define as robustness the ability of a given methodology to generate accurate segmented volumes under varying acquisition and image reconstruction conditions. This robustness is determined as the variability of the segmentation results when a method is applied without prior optimization on images acquired using various scanners, and for each scanner under various contrast and noise conditions, using different reconstruction and associated correction algorithms. A dataset consisting of multiple phantom acquisitions performed on various scanner models (see section 2.3) was used for this task. These phantom studies were used to assess

robustness as they are consistently employed for optimization purposes with most of the functional volume segmentation algorithms.

Within the context of this study, repeatability is defined as the ability of a given algorithm to reach the same result when applied multiple times on a single image. In such a task, deterministic fixed threshold approaches will always give the same result. On the other hand, more advanced methods are susceptible to give different results when applied multiple times on the same image. For example, the adaptive thresholding segmentation may depend on a manually drawn background ROI and may thus result in variable delineations depending on the choice of this ROI. Finally, manual delineation may be considered as the least repeatable, even when considering a single operator (intra-operator variability). In order to compare the performance of the different segmentation algorithms considered in terms of repeatability, we used a series of simulated tumor images [22], as well as fifteen different clinical cases (see section 2.3).

2.3 Validation studies

Four different PET/CT scanners currently used in clinical practice were considered for the robustness study; namely the Philips Gemini and Gemini TF (Philips Medical Systems, Cleveland, OH USA), the Siemens Biograph (SIEMENS Medical Solutions, Knoxville, USA) and the GE Discovery LS (GE Healthcare, Milwaukee, USA). In each case, acquisitions of the IEC phantom containing spheres of various diameters (10, 13, 17, 22, 28, 37 mm) filled with 18F and placed on a hot uniform background were carried out. A standard protocol was designed to generate the following acquisitions for each scanner model: (a). two different SBR (4:1 and 8:1), (b). three different acquisition durations (1, 2 and 5 min) to study the effect of

noise, and (c). two different voxel volumes used in the reconstruction (between 2x2x2mm^3 and 4.3x4.3x4.25mm^3). All acquisitions were performed in 3D mode and listmode format facilitating the generation of 1, 2 and 5 minutes realizations from one single five minutes acquisition. In addition to the standard CT acquisition used for attenuation correction, a CT scan at high resolution was acquired for each PET/CT acquisition in order to generate (after registration) a ground-truth defining the true spatial extent (the interior of the sphere) of the tracer uptake at the voxel-by-voxel level [14]. This is subsequently used to compute the accuracy of each algorithm through classification errors (see section 2.4).

Routine clinical image reconstruction protocols were used for all scanners. For the Philips GEMINI and GEMINI TF, data were reconstructed using the RAMLA 3D (2 iterations, relaxation parameter of 0.05 and a 5mm FWHM 3D Gaussian post-filtering) and the TF ML-EM algorithm respectively. In the case of the Siemens Biograph and GE Discovery LS, images were reconstructed with Fourier rebinning (FORE) followed by OSEM (4 iterations, 8 subsets (4i/8s) with a 5mm FWHM 3D Gaussian post-filtering and 2i/8s for the Biograph and Discovery systems respectively). All acquisitions were corrected for attenuation (using the corresponding CT image), as well as for scattered and random coincidences. Table I contains a summary of the parameters for each of the datasets obtained using the different scanners considered. Figures 1 and 2 illustrate the various images obtained. Note that in the case of the Philips GEMINI acquisitions, the 37mm sphere was not in the same plane as the others, thus appears visually smaller in the selected slice, while the 28mm sphere was missing in the phantom used for the GE Discovery LS acquisitions.

Regarding the repeatability study, two different datasets were used. The first one consisted of ten tumors extracted from a database of realistically simulated PET scans based on clinical whole body images using the NCAT (NURBS cardiac-torso) phantom, a model of the Philips GEMINI scanner and GATE (Geant4 Application for Tomography Emission). The procedure for the generation of these images, reconstructed using OPL-EM (7i/1s) with 4x4x4mm^3 voxels, has been previously described in detail [22]. In the second part of the repeatability study a number of clinical cases were selected from datasets acquired on various scanner models: 4 esophagus and 4 follicular lymphoma patients were acquired on the Philips GEMINI PET/CT scanner with 2min acquisition per bed position, 60min after FDG injection of 6MBq/kg. 3 Non-Small Cell Lung Cancer (NSCLC) were acquired on the Siemens Biograph (5min per bed position, 45min after 5MBq/kg of FDG injection) and the GE Discovery LS (3min per bed position, 60min after 5MBq/kg of FDG injection) respectively.

2.4 Analysis

For the phantom images used in the robustness study each sphere was processed separately. The images corresponding to the region containing each sphere were segmented in two classes (*sphere* and *background*), using each of the methods under evaluation (FCM, FLAB, T42 and T50). A voxel-to-voxel ground-truth based on the corresponding CT datasets as described previously [14], was used in the robustness evaluation of the different methodologies considered, through the determination of the segmentation accuracy with the computation of classification errors (CE):

$$CE = \frac{card\{t \mid c_t \neq x_t\}}{card\{t \mid x_t = 1\}} \times 100 \qquad (2)$$

where, c_t is the class assigned by the classification of voxel t, and x_t is its true class ($x_t = 1$ for the sphere and $x_t = 0$ for the background) and card{} is the cardinal. The errors are computed based on all misclassified voxels, either background voxels classified as the sphere or *vice versa,* divided by the total number of voxels defining the sphere volume.

Mean CE and associated standard deviation (SD) were obtained for each sphere and for each segmentation approach, thus providing a measure of the robustness of the different segmentation algorithms, when applied without specific optimization for a given scanner model or associated reconstruction algorithm under different imaging conditions (contrast and noise). The 10 mm sphere was not included in the analysis because it was not clearly visible in several of the phantom acquisitions and therefore not possible to segment particularly when using 4x4x4mm^3 and 5x5x5mm^3 reconstruction voxel size by any of the segmentation algorithms considered. Adaptive thresholding could not be compared directly with the other methodologies since it is optimized on each of the individual scanner datasets, with the parameters (a,b) optimized for each imaging device shown in table II. However, in order to assess the robustness of such approaches depending on the imaging system used we applied the adaptive thresholding using the parameters optimized on other scanners to the image datasets acquired with the Siemens Biograph.

For the repeatability evaluation, the simulated and clinical tumors were segmented ten times each with FCM, FLAB, and TSBR (fixed thresholding was not included since it always gives the same volume). In addition, manual delineation was carried out by two nuclear medicine experts with similar experience (more than 10 years) and training. More specifically the two experts were instructed to delineate the elevated uptakes in the images by performing ten different slice-by-slice manual

delineations for the different lesions considered in a randomised fashion, ensuring a minimum of a week between two consecutive segmentations of the same lesion. All these manual segmentations were carried out under the same conditions of full range contrast display. The mean percentage variability and associated standard deviation with respect to the mean segmented volume was computed for each of the lesions and segmentation approaches across the ten executions and across the ten manual delineations, in order to assess the repeatability of the approaches for each of the images. The repeatability of the manual delineations from the two experts were compared separately (intra-observer variability) and to each other (inter-observer variability).

3. Results

Classification errors representing segmentation accuracy, computed for each sphere are shown in figure 4(a), considering the entire range of systems used for acquisition and the different parameters in terms of contrast, acquisition duration and voxel size. For all the systems considered the relative impact of the different acquisition (contrast, duration) and reconstruction (voxel size) parameters is demonstrated in figures 4(b), 4(c) and 4(d) respectively. Table III contains the mean errors and standard deviations computed across the different spheres taken separately (as shown in figure 4(a)) and all together for the different imaging devices and acquisition configurations considered.

For the entire range of sphere sizes (37 to 13 mm), better accuracy and variability through smaller overall mean errors and SD can be seen for the FLAB algorithm (8.7±4.5%) relative to the other advanced segmentation algorithm

(27.8±25.6% for FCM) as well as relative to the fixed threshold approaches (20.3±18.5% and 42.6±51.6% for T50 and T42 respectively). These latter were also more sensitive to variations of the parameters as shown in figure 4(a). The results suggest that T50 is clearly more robust than T42 (SD of 19% compared to 52%). This is explained by the fact that the 50% threshold is more restrictive and hence leads to smaller over-estimation for the smallest spheres volumes, that the 42% threshold may grossly over-estimate (>100% errors for the most challenging imaging conditions). On the other hand, T50 leads to larger CE for the two larger spheres, as it tends to under-estimate their volumes by only including the central high intensity voxels of the sphere. Considering the FCM algorithm, the results demonstrate that it is unable to accurately segment spheres smaller than 2cm in diameter, leading to large overall mean errors when considering the performance over all of the sphere sizes, although it exhibits lower variability than the fixed threshold approaches for the majority of the spheres with a size >2cm. As Figure 4(b) demonstrates, whereas FLAB exhibits small variability with respect to contrast changes, all other methodologies, especially T42 and FCM exhibit higher sensitivity to such changes. T50 on the other hand, is less sensitive to contrast changes with respect to the mean error but exhibits larger variability for lower contrast. Figure 4(c) illustrates the resilience to shorter acquisition (hence higher noise levels) for each methodology. FLAB demonstrates very low variability with shorter acquisitions, whereas all other methodologies show higher variability with significantly larger mean errors and standard deviations. Finally, only small improvements were found for each methodology (except for T50) when using smaller voxels (see figure 4(d)).

The optimized (a,b) parameters of the TSBR for each scanner model are shown in table II. The mean classification error across the 13-37mm spheres range,

associated to each scanner was between 9.7 and 13.1% with associated standard deviations from 2.8 to 5.2%. When applying the (a,b) parameters of the Philips GEMINI, Philips GEMINI TF and Discovery LS datasets to the Siemens Biograph dataset, this mean error rose from 13.1±5.2% to 21.7±7.1%, 23.4±7.6% and 19.1±6.4% respectively.

Concerning the repeatability results, table IV contains the mean variability and SD around the mean segmented volume across the ten manual delineations performed from each of the two nuclear medicine experts, and 10 repeated executions of the FLAB, FCM and TSBR algorithms. FLAB demonstrated highly repeatable results in all of the studied cases, with negligible variability (<1%) around the mean segmented 3D volume across the different repeated executions for both the simulated and the clinical datasets. FCM also lead to satisfactory repeatability results (0.8±0.6% on simulated tumors and 1.7±1.9% on clinical cases). In comparison, the use of the TSBR led to more than twice as high variability (3.4±2.8% and 3.8±3.1% for the simulated tumors and clinical cases respectively) which is most certainly due to the background ROI manual definition. By contrast manual segmentation performed by the two experts showed high intra-observer variability on simulated tumors (13.4±17.3% and 11.7±18.4% for expert 1 and 2 respectively), and even larger variability on clinical images (19.6±15.2% and 22.1±13.6% for expert 1 and 2 respectively). Inter-observer variability (variability of observer 2 with respect to mean volume of observer 1) was 16.4±21.8% and 24.7±17.6% for the simulated tumors and clinical cases respectively. Figure 4 illustrates one example of some of the delineations obtained by the manual segmentation and automatic approaches.

4. Discussion

Functional tumor uptake volumes delineation represents today an area of interest for multiple clinical (routine and research) applications of PET imaging, such as response to therapy studies and radiotherapy treatment planning. In all of these applications, the robustness and repeatability with which functional uptake volumes can be determined under different imaging conditions plays a predominant role, allowing a level of confidence to be established in the use of such tumor volume measurements in clinical practice [18]. Although several promising advanced algorithms have been recently proposed [11-15,20], methodologies currently used in clinical practice are based on the use of manual delineation or fixed and adaptive thresholding [6-8]. The major drawback of manual delineation is its high inter- and intra-observer variability in addition to being time consuming. On the other hand, the currently considered state of the art adaptive threshold based algorithms have been shown to accurately define functional volumes under certain imaging conditions of spherical and homogeneous activity distribution lesions. However, they require specific parameters optimization and are thus system-dependent. In addition, the adaptive thresholding approaches usually involve some user interaction to select background regions of interest, which can potentially lead to user introduced variability. In the present study we have focused on the evaluation under different imaging conditions of the level of robustness and repeatability of different functional volume segmentation algorithms, including current state-of-the-art in clinical practice.

In terms of robustness, the use of images from different commercial PET/CT systems acquired under typical clinical acquisition conditions resulted in large variability in the performance of the different segmentation algorithms evaluated. Across all of the images and spheres considered, a fixed threshold of 42% of the

maximum resulted in the largest variability of the segmented functional volumes (±15-60%) across the different images considered for spheres <3cm in diameter. On the other hand, a fixed threshold of 50% was closer in terms of variability (±20%) with that of one of the advanced segmentation algorithms included in this work (FCM). Finally, the FLAB algorithm was the most robust of all evaluated algorithms leading to the smallest variability (±5%), with no particular dependence on acquisition (duration, contrast) and processing parameters (reconstructed voxel size). The 42% fixed threshold and the FCM algorithm were the most sensitive to contrast and the acquisition duration, across the different scanners used. In terms of variability across the different images used, the 50% fixed threshold demonstrated the most significant variability dependence on lesion contrast. Finally, regarding the use of adaptive thresholding (TSBR), applying this approach to acquisitions performed on a different scanner than the one used to optimize its parameters led to higher mean errors of <25%.

In terms of repeatability, all algorithms considered exhibited mean differences of <5%, although only FLAB came close to the perfect repeatability that can be achieved by a deterministic approach such as a fixed threshold. Finally, the repeatability of both threshold and automatic segmentation approaches was superior to that of manual delineation (variability >15-20% for both the clinical and simulated tumors).

The overall better accuracy (lower mean errors) and smaller variability (lower standard deviation) associated with the FLAB algorithm across the different images considered demonstrates its ability, without the need of any scanner-specific optimization, to robustly deal with the different image quality resulting from the use of different reconstruction and correction algorithms as well as sensitivities associated

with different systems. This of course should be considered within the context of the limited absolute accuracy of binary threshold-based approaches shown in this and previous studies. The accuracy of threshold-based approaches is particularly limited for non-homogeneous in form and activity distribution lesions resulting, as previously shown [15], in large under or over-estimation of the overall tumor spatial extent.

The present study also demonstrated that the use of any of the segmentation algorithms significantly reduces intra- and inter-observer variability associated with manual delineation. However, one should keep in mind that automated segmentation algorithms are not able to differentiate between similar levels of physiological and pathological elevated tracer uptakes. Therefore physician involvement is still imperative and desirable, especially regarding the detection and selection of elevated tracer uptakes corresponding to pathological findings that are to be subsequently accurately delineated.

5. Conclusion

This study has demonstrated significant differences in the robustness and reproducibility of functional volume measurements depending on the segmentation algorithm used. The advantage of employing advanced segmentation algorithms is an improvement in overall elevated activity delineation across the different range of image quality that can be encountered today in clinical practice, without the need for system-dependent optimization procedures. In addition, their high level of repeatability allows achieving similar performance to that of deterministic threshold based approaches. Therefore such advanced image segmentation algorithms may provide robust and reliable tools to aid physicians as an initial guess in determining functional volumes in PET.

Acknowledgments

This work was financially supported by the French National Research Agency (ANR) under contract ANR-08-ETEC-005-01. We would like to thank the following clinical centers and associated members for some of the phantom and patient datasets used in this study: Nuclear Medicine departments of CHU Brest, France (Alexandre Turzo), CHU Sud-Amiens, France, (Pascal Bailly, Joel Daouk), and St Bartholomew's Hospital, London, UK (Iain Murray).

References

(1) Lucignani G. SUV and segmentation: pressing challenges in tumor assessment and treatment. *Eur J Nucl Med Mol Im* 2009;36:715-720.

(2) Jarritt H, Carson K, Hounsel AR, Visvikis D. The role of PET/CT scanning in radiotherapy planning. *British Journal of Radiology* 2006;79**(**S):27-35.

(3) Pan T, Mawlawi O. PET/CT in radiation oncology. *Med Phys* 2008;35(11):4955-4966.

(4) Krak NC, Boellaard R, et al. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Im* 2005;32:294-301.

(5) Jerusalem G, Hustinx R, Beguin Y, Fillet G. The value of positron emission tomography (PET) imaging in disease staging and therapy assessment. *Ann Oncol* 2002;13(S4):227-234.

(6) Erdi E, Mawlawi O, Larson SM, et al. Segmentation of Lung Lesion Volume by Adaptative Positron Emission Tomography Image Thresholding. Cancer 1997;80(S12):2505-2509.

(7) Daisne JF, Sibomana M, Bol A, et al. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radioth Oncol* 2003;69:247-250.

(8) 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(8):1342-8.

(9) Biehl KJ, Kong MF, Dehdashti F, Jin JY, Mutic S, El Naqa I, 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;47:1808-1812.

(10) Oellers M, Bosmans G, Van Baardwijk A, et al. The integration of PET-CT scans from different hospitals into radiotherapy treatment planning *Radiother Oncol* 2008;87(1):142-146.

(11) El Naqa I, Yang D, Apte A, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. *Med Phys* 2007;34(12):4738-4749.

(12) Montgomery DWG, Amira A, and Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Medical Physics* 2007;34(2):722-736.

(13) Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur J Nucl Med Mol Im* 2007;34:1427-1438.

(14) Hatt M, Turzo A, Roux C, et al. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. *IEEE Trans Med Im* 2009;28(6):881-893.

(15) 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(1):301-308.

(16) Dunn JC. A fuzzy relative of the isodata process and its use in detecting compact well-separeted clusters. *J Cybernet* 1974;31:32-57.

(17) Koh WJ, Rasey JS, Evans ML, et al. Imaging of hypoxia in human tumors with [F-18]fluoromisonidazole. *Int J Radiat Oncol Biol Phys* 1992;22(1):199-212.

(18) Hatt M, Cheze Le Rest C, Aboagye EO, et al. Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements. J Nucl Med. 2010;51(9):1368-1376.

(19) Zhu W, Jiang T. Automation segmentation of PET image for brain tumors *in IEEE NSS-MIC Conf Rec* 2003;4:2627–2629.

(20) Belhassen S and Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. Med Phys 2010;37(3):1309-1324.

(21) Celeux G, Diebolt J. L'algorithme SEM : Un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités. *Revue de Statistique Appliquée* 1986;34(2).

(22) Le Maitre A, Segars WP, Marache S, et al. Incorporating patient specific variability in the simulation of realistic whole body 18F-FDG distributions for oncology applications. *Proceedings of the IEEE special issue on computational anthropomorphic anatomical models* 2009;97(12):2026-2038.

PET/CT scanner models and acquisition parameters				
PET/CT system	**Contrast**	**Voxel size**	**Duration (min)**	**Recon**
Philips Gemini	4:1 8:1	2 x 2 x 2 4 x 4 x 4	1, 2, 5	RAMLA 3D
Philips Gemini TF	4:1 8:1	2 x 2 x 2 4 x 4 x 4	1, 2, 5	TF ML-EM
Siemens Biograph	4:1 8:1	2 x 2 x 2 5.33 x 5.33 x 2	1, 2, 5	FORE-OSEM
GE Discovery LS	4:1 8:1	1.95 x 1.95 x 4.25 4.3 x 4.3 x 4.25	1, 2, 5	FORE-OSEM

Table I

Adaptive thresholding parameters for each scanner				
PET/CT system	**TSBR a param**	**TSBR b param**	**Minimum mean associated classification error (%)**	**Standard deviation of classification error**
Philips Gemini	40.1	59.7	10.8	3.3
Philips Gemini TF	38.6	61.4	9.7	2.8
Siemens Biograph	41.7	57.6	13.1	5.2
GE Discovery LS	42.0	56.8	11.1	3.7

Table II

Robustness results obtained across the entire range of scanner models and acquisition parameters for each method			T42	T50	FCM	FLAB
All spheres (37-13 mm)		Mean CE (%)	42.6	20.3	27.8	8.7
		Standard dev.	51.6	18.5	25.6	4.5
Figure 4(a)	37 mm	Mean CE (%)	10.5	16	11.4	8.4
		Standard dev.	5.3	7.9	5.3	2.8
	28 mm	Mean CE (%)	17	15.9	11.7	8.4
		Standard dev.	13.8	7.5	5.7	3.6
	22 mm	Mean CE (%)	23	15.6	13.4	7.9
		Standard dev.	20.7	9.8	7.1	3.3
	17 mm	Mean CE (%)	49.1	21.5	31.6	7.2
		Standard dev.	35	13.8	12.7	4.9
	13 mm	Mean CE (%)	113.6	32.7	70.9	11.6
		Standard dev.	62.1	33.1	20.9	5.9

Table III

Repeatability results for each methodology and manual observers				
Method	**Simulated cases**		**Clinical cases**	
	Mean variability (%)	**Standard deviation**	**Mean variability (%)**	**Standard deviation**
FLAB	0.5	0.3	0.9	0.5
FCM	0.8	0.6	1.7	1.9
Fixed thresholding	0	0	0	0
Adaptive thresholding	3.4	2.8	3.8	3.1
Manual delineation (expert 1)	13.4	17.3	19.6	15.2
Manual delineation (expert 2)	11.7	18.4	22.1	13.6
Manual delineation (expert 2 w/r to 1)	16.4	21.8	24.7	17.5

Table IV

Table captions

Table I: Overview of all the parameters considered for each scanner model.

Table II: Optimized parameters a and b of the adaptive thresholding for each scanner model, with the minimum mean classification errors and their associated standard deviations across the entire range of configurations.

Table III: Robustness evaluation: mean classification error and associated standard deviation computed for each methodology across the entire range of sphere phantom acquisitions

Table IV: Repeatability evaluation: variability and standard deviation around the mean segmented volume for repeated (10 times) delineations of simulated and clinical tumors.

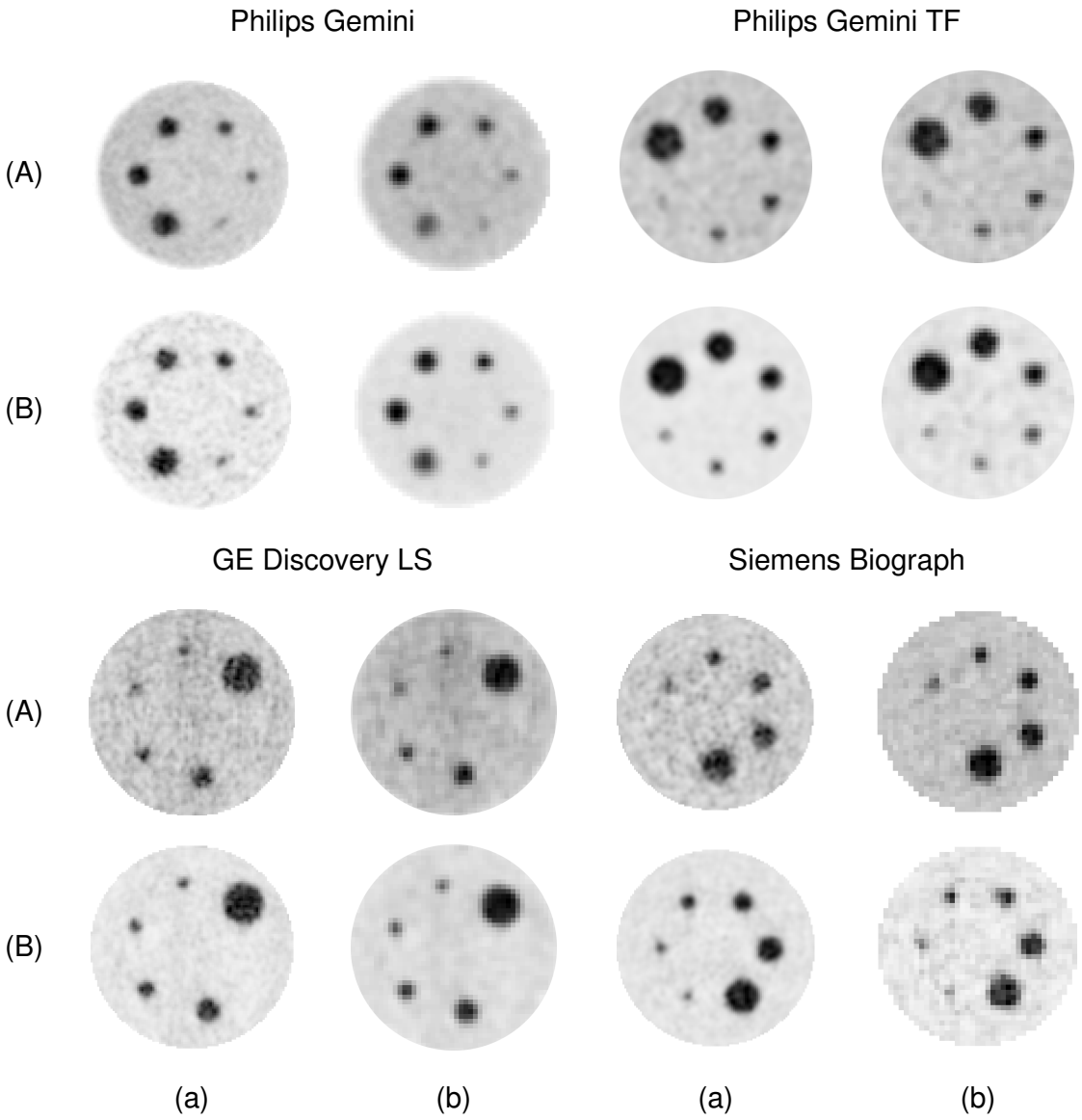
Philips Gemini
Philips Gemini TF
(A)
(B)
GE Discovery LS
Siemens Biograph
(A)
(B)
(a)
(b)
(a)
(b)

Figure 1

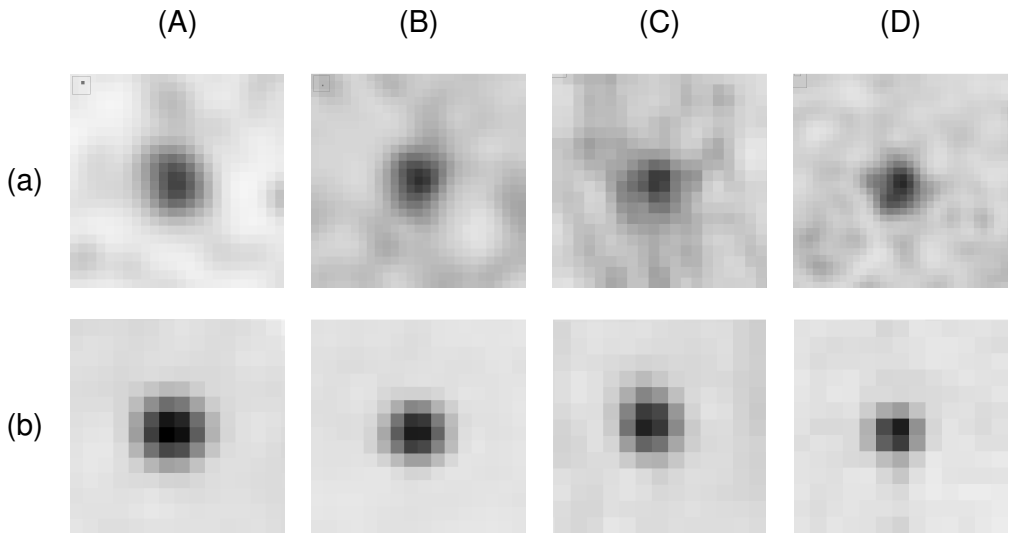

Figure 2

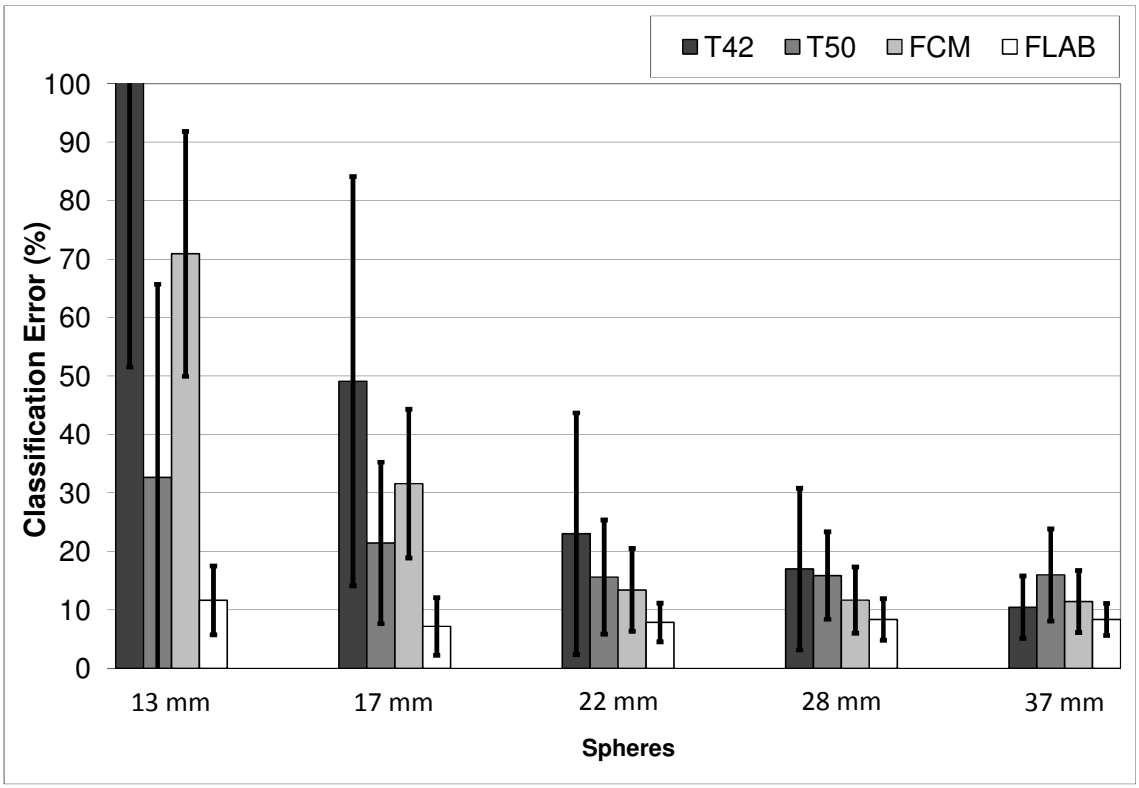

Figure 3(a)

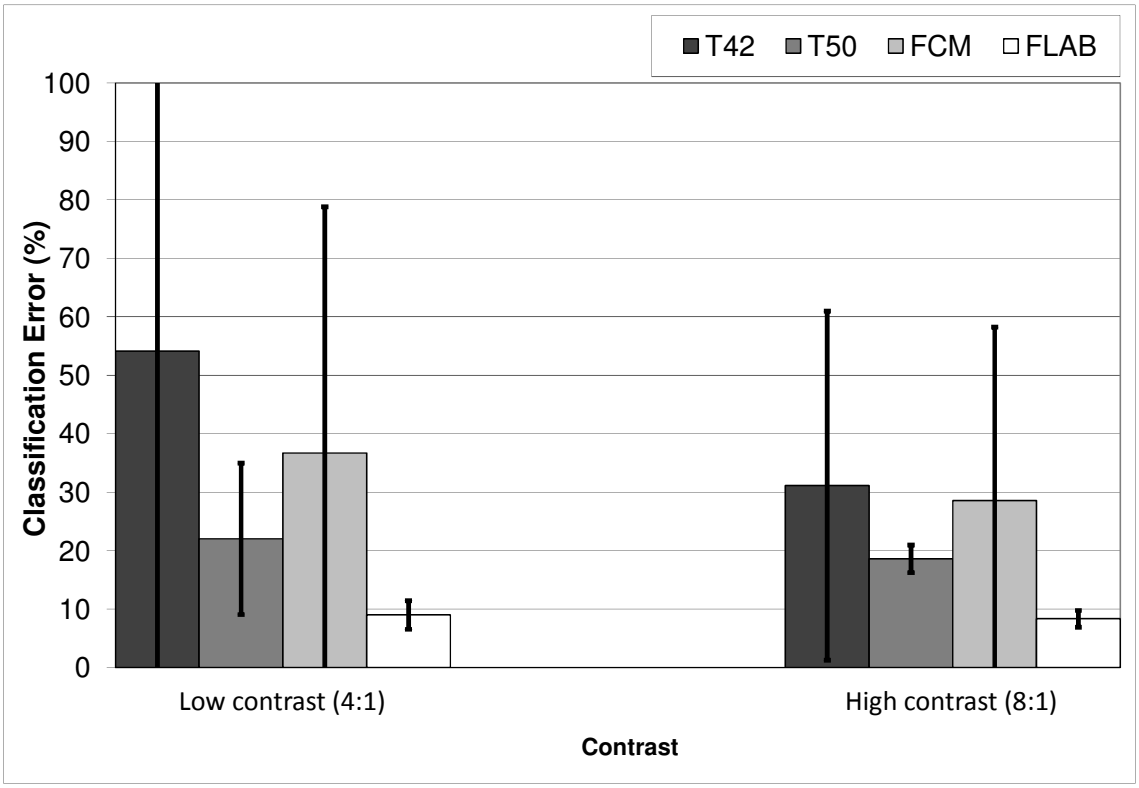

Figure 3(b)

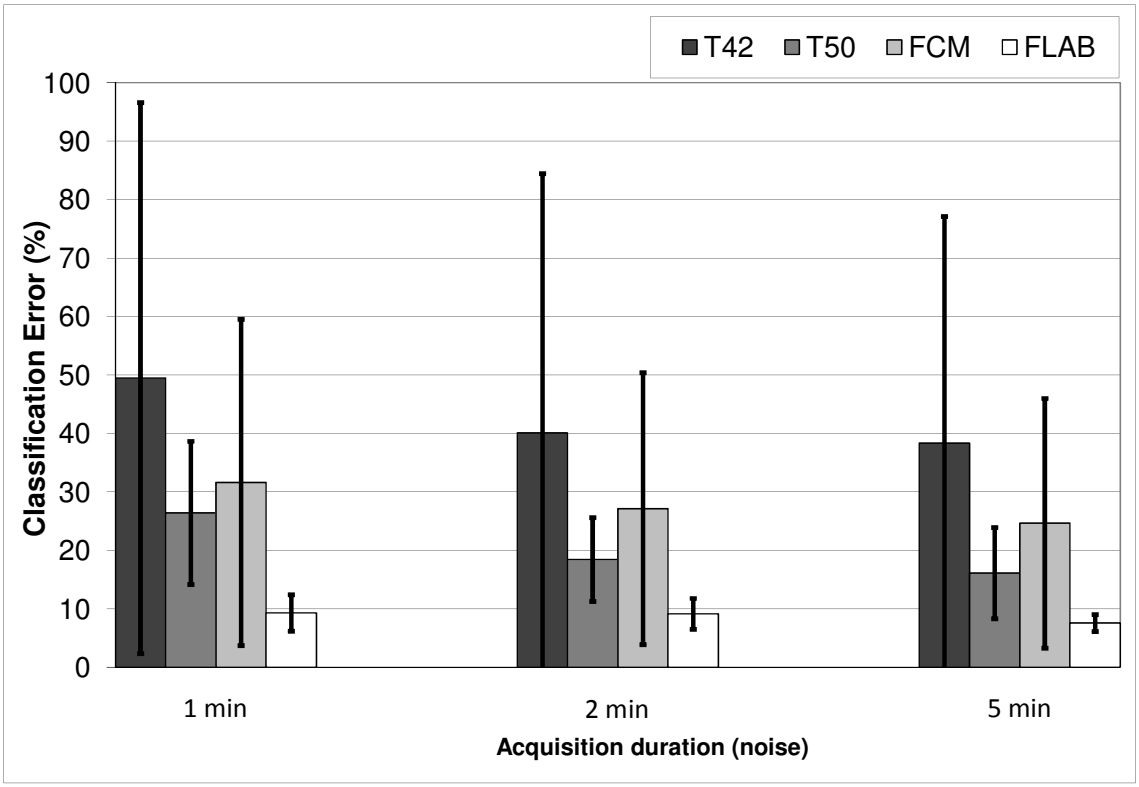

Figure 3(c)

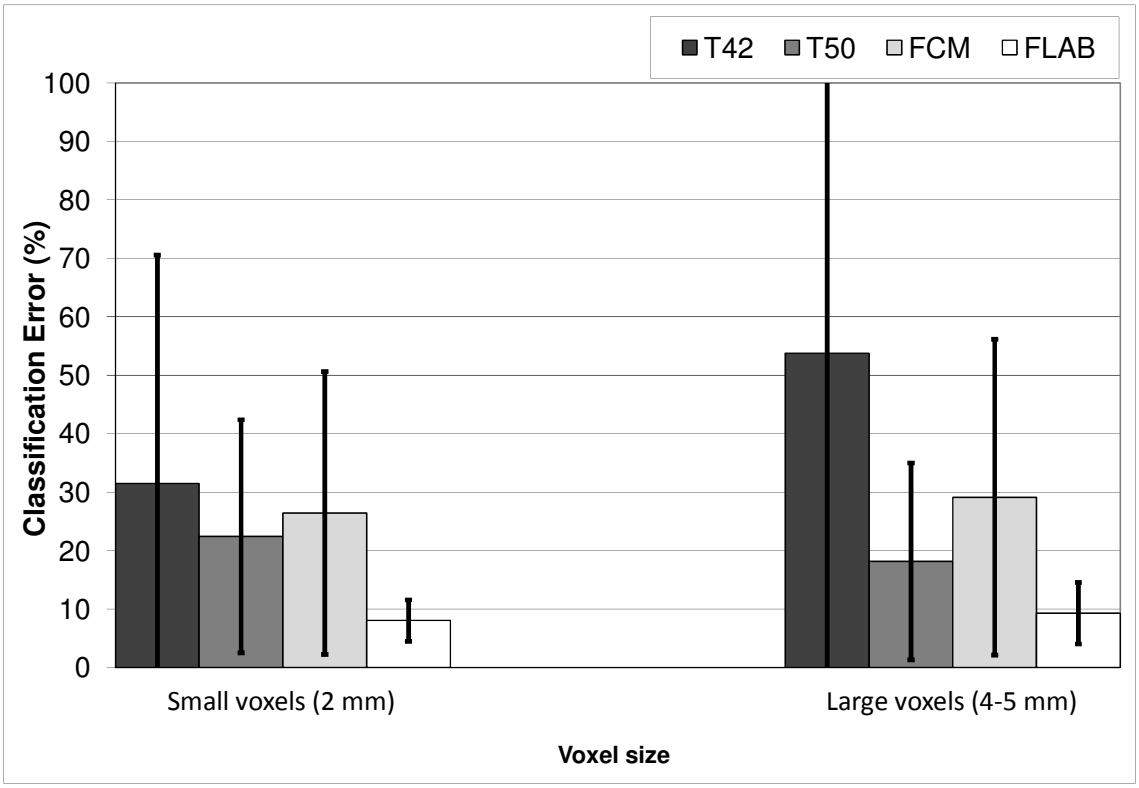

Figure 3(d)

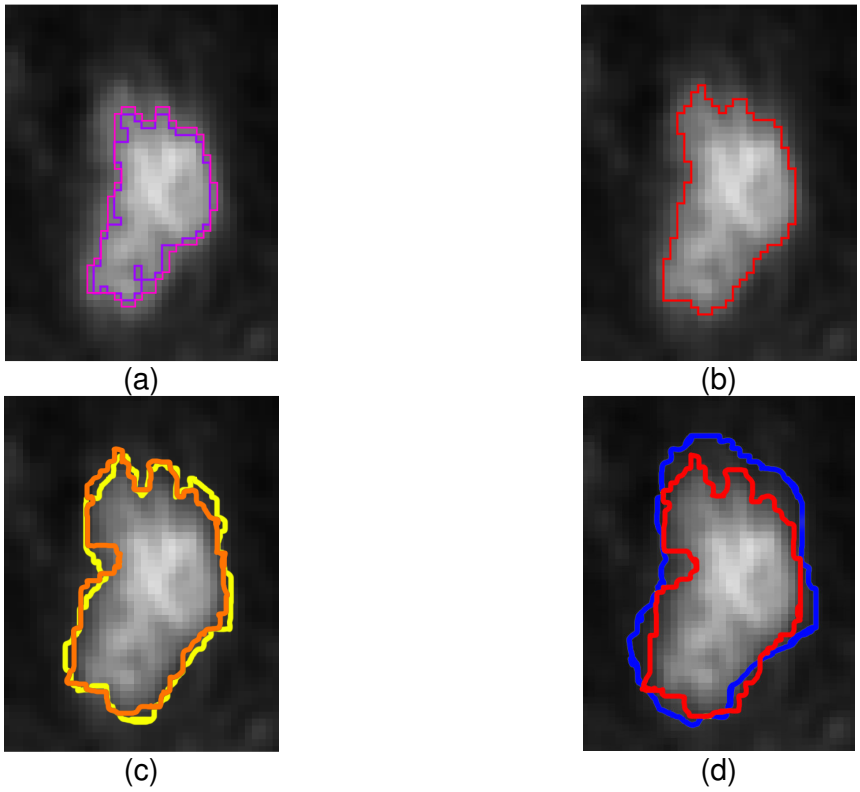

Figure 4

Figure Captions

Figure 1: 2D phantom slices, through the centre of the spheres, for the different systems and imaging conditions. Contrast ratios: rows (A) 4:1 and (B) 8:1. Voxel sizes: columns (a) small voxels and (b) large voxels (see table I).

Figure 2: Illustration of the variability considering the 17mm sphere across all four scanner models for two different opposing configurations: (a): 4:1 contrast, small voxels and 1min acquisition. (b): 8:1 contrast, large voxels and 5min acquisition. (A) Philips Gemini, (B) Philips Gemini TF, (C) Siemens Biograph, and (D) GE Discovery LS.

Figure 3: Mean classification errors and standard deviation (error bars) for each methodology with respect to (a) sphere diameter, (b) contrast, (c) acquisition duration, and (d) voxel size, computed across the different scanner models.

Figure 4: Illustration on one image slice of tumor delineations obtained using: (a) adaptive thresholding with two different background ROIs (6% difference), (b) FLAB delineation, (c) two fairly consistent manual delineations (9% difference) from the same observer and (d) two highly different (37% difference) manual delineations from two different observers.