Received Date: 07-Jun-2016

Revised Date: 09-Dec-2016

Accepted Date: 04-Jan-2017

Article Type: Task Group Report

Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM Task Group No. 211

Mathieu Hatt[1], John Lee[2], Charles R. Schmidtlein[3], Issam El Naqa[4], Curtis Caldwell[5], Elisabetta De Bernardi[6], Wei Lu[3], Shiva Das[7], Xavier Geets[2], Vincent Gregoire[2], Robert Jeraj[8], Michael P. MacManus[9], Osama R. Mawlawi[10], Ursula Nestle[11], Andrei B. Pugachev[12], Heiko Schöder[3], Tony Shepherd[13], Emiliano Spezi[14], Dimitris Visvikis[1],Habib Zaidi[15], Assen S. Kirov[3*]

[1] INSERM, UMR 1101, LaTIM, University of Brest, IBSAM, Brest, France

[2] Université catholique de Louvain (IREC/MIRO) & FNRS, 1200 Brussels, Belgium

[3] Memorial Sloan-Kettering Cancer Center, New York, NY 10065

[4] University of Michigan, Ann Arbor, MI, 48103

[5] Sunnybrook Health Sciences Center, Toronto, ON M4N 3M5, Canada

[6] University of Milano-Bicocca, Monza, Italy

[7] University of North Carolina, Chapel Hill, NC 27599

[8] University of Wisconsin, Madison, WI 53705

[9] Peter MacCallum Cancer Centre, Melbourne, Australia

[10] MD Anderson Cancer Center, Houston, TX 77030

[11] Universitätsklinikum Freiburg, 79106 Freiburg, Germany

[12] University of Texas Southwestern Medical Center, Dallas, TX 75390

[13]Turku University Hospital, 20521 Turku, Finland

[14] School of Engineering, Cardiff University, Cardiff, Wales, United Kingdom

[15] Geneva University Hospital, Geneva, CH-1211, Switzerland

---

*Corresponding author
Department of Medical Physics,
Memorial Sloan-Kettering Cancer Center
1275 York Avenue
New York, NY 10065
Tel.: (212) 639-7126
Fax: (212) 717-3010
E-mail: kirova@mskcc.org

Abstract

Purpose: The purpose of this educational report is to provide an overview of the present state-of-the-art PET auto-segmentation (PET-AS) algorithms and their respective validation, with an emphasis on providing the user with help in understanding the challenges and pitfalls associated with selecting and implementing a PET-AS algorithm for a particular application.

Approach: A brief description of the different types of PET-AS algorithms is provided using a classification based on method complexity and type. The advantages and the limitations of the current PET-AS algorithms are highlighted based on current publications and existing comparison studies. A review of the available image datasets and contour evaluation metrics in terms of their applicability for establishing a standardized evaluation of PET-AS algorithms is provided. The performance requirements for the algorithms and their dependence on the application, the radiotracer used and the evaluation criteria are described and discussed. Finally, a procedure for algorithm acceptance and implementation, as well as the complementary role of manual and auto-segmentation are addressed.

Findings: A large number of PET-AS algorithms have been developed within the last 20 years. Many of the proposed algorithms are based on either fixed or adaptively selected thresholds. More recently, numerous papers have proposed the use of more advanced image analysis paradigms to perform semi-automated delineation of the PET images. However, the level of algorithm validation is variable and for most published algorithms is either insufficient or inconsistent which prevents recommending a single algorithm. This is compounded by the fact that realistic image configurations with low signal-to-noise ratios (SNR) and heterogeneous tracer distributions have rarely been used. Large variations in the evaluation methods used in the literature point to the need for a standardized evaluation protocol.

Conclusions: Available comparison studies suggest that PET-AS algorithms relying on advanced image paradigms provide generally more accurate segmentation than approaches based on PET activity thresholds, particularly for realistic configurations. However, this may not be the case for simple shape lesions in situations with a narrower range of parameters, where simpler methods may also perform well. Recent algorithms which employ some type of consensus or automatic selection between several PET-AS methods have potential to overcome the limitations of the individual methods when appropriately trained. In either case, accuracy evaluation is required for each different PET scanner and scanning and image reconstruction protocol. For the simpler, less robust approaches, adaptation to scanning conditions, tumor type and tumor location by optimization of parameters is necessary. The results from the method evaluation stage can be used to estimate the contouring uncertainty. All PET-AS contours should be critically verified by a physician. A standard test, i.e., a benchmark dedicated to evaluating both existing and future PET-AS algorithms needs to be designed, in order to aid clinicians in evaluating and selecting PET-AS algorithms and to establish performance limits for their acceptance for clinical use. The initial steps towards designing and building such a standard are undertaken by the task group members.

TABLE OF CONTENTS

I. INTRODUCTION

Positron emission tomography (PET) has the potential to improve the outcome of cancer therapy because it allows the identification and characterization of tumors to be conducted based on their metabolic properties[1], which are inherently tied to cancer biology. PET is helpful in delineating the tumor target for radiation therapy, in quantifying tumor burden for therapy assessment, in determining patient prognosis and in detecting and quantifying recurrent or metastatic disease. This is especially true when the cancer lesion boundaries are not easily distinguished from surrounding normal tissue in anatomical images. Combined PET/CT (computed tomography) provides both anatomical/morphological and functional information in one imaging session. In addition to segmentation, this allows for the division of the tumors into subregions based on metabolic activity, which could potentially be used to treat/evaluate these subregions differentially (e.g., by increasing the dose to the more aggressive and radioresistant sub-volumes, an approach known as "dose painting"[2]). Accurate delineation of the metabolic tumor volume in PET is important for predicting and monitoring response to therapy. Apart from standardized uptake value (SUV) measurements, [3,4] other parameters (e.g., total lesion glycolysis (TLG) or textural and shape features, as well as tracer kinetic parameters) with complementary/additional predictive/prognostic value can be extracted from PET images.

For radiation therapy, leaving parts of the tumor untreated, because its extent is underestimated by anatomic imaging, or conversely irradiating healthy tissue because boundaries between the tumor and the adjacent normal tissue cannot be defined, can result in suboptimal response and/or (possibly severe) adverse side-effects. It has been shown in several clinical studies that PET, using the [18F]2-fluoro-2-deoxy-D-glucose (Fluorodeoxyglucose) radiotracer (^{18}F-FDG PET), has led to changes in clinical management for about 30% of patients[5-7]. Other studies, involving non-small-cell lung cancer (NSCLC)[8, 9, 10] and head and neck (H&N) cancer[11] have demonstrated that the incorporation of PET imaging in radiotherapy planning can result in significant changes (either increase or decrease) in treatment volumes.

In addition, the quantitative assessment of the metabolically active tumor volume, may provide independent prognostic or predictive information. This has been shown in several malignancies, including locally advanced esophageal cancer[12], non-Hodgkin lymphoma[13], pleural mesothelioma[14], cervical and H&N cancers[15] and lung cancer[16].

These promising data impose the need to establish and validate algorithms for the segmentation of PET metabolic volumes before and during treatment. The gross tumor volumes (GTV) defined by PET are intended to contain the macroscopic extend of the tumors. Currently, inaccuracies in defining PET-based GTV arise from variations in the biological processes determining the radiotracer uptake, as well as from physical and image acquisition phenomena which affect the reconstructed PET images[4, 17-22]. Furthermore, uncertainty can be introduced by the segmentation process itself. It has been shown that volume differences of up to 200% can arise from using different GTV contouring algorithms[23].

Regardless of these uncertainties, many radiation oncology departments have started using PET/CT for lesion delineation in radiation treatment planning (RTP)[1, 7-9, 11,24]. Numerical auto-segmentation techniques can be used for guidance in the delineation process, have

been shown to reduce intra- and inter-observer variations[25, 26] and some commercial vendors are now offering tools for semi-automatic delineation of tumor volumes in PET images for radiotherapy planning or response assessment. While these approaches may work reasonably well when applied in conjunction with anatomical imaging and clinical expertise, their accuracy and limitations have not been fully assessed.

Due to the complexity of the problem of PET based tumor segmentation and due to the abundance of potentially applicable numerical approaches, a large variety of automatic, semiautomatic and combined PET-AS approaches have been proposed over the past 20 years[27-30] Multiple semi-automatic approaches derived from phantom data as well as fully automated algorithms differing in terms of the algorithmic basis, fundamental assumptions, clinical goals, workflow, and accuracy have been proposed. In addition, algorithms for segmenting combinations of images from PET and other imaging modalities have appeared in literature[31-34]. The majority of these approaches have been tested on either simplistic phantom studies or patient datasets where the ground truth is largely unknown. Finally, only a few of these algorithms have been tested for their ability to segment lesions with irregular shapes or non-uniform activity distributions, which are essential for the implementation of accurate delineation protocols. In addition, most methods have been evaluated using different datasets and protocols, which makes comparing the results difficult, even impossible. As a result, in essence, there is currently no commonly adopted technique for reliable, routine, clinical PET image auto-segmentation.

In this educational report we provide a description with examples of the main classes of PET-AS algorithms (section II), highlight the advantages and the limitations of the current techniques (section III) and discuss possible evaluation approaches (section IV). In that section the types of available image datasets and the existing approaches for contour evaluation are discussed with the intention of laying out a basis for a standard for effective evaluation of PET auto-segmentation algorithms. The clinician interested in the practical aspects of PET segmentation may find most useful section V, which highlights the biological, physiological and image acquisition factors affecting the performance of the PET-AS methods, as well as preliminary guidelines for their acceptance and implementation.

II. DESCRIPTION AND CLASSIFICATION OF THE ALGORITHMS

The following is a glossary of the abbreviation, definitions and notations used in this report:

Abbreviations:

ARG – Adaptive Region Growing
ATS – Adaptive Threshold Segmentation
BTV – Biological Target Volume
CT – Computed Tomography
CTV – Clinical Target Volume
DSC - Dice Similarity Coefficient
DWT- discrete wavelet transforms
FTS – Fixed Threshold Segmentation
EM – Expectation Maximization
FCM - Fuzzy *C*-Means
FDG – [18F]2-fluoro-2-deoxy-D-glucose (Fluorodeoxyglucose)
FLT – ^{18}F-3′-fluoro-3′-deoxy- L-thymidine
F-MISO – ^{18}F-fluoromisonidazole
FOM – Figure of Merit
GTV – Gross Tumor Volume
H&N – Head and Neck
ML - Maximum Likelihood
MRI – Magnetic Resonance Imaging
MVLS – Multi Valued Level Sets
NEMA - National Electrical Manufacturers Association
NGTDM - neighborhood gray-tone difference matrices
NSCLC – Non-Small Cell Lung Cancer
PET – Positron Emission Tomography
PET-AS – PET Auto - Segmentation
PPV - Positive Predictive Value
PSF – Point Spread Function
PTV - Planning Target Volume
PVE – Partial Volume Effect
ROC - Receiver Operating Characteristic
SNR – Signal to Noise Ratio

Definitions:

ROI (Region of Interest): A 2D or 3D region drawn on an image for purposes of restricting and focusing analysis to its contents. It is closely related to and often used interchangeably with volume of interest (VOI).

SUV (Standardized Uptake Value): A measure of the intensity of radiotracer uptake in an object (lesion or body region) or region of interest; measured activity in that region is normalized to the injected activity and some measurement of patient size, most commonly weight (mass).

TLG (Total Lesion Glycolysis): The integral of the FDG-SUV over the volume, also equal to the product of the mean FDG-SUV and the volume. The same paradigm can be applied to

other radiotracers and is called for instance Total Proliferative Volume, or Total Hypoxic Volume in the case of FLT or FMISO, respectively.

VOI (Volume of Interest): A 3D region defined in a set of images for purposes of restricting and focusing analysis to its contents. It is closely related to and often used interchangeably with region of interest (ROI).

Notation:

I: The image set.

I_{VOI}: The VOI in image set I that the segmented region is taken from.

$\hat{\Omega}$: The segmented region from image set I.

I_i: The intensity of the i^{th} element (image pixel) from image set I. This element is commonly normalized with respect to activity and weight to SUV but can be further normalized.

ξ_i: Normalized uptake for the i^{th} voxel (see Appendix II).

T(Threshold): Commonly used in threshold segmentation. It defines the value at which a voxel is segregated between one set and another.

T^* (Threshold): The estimated segmentation threshold.

$V(T)$: Volume as a function of threshold.

V_{known}: The known volume of a segmented object.

x_i: The position of the i^{th} voxel.

c_k^n: The cluster center of the k^{th} cluster at the n^{th} iteration.

u_{ik}^n: The membership probability of the i^{th} pixel in the k^{th} cluster at the n^{th} iteration.

N : The number of images sets/modalities.

$c_i^{+/-}$: The internal/external $(+/-)$ mean intensities of the enclosed contour region at level set = 0 in the i^{th} image set.

$\lambda_i^{+/-}$: User defined importance weights for inclusion/exclusion $(+/-)$ from a region defined by the enclosed contour at level set in the i^{th} image set.

ϕ : A level set function.

Ω: The domain of the image.

II.A. Possible classifications

The first objective of this document is to provide introductory information about the different classes of PET auto-segmentation (PET-AS) algorithms. Classifications of PET-AS algorithms can be based on several different aspects:

1. The segmentation/image processing algorithm employed and its assumptions and complexity;
2. The use of pre- and post- processing steps;
3. The level of automation;

The first classification, relying on the type of image segmentation paradigm (e.g., simple or adaptive thresholding, active contours, statistical image segmentation, clustering, etc.), has been used in previous reviews[27, 29, 35, 36]. In most cases, detailed descriptions of the numerical algorithms and their assumptions and limitations are given.

The second classification is based on the use of pre- and post- processing steps. Most algorithms do not use pre-processing steps, although some use either denoising or deconvolution image restoration techniques before the segmentation or as part of the algorithm itself[37, 38]. Other algorithms require either an image-based database[28, 39, 40] to build

a classifier (*i.e.* learning algorithms), or phantom acquisitions for the optimization of parameters (*i.e.* adaptive threshold algorithms).

Regarding the third classification based on automation, Udupa, et al.[41] divide image segmentation into two processes: recognition and delineation, and point to the "essential" need of "incorporation of high level expert knowledge into the computer algorithm, especially for the recognition step." For this reason, most existing algorithms rely on the identification of the tumor first, by the user drawing a volume of interest (VOI) around the tumor to delineate (denoted from here onwards as "standard user interaction", see Table 1), whereas other approaches require the identification of the tumor after the segmentation process in the resulting map (e.g., Belhassen, et al.[28]). Other examples of manual interaction are user-definition of background regions (used by some of the adaptive threshold algorithms), manual selection of markers to initialize the algorithm [42], or the manual input of parameters in case of failure of the automatic initialization[43]. Furthermore, the level of automation can be quite difficult to assess, as factors such as the requirement of building a classifier for each image region, the individual optimization for each combination of scanner system/reconstruction algorithm, the selection and validation of the parameters of the optimization approach, or finally, the detection of lesions to segment, are usually not included in these assessments. In practice, all algorithms require some level of user interaction.

In the following section we used the first classification scheme with emphasis on the algorithm complexity.

II.B. Classes of algorithms for PET auto-segmentation

II.B.1 Fixed and Adaptive Threshold algorithms

Segmentation via a threshold is conceptually very simple. It consists of defining a specific uptake (often expressed as a fixed fraction or percentage of SUV) between the background and imaged object's intensities (tracer uptake) and then using that intensity to partition the image and recover the true object's boundaries. All voxels with intensities at or above the threshold are assigned to one set while the remaining voxels are assigned the other (Appendix I). The details of how the threshold and the uptake values are normalized are discussed in Appendix II.

The decision to use threshold segmentation is generally based upon its simplicity and the ease of implementation. Threshold segmentation carries a number of implied assumptions that should be understood and accounted for. These are:

- The true object has a well-defined boundary and uniform uptake near its boundary, i.e., the image is bi-modal.
- The background intensity is uniform around the object.
- The noise in the background and in the object is small compared to the intensity change at the tumor edge.
- The resolution is constant near the edges of the object.
- The model, used to define the threshold, is consistent with its application, e.g., a segmentation scheme designed for measuring tumor volume may not be appropriate for radiation therapy and vice versa (see section V.A).

In practice, these assumptions rarely hold and some effort is required to determine their validity/acceptability in the context of the intended application.

Several points above are illustrated in a review of PET segmentation by Lee[29]. In this review, the effect of the thickness of the phantom wall on estimating the segmentation threshold and the dependence of this effect on the Point Spread Function (PSF) of the PET scanner, were shown via mathematical analysis. This work also showed that to obtain the correct threshold on a phantom with cold walls (a certain thickness of material without any uptake, such as the plastic surrounding spheres in physical phantoms), a lower threshold is required than for the case without walls, and that due to the limited PET spatial resolution, small volumes require larger thresholds. This was further investigated recently, demonstrating the important impact of cold walls on the segmentation approaches, and the potential improvement brought by thin-wall inserts [44] [45]. A similar result was shown by Biehl, et al [46], who concluded that for NSCLC the optimal threshold for their specific scanner and protocol was related to volume as shown in Appendix I. Finally, it is interesting to note that, given knowledge of the local PSF and the assumption of uniform uptake, the threshold for the lesion's boundary can be estimated analytically, with the result being independent from the tumor to background ratio, provided the background has been subtracted beforehand[47].

Generally, threshold segmentation can be loosely categorized into two separate categories: fixed threshold segmentation (FTS) and adaptive threshold segmentation (ATS). In FTS a general test/model of the problem is developed and a set of parameters is estimated by minimizing the error of the model in order to deduce an "optimal" threshold, T^*. The threshold value may be dependent or not (e.g. 42% of peak lesion activity[48] or SUV=2.5[49]) on the tumor to background ratios. Other tumor or image aspects are generally ignored. For ATS, an objective function is chosen that generates a threshold based upon the properties of each individual tumor/object and PET image. In this case, rather than depending on simple measures, such as tumor to background ratio, the threshold calculation depends upon an ensemble of lesion properties such as volume [46] [48, 50-52] or SUV mean-value[53], and thus makes the threshold segmentation process iterative.

Both FTS and ATS have been discussed in several recent literature reviews of general segmentation in PET[29, 36, 39, 54] Each of these reviews provides a fairly complete literature survey of the state of various threshold segmentation algorithms. In addition, the review by Zaidi and El Naqa[54] provides a brief description and summary of the rationale of many ATS algorithms. These are summarized in Table A1 of Appendix I.

II.B.2. Advanced algorithms

A list of some of the advanced PET-AS algorithms published is given in Table 1, with a focus on the evaluation protocols that were followed. Below they are divided into three subcategories (advanced algorithms applied directly to PET images, approaches combined with image processing or reconstruction, and those dealing with multiple imaging modalities), which are discussed in separate subsections B.2 to B.4.

II.B.2. a) Gradient-based segmentation

The underlying assumption in threshold-based delineation (II.B.1) is that the uptake within the target is significantly different from that in the background. With this idea in mind, the gradient naturally finds the transition contour that delineates a high-uptake volume from the surrounding low uptake regions. The immediate advantage of this alternative method is that uptake inside and outside the target need not be uniform for successful segmentation, nor need it be constant along the contour.

In practice, the method consists of computing the gradient vector for each voxel and then using it to form a new image composed of the gradient magnitude values. Segmentation based on gradient information is an important part of what the human visual system does when looking at natural scenes. The difficulty lies in interpreting the gradient image in order to translate the relevant information into target contours. The general idea is to locate and follow the crests of the gradient magnitude. The points where the gradient is the largest in magnitude (where the second derivative, or Laplacian, is null) correspond to the target contours. There are several ways to locate the crests. For instance, adaptive contours or "snakes" with various smoothness constraints can be programmed in such a way that the contours are attracted towards the crest.[55] Another very popular way to track the gradient crests is the watershed transform. It considers the gradient image as a landscape in which the gradient crests are mountain chains. Then it "floods" the landscape and keeps a record of the boundaries of all hydrographic basins that progressively merge as the water level rises. The hierarchy of all basins can be displayed as a tree in a dendrogram. Clustering tools can help in identifying the branch that gathers all basins corresponding to the target.

The quality of gradient-based segmentation depends on the accuracy and precision of the gradient information, which can be biased by spatial resolution blur. For objects with a concave or convex surface, the uptake spill-in and spill-out caused by blur tends to slightly shift, smooth, and distort the real object boundary. This effect can be partially compensated for with deblurring methods, such as deconvolution algorithms and some tools for Partial Volume Effect (PVE) correction. The gradient computation also amplifies the image noise. Therefore, denoising tools are needed as well, provided they do not decrease the image resolution.

The algorithm described by Geets, et al.[37] relies on deblurring and denoising tools prior to segmentation. The deblurring parameters are adjusted according to the resolution of the PET system and are therefore PET-camera dependent. The watershed transform is applied to the gradient magnitude image and a clustering technique creates a hierarchy of basins. The user can choose the tree branch associated with the high-uptake region in the images expected to correspond to the target volume. In the case of a low signal-to-background ratio (surrounding inflammation, other causes of tracer concentration, uptake reduction due to treatment), the hierarchy can get more complicated and the branch corresponding to the target volume might be difficult to isolate. This usually indicates that the images do not convey enough information for the target volume to be accurately delineated. This approach has been validated using phantom PET acquisitions as well as clinical datasets of both H&N[37] and lung[56] tumors with tri-dimensional (3D) histopathology reconstructions as ground truths.

II.B.2. b) Region growing and adaptive region growing

Region growing algorithms start from a seed region inside the object and progressively include the neighboring voxels to the region if they satisfy certain similarity criteria[57-59]. Similarity is often calculated based on image intensity, but can be based on other features such as textures. Let $I(\boldsymbol{x})$ represent the image intensity at $\boldsymbol{x}$. The similarity criteria can be a fixed interval: $I(\boldsymbol{x}) \in$ [lower, upper], or a confidence interval: $I(\boldsymbol{x}) \in [m - f\sigma, m + f\sigma]$, where m and σ are the mean intensity and standard deviation of the current region, and f is a factor defined by the user[59]. Region growing with a fixed interval is essentially a connected threshold algorithm. Small f restricts the inclusion of voxels to only those having very similar intensities to the mean in the current region, and thus can result in under growth. Large f

relaxes the similarity criteria, and thus may result in over growth into neighboring regions. It is often difficult, if not impossible to identify experimentally an optimal *f* for all objects. For example, four different *f* values were experimentally determined based on the maximum intensity and its location using phantoms by Day, et al.[60]. The authors noted that these *f* values are specific to their clinic.

To overcome this limitation, an adaptive region growing (ARG) algorithm that can automatically identify *f* for each specific object in PET was proposed by Li, et al.[55]. As illustrated in Fig. 1, in ARG *f* is varied from small to large values so that the grown volume changes from the small seed region to the entire image. A sharp volume increase occurs at a certain *f** where the region grows just beyond the object (hot phantom or tumor) into the background (warm water or normal tissue). Because the background typically consists of large homogeneous regions, a great number of voxels are added to the current region at this transition point. The ARG algorithm automatically identifies *f** for which the volume would be increased by more than 200% at the next iterative value of *f*. The resulting volume *V** was proven to be quite an accurate representation of the homogeneous object. The quality of the segmentation performed by ARG depends mainly on the homogeneity of the background and the contrast between the tumor and background. The performance of ARG in segmenting tumors with various levels of heterogeneous uptake still needs to be studied. The ARG algorithm does not have any parameters that require experimental determination. It uses the intrinsic contrast between a tumor and its neighboring normal tissue in each image to determine the tumor boundary. Therefore, it can be directly applied to various imaging conditions such as different scanners or imaging protocols.

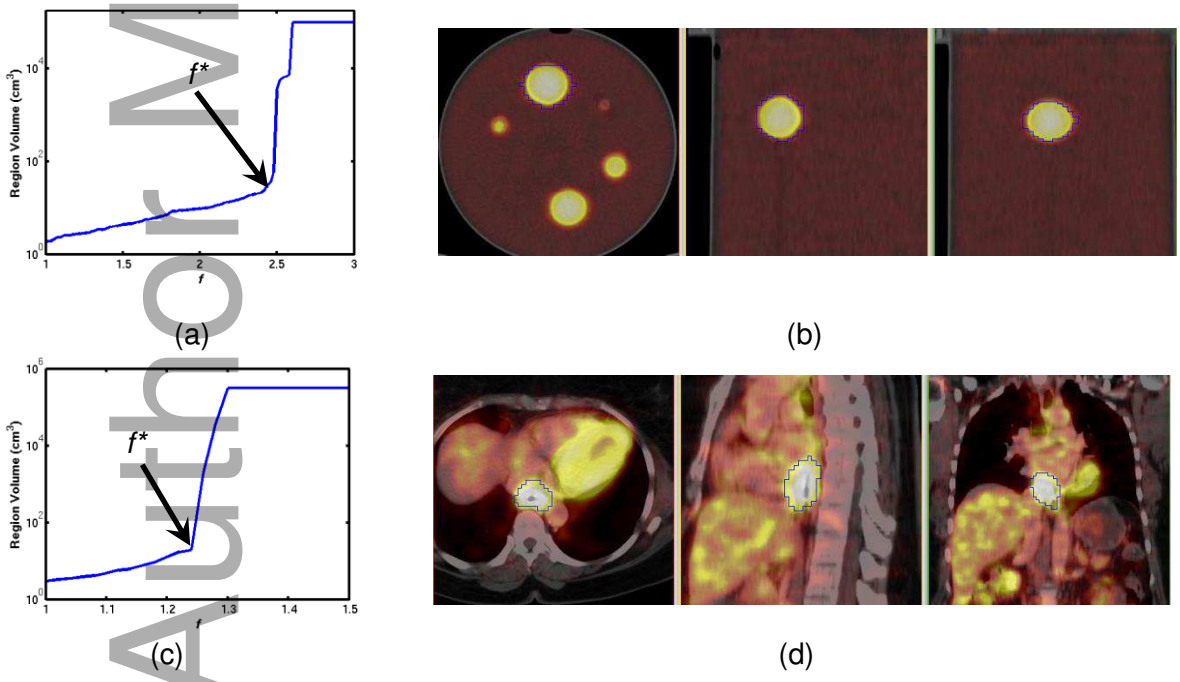


Fig.1. An illustration of applying the adaptive region growing (ARG) algorithm to PET: (a) plot of segmented volume growing as a function of *f*, the arrow indicates the location of the transition point *f** for a spherical lesion in a PET/CT of a phantom (b): the thin blue contour indicates the delineated volume *V**. (c) - (d) selection of *f** and results for an esophageal tumor.

Another approach based on adaptive region growing has been recently proposed by Hofheinz, et al.[61], in which the approach was made able to deal with heterogeneous distributions. The method is based on an adaptive threshold, in which instead of a lesion-specific threshold for the whole ROI, a voxel-specific threshold is computed locally in the close vicinity of the voxel. The absolute threshold T_{abs} for the considered voxel is then obtained based on a parameter T previously determined with phantom measurements (T=0.39): $T_{abs}=T\times(R-Bg)+Bg$, where R is a tumour reference value (e.g., ROI maximum) and Bg is the background. Region growing algorithms use statistical properties (mean and standard deviation) of the region to stop the iterative process[60]. The algorithms, which exploit the statistical properties of a noisy function and a noisy argument and rely on probabilistic calculations, are described in the next subsection.

Table 1. Summary of the main characteristics of some representative advanced PET-AS algorithms (not an exhaustive list) and their respective evaluation.

Reference(s)	Image segmentation paradigm(s) used	User interaction [1]	Pre- and post-processing steps	Aimed application[2]	Validation data and ground truth [3]	Accuracy evaluation on realistic tumors[5]	Robustness evaluation[6]	Repeatability evaluation
Tylski, et al. 2006 [62]	Watershed	Std + multiple markers placement	None	Global	PA(1): Vol. and CiTu images	No	No	No
Werner-Wasik, et al. *2012* [63]	Gradient-based	Std + initialization using drawn diameters	Unknown	Global	PA(5): Diam. 31 MCST: Vol.	Yes	Yes	Yes
Geets, et al. 2007[37]	Gradient-based	Std + initialization	Denoising and deconvolution steps	Global	PS(1) and PA(1): Vol. + Diam. 7 CiTuH: Complete	No	No	No
El Naqa, et al. 2008[64]	AT + active contours	Std + several parameters to set	None	Global	PA(1): Vol. 1 CiTu - Ø	Yes	No	Yes
El Naqa, et al. 2007[31]	Multi-modal (PET/CT) active contours	Std + initialization of the contour shape, selection of weights	Normalization and registering of PET and CT images, deconvolution of PET images.	GTV definition on PET/CT	PA(1): Vol. 2 CiTu: MC(1), FT	Yes	No	No
Dewalle-Vignion, et al. 2011[65]	Possibility theory applied to MIP projections	Std	None	Global	PA(1): Vol. 5 MCST: Vox. 7 CiTuH: Complete	Yes	No	No
Belhassen and Zaidi 2010[28]	Improved Fuzzy C-Means (FCM)	A posteriori interpretation of resulting classes in the segmentation of entire image	Denoising, wavelet decompositions	Global	3 AST: Vox. 21 CiTuH: Diam. 7 CiTuH: Complete	Yes	No	No
Aristophanous, et al. 2007[66]	Gaussian mixture modeling without spatial constraints	Std + initialization of the model and selection of the number of classes	None	Pulmonary tumors	7 CiTu: Ø	No	No	Yes
Montgomery, et al. 2007[67]	Multi scale Markov field segmentation	A posteriori interpretation of resulting segmentation on the entire image	Wavelet decompositions	Global	PA(1) : Vol. 3 CiTu: Ø	No	No	No
Hatt, et al. 2007[68]	Fuzzy Hidden Markov Chains	Std	None	Global	PS(1) and PA(2): Vox.	No	No	No

Hatt, et al. 2009[69], 2010 [70],2011[43]	Fuzzy locally adaptive Bayesian	Std	None	Global	PS(1) and PA(4): Vox. 20 MCST: Vox. 18 CiTuH: Diam.	Yes	Yes	Yes
Day, et al. 2009[60]	Region growing based on mean and SD of the region	Std + optimization on each scanner	None	Rectal tumors	18 CiTu: MC(1)	No	No	No
Yu, et al. 2009[40], Markel, et al. 2013[71]	Decision tree built based on learning of PET and CT textural features	A posteriori interpretation of resulting segmentation on the entire image	Learning for building the decision tree	GTV definition of H&N and lung tumors	10 CiTu: MC(3) 31 CiTu: MC(3)	No	No	No
Sharif, et al. 2010[72], 2012[73]	Neural network	A posteriori interpretation of resulting segmentation on the entire image.	Learning for building the neural network	Global	PA(1): Vol. 3 AST: Vox. 1 CiTuH: Diam.	No	No	Yes
Sebastian, et al. 2006[74]	Spherical Mean shift	Std	Resampling in a different spatial domain	Global	280 AST: Vox.	No	No	No
Janssen, et al. 2009[75]	Voxels classification based on time-activity curve	Std + Initialization and choice of the number of classes	Only on dynamic imaging, denoising and deconvolution steps	Rectal tumors in dynamic imaging	PA(1): Vol. + Diam. 21 CiTu: MC(1)	No	No	No
De Bernardi, et al. 2010[76]	Combined with PVE (image reconstruction)	Std + initialization	PSF model of the scanner and access to raw data required	Global	PA(1): Vol.	No	Yes	No
Bagci, et al. 2013[33]	Multimodal random walk	Std	Multimodal images registration	Global for PET/CT or PET/MR	77 CiTu: MC(3) PA(1)	Yes	No	No
Onoma, et al. 2014 [77]	Improved random walk	Std	None	Global	PA(1), 4 AST: Vox, 14 CiTu: MC(2)	Yes	No	No
Song, et al. 2013[32]	Markov field + graph cut	Std	None	Global	3 CiTu: MC(3)	Yes	No	No
Hofheinz, et al. 2013[61]	Locally adaptive thresholding	Std + one parameter to determine on phantom acquisitions	None	Global	30 AST: Vox.	Yes	No	No
Abdoli, et al. 2013[78]	Active contour	Std + several parameters to optimize	Wavelet decompositions	Global	1 AST: Vox. 9 CiTuH: Complete. 3 CiTuH: Complete 2 CiTuH: Complete	No	Yes	No
Mu, et al. 2015 [79]	Level set combined with PET/CT Fuzzy C-Means	Std	None	Specific to cervix	7 AST: Vox, 27 CiTu: MC(2)	Yes	No	No
Cui, et al. 2015 [80]	Graph cut improved with topology modeling	Std + one free parameter previously optimized	PET/CT registration	Specific to lung tumors and PET/CT	20 PA(1), 40 CiTu(2)	Yes	No	No

Lapuyade-Lahorgue, et al. 2015 [81]	Generalized fuzzy C-means with automated norm estimation	Std	None	Global	PA(4): Vol. 34 MCST: Vox. 9 CiTu: MC(3).	Yes	Yes	Yes
Devic et al, 2016 [82]	Differential uptake volume histograms for identifying biological target sub-volumes	Selection of three ROIs encompassing PET avid area; iterative decomposition of differential uptake histograms into multiple Gaussian functions	None	Isolation of glucose phenotype driven biological sub-volumes specific to NSCLC,	None	No	No	No
Berthon et al, 2016 [83]	Decision tree based learning using nine different segmentation approaches (region growing, thresholds, FCM, etc.) with the goal of selecting the most appropriate method given the image characteristics	Std	Learning on 100 simulated cases to train/build the decision tree	Global	85 NSTuP: Vox.	Yes	No	No
Schaefer et al, 2016 [84]	Consensus between contours from 3 segmentation methods (contrast-oriented, possibility theory, adaptive thresholding) based on majority vote or STAPLE	Std	None	Global	22 CiTuH; Complete 10 CiTu: MC(4) 10 CiTu: MC(1)	Yes	No	Yes

Table legend

[1] std = « standard » interaction (i.e. the metabolic volume of interest is first manually isolated in a region of interest that is used as an input to the algorithm.)

[2] global = not application specific

[3] PA(x) = Phantom (spheres) Acquisitions on x different scanners; PS(x) = Phantom (spheres) Simulations on x different scanners; AST = Analytically Simulated Tumors; MCST = Monte Carlo Simulated Tumors; NSTuP = Non spherical tumors simulated in phantoms (thin-wall inserts, printed phantoms, etc.). CiTu = Clinical Tumors; CiTuH = Clinical Tumors with Histopathology

Vol. = only volume; Vox. = voxel-by-voxel; Diam = histopathology maximum diameter; Complete = 3D histopathology reconstruction; MC(x) = manual contouring by x experts; FT = fixed threshold, AT = adaptive threshold

[4] Highly heterogeneous, complex shapes, low contrasts and rigorous ground truth.

[5] Requires multiple acquisitions on different systems and a large number of parameters.

[6] With respect to different scanners and protocols

II.B.2. c) Statistical

Statistical image segmentation

Statistical image segmentation aims at classifying pixels/voxels and creating regions in an image or volume based on the statistical properties of these regions and voxels, by relying

on probabilistic calculations and estimation for the decision process. Numerous approaches have been proposed; most are based on Bayesian inference. In essence, it is assumed that the observed image Y (usually taking its values in the set of real numbers) is a noisy and degraded version of a ground truth field X (usually taking its values in several classes C). Therefore, X has to be estimated from Y, assuming that X and Y can be modeled as realizations of random variables. These algorithms usually combine an iterative estimation procedure of the parameters of interest, since parameters defining the distributions of X and Y are not known in real situations. In addition, a decision step to classify voxels (i.e. assigning a label among the possible values of X to each voxel, based on its observation Y) and the estimated distributions of X and Y, are required. Hence, the voxel classification is carried out based on the previously estimated statistical properties and the resulting probabilities for each voxel to belong to a specific class or region.

Spatial and observation models

The parameters of interest are usually defined within both a spatial model of X (also called *a priori model*) and an observation model of Y (also called a *noise model*). Most spatial models are based on Markovian modeling of the voxels field, such as Markov chains, fields, or trees, although simpler spatial neighboring definitions (blind, adaptive or contextual) also exist[85]. Noise models are used to model uncertainty in the decision to classify a given voxel, and are most often defined using Gaussian distributions, but more advanced noise models have also been proposed, allowing for the modeling of correlated, multi-dimensional and non-Gaussian noise distributions[86]. Parameters estimation is usually carried out using algorithms such as Expectation Maximization (EM), Stochastic EM (SEM) or Iterative Conditional Estimation (ICE), depending on the assumptions of the model. These methods have been demonstrated to provide robust segmentation results in several imaging applications, such as astronomical, satellite or radar images, by selecting appropriate noise models.

Adaptation to PET image segmentation

Some of the algorithms above, have been applied to PET image segmentation. One example is the use of a multi-resolution model applied to wavelet decomposition of the PET images within a Markov field framework[67]. Another approach is a mixture of Gaussian distributions for classification without spatial modeling[66]. Although these models are robust for noisy distributions of voxels (each voxel has an assigned label, but its observation is noisy), they do not explicitly take into account imprecision of the acquired data (a given voxel can contain a mixture of different classes). Therefore, they do not include the modeling of the fuzzy nature of PET images. As a result, to be applied efficiently to PET images, which are not only intrinsically noisy but also blurry due to Partial Volume Effect (PVE), more recent models can be used that allow the modeling of the imprecision within the statistical framework, using a combination of “hard” classes and a fuzzy measure. In such a model, the actual image, X does not take its values in a set number of classes, but in a continuous [0,1] interval: the fuzzy Lebesgue measure being associated with the open interval (0,1) and the Dirac measure being associated with {0} and {1}[87]. Such a model has been proposed using Markov chains[88] and fields[85] and also using local neighborhoods without Markovian modeling. These models retain the flexibility and robustness of statistical and Bayesian algorithms versus noise, with the added ability to deal with more complex distributions, due to the presence of both hard and fuzzy classes in the images. The Fuzzy Locally Adaptive Bayesian (FLAB) method takes advantage of this model[69], which had previously been proposed within the context of Markov chains[68]. In addition, FLAB modeling has been extended to take into account heterogeneous uptake distributions by considering three

classes and their associated fuzzy transitions instead of only two classes and one fuzzy transition. The extended FLAB model has been validated on phantom acquisitions and simulated tumors, as well as clinical datasets[70].

II.B.2. d) Learning and texture-based segmentation algorithms

For PET image segmentation, the learning task consists of discriminating tracer uptake in lesion voxels (foreground) from surrounding normal tissue voxels (background) based on a set of extracted features from these images[28]. Two common categories of statistical learning approaches have been proposed: supervised and unsupervised[89, 90]. Supervised learning is used to estimate an unknown (*input*, *output*) mapping from known (labeled) samples called the training set (e.g., classification of lesions given a database of example images). In unsupervised learning, only input samples are given to the learning system without their labels (e.g., clustering or dimensionality reduction).

In machine learning and classification there are two steps: training and testing. In the training step, the optimal parameters of the model are determined given the training data and its best in-sample performance is assessed. This is usually followed by a validation step, aimed at optimal model selection. The testing step then specifically aims to estimate the expected (out-of-sample) performance of a model with respect to its chosen training parameters. A recent example of such a development is the ATLAAS method [83], which is an automatic decision tree that selects the most appropriate PET-AS method based on several image characteristics, achieving significantly better accuracy than any of the PET-AS methods considered alone. There are also numerous other types of machine learning techniques that could be applied to PET segmentation, such as random forest, support vector machines, or even deep learning techniques [91], which have been applied to the task of image segmentation in other modalities such as MRI or CT [92, 93]. Although these approaches are promising for the future of PET image segmentation, the use of these techniques for PET is currently rather scarce in the literature [94]. Today these techniques are exploited to classify patients in terms of outcome based on characteristics extracted from previously delineated tumors[95, 96].

PET-AS algorithms can be trained on pathological findings or physician contours. The advantage of training an algorithm using these contours is that additional information, not present in the PET image, is taken into account since the physician draws contours based on additional a priori information (anatomical imaging, clinical data, etc.). On the other hand, training algorithms using physician contours can be biased by the particular physician's background, goals, or misconceptions.

One of the most used approaches to extract image features that can be used for segmentation is texture analysis. Uptake heterogeneity in PET images can be characterized by using regional descriptors such as textures. Unlike intensity or morphological features, textures represent more complex visual patterns composed of entities or sub-patterns, that have unique characteristics of brightness, color, slope, size, etc.[97] "Image texture" can refer to the relative distribution of gray levels within a given image neighborhood. It integrates intensity with spatial information resulting in higher order histograms when compared to common first-order intensity histograms. Texture-based algorithms heavily use image statistical properties; however, since human visual perception often relies on subtle visual properties, such as texture, to differentiate between image regions of similar gray level

intensity, they are separated from the iterative, model-based approaches described in the previous section.

Furthermore, the human visual system is limited in its ability to distinguish variations in gray tone and is subject to observer bias. Variation in image texture can reflect differences in underlying physiological processes such as vascularity or ordered/disordered growth patterns. The use of automated computer algorithms to differentiate tumor from normal tissue based on textural characteristics offers an objective and potentially more sensitive algorithm of tumor segmentation than those based on simple image thresholds. Among the methods that have been suggested to calculate image texture features are those based on (1) Gabor filters, (2) discrete wavelet transforms (DWT), (3) the co-occurrence matrix, (4) neighborhood gray-tone difference matrices (NGTDM) and (5) run-length matrices.

Gabor filters[98] and DWT[99] measure the response of images to sets of filters at varying frequencies, scales and orientations. The Gabor filter (a Gaussian phasor), using a bank of kernels for each direction, scale and frequency, can produce a large number of non-orthogonal features, which makes processing and feature selection difficult. DWTs take a multi-scale approach to texture description. Orthogonal wavelets are commonly used resulting in independent features. DWT, however, have had more difficulty discriminating fractal textures with non-stationary scales[100].

The co-occurrence matrices proposed by Haralick, et al. [101] and spatial gray level dependence matrix (SGLDM) features, are based on statistical properties derived from counting the number of times pairs of gray values occur next to each other. These are referred to as "second-order" features because they are based on the relationship of two voxels at a time. The size of a co-occurrence matrix is dependent on the number of gray values within a region. Each row (*i*) and column (*j*) entry in the matrix is the number of times voxels of grey values *i* and *j* occur next to each other at a given distance and angle. Higher order features refer to techniques that take into account spatial context from more than two voxels at a time. Amadasun and King proposed several higher order features based on NGTDM[102]. For every gray level *i*, the difference between this level, and the average neighborhood around it, is summed over every occurrence to produce the i^{th} entry in the NGTDM.

Another category of higher order features makes use of "run-length matrices." In this case, analysis of the occurrence of consecutive voxels in a particular direction with the same grey level is used to extract textural descriptors such as energy, homogeneity, entropy, etc.[103] However, run-length matrices are a computationally intensive means of deriving texture descriptors[103].

Although textural features have been used to characterize uptake heterogeneity within tumors after the segmentation step[15,96] their use as a means of automatic segmentation can also provide additional information beyond simple voxel intensity that may improve the robustness of delineation criteria. This has been shown in multiple modalities including ultrasound (US) [104] and MRI[105]. PET and CT textures in the lung have been used in a series of applications including differentiating between malignant and benign nodes[106, 107], judging treatment response[15, 16], diagnosing diffuse parenchymal lung disease[108-110], determining tumor staging, detection and segmentation[111]. With dual modality PET/CT systems (also PET/MRI in the near future[112, 113]), it is also possible to make use of image textures from PET and CT (MRI) in combination to improve image segmentation results. However, this leads to including anatomy for tumor volume characterization, instead of

characterizing the functional part of the tumor only. In two separate studies, combinations of PET and CT texture features in images of patients with H&N cancer[114] and those with lung cancer[71] improved tumor segmentation with respect to the dual modality ground truth, versus using PET and CT separately. This is discussed in more detail in section II.B.4 below.

Within the learning category would also fall the recent approaches to account for a set or contours generated via multiple automatic methods, through averaging/consensus methods, [84] statistical methods such as the "inverse–ROC (receiver operating characteristic)" approach,[115] STAPLE (simultaneous truth and performance level estimation)-derived methods[116], majority voting[117], or decision tree[83] to generate a surrogate of truth. Most of these methods would need some type of "training" or preliminary determination of parameters for the particular type of lesions and may therefore avoid the limitations of the individual methods used.

II. B. 3. Combined with image processing and/or reconstruction

The limited and variable resolution of PET scanners, which results in anisotropic and spatially variant blur affecting PET images, leads to PVE, spill-in and spill-out of activity in nearby tissues[17] and is therefore one of the main challenges for segmentation and for uptake quantification of oncologic lesions. In principle, all the segmentation strategies not explicitly intended for blurred images, but widely used for imaging modalities less affected by PVE than PET (e.g., thresholding, region growing, gradient-based algorithms, etc.),[118] can be applied to PET images after a PVE recovery step[119]. PVE recovery can be performed after[120-124] or during image reconstruction with algorithms taking into account a model of the scanner PSF[125-127]. These images however, should be handled with caution since PVE recovery techniques can introduce artifacts (e.g., variance increase related to Gibbs phenomena). The accuracy of PVE-recovered images can be improved by introducing regularizations such as a priori models, constraints or iteration stopping rules. An approach of this kind has been followed by Geets, et al.[37] (described in section II.B.2.a) where a gradient-based segmentation algorithm was applied on deblurred and denoised images. In order to avoid Gibbs phenomena artifacts near the edges, deconvolution was refined with constraints on the deconvolved uptake.

An alternative approach to account for blur is to model it explicitly in the segmentation procedure. For example, FLAB[69], described in section II.B.2.c, or FHMC (Fuzzy Hidden Markov Chains)[68], parameterize a generic form of uncertainty in order to assign special intermediate classes for the blurry borders of the main classes. Such algorithms, if combined with a post-segmentation PVE recovery technique for objects of known dimension/shape, like recovery coefficients, geometric transfer matrix[17] or VOI-based deconvolution[128], may also be able to provide an estimate of PVE-recovered lesion uptake inside the delineated borders[129].

Another means to account for PVE recovery in segmentation is to model it in an iterative process. The lesion border estimate can be iteratively refined using the result of the PVE recovery inside the lesion area and vice versa. Such an approach can potentially improve the estimation accuracy while providing a joint estimate of lesion borders and uptake. This approach was originally proposed by Chen, et al. for spherical objects[130]. More recently De Bernardi, et al. have further developed the idea by proposing a strategy that combines segmentation with a PVE recovery step obtained through a targeted maximum likelihood (ML) reconstruction algorithm with PSF modeling in the lesion area[38]. A scheme of the approach is shown in figure 2.

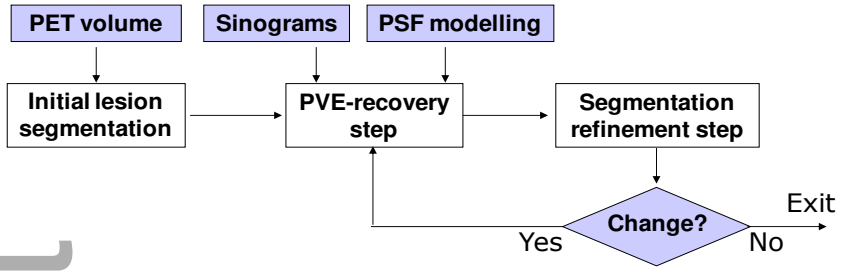


Fig.2. A schematic representation of the algorithm proposed by De Bernardi, et al.[38], which combines segmentation and PVE recovery within an iterative process.

To reduce blur in the latter approach, algorithms using transition regions between lesion and background are employed. These regions correspond to spill-out due to PVE and are modeled by regional basis functions in the PVE recovery reconstruction step. The reconstruction adjusts the activity inside each region according to the ML convergence with respect to the sinogram data. The subsequent segmentation refinement step acts on the lesion borders in the improved image, until borders no longer change. A requirement of the algorithm is that a model of the scanner PSF and access to raw data are available. Conversely, the advantage is that a joint estimate of lesion borders and activity can be obtained.

In the work of De Bernardi, et al.[38] the segmentation was obtained by using *k*-means clustering and the refinement was achieved by smoothing the result with the local PSF and by re-segmenting. The algorithm, suited for the simplest case of homogeneous lesions, was validated in a sphere phantom study. More recently, an improved strategy was proposed, in which the segmentation is performed with a Gaussian Mixture Model and PVE recovery is performed on a mixture of regional basis functions and voxel intensities. The algorithm was validated on a phantom in which lesions are simulated with zeolites (see section IV.C.1)[131].

II.B.4. Segmentation of multimodality images

Multimodality imaging is of increasing importance for cancer detection, staging and monitoring of treatment response[132-136].

In radiotherapy treatment planning, significant variability can occur when multiple observers contour the target volume[137]. This interobserver variability has been shown to be reduced by combining information from multimodality imaging and performing single delineations on fused images, such as CT and PET, or MRI and PET[25, 138-142]. However, traditional visual assessment of multimodality images is subjective and prone to variation. Alternatively, algorithms have been proposed for integrating complementary information into multimodality images by extending semi-automated segmentation algorithms into an interactive multimodality segmentation framework to define the target volume[31-34].

Consequently, the accuracy of the overall segmentation results would be improved, although, as a word of caution, it should be emphasized that the goal may be different from mono-modality delineation and its realization would depend on the application endpoint combined with the clinical association objective of the different image modalities. For instance, in radiotherapy planning, the main rationale behind the use of combining several images of different modalities to define the GTV is that they complement each other by combining different aspects of the underlying biology, physiology and/or anatomy. However,

in reality, this may not be the case for all patients and all pathologies, for example the lesion may not be seen in the additional modality, or may exhibit an artifact. In addition, misregistration between the different modalities and respiratory motion may lead to a potentially erroneous GTV if the images were simply fused without careful consideration of geometric correspondence and the logic by which the different image data are combined (union, intersection or other forms of fusion).

Exploitation of multimodal images for segmentation has been applied to define myocardial borders in cardiac CT, MRI and ultrasound using a multimodal snake deformable model[143]. Another example is the classification of coronary artery plaque composition from multiple contrast MRI images using a *k*-means clustering algorithm[144]. To define tumor target volumes using PET/CT/MRI images for radiotherapy treatment planning, a multi-valued deformable level set approach was used as illustrated in Figure 3[31]. This approach was extended further later on using the Jensen Renyi divergence as the segmentation metric [34].

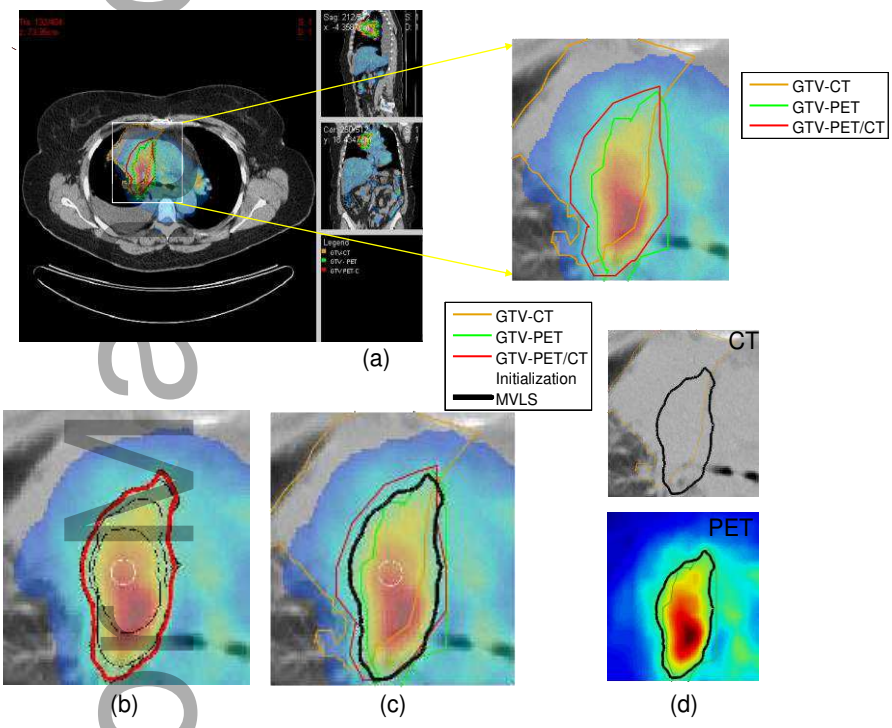


Fig. 3. (a) PET/CT images of a patient with lung cancer in case of atelectasis (lung collapse), with manual segmentation for CT (orange), PET (green) and fused PET/CT (red). (b) The multi-valued level sets (MVLS) algorithm initialized (white circle), evolved contours in steps of 10 iterations (black), and the final contour (red). (c) MVLS results shown along with manual contour on the fused PET/CT. (d) MVLS contour superimposed on CT (top) and PET (bottom). Reproduced with permission from El Naqa, et al.[31].

Mathematically, approaches that aim at simultaneously exploiting several image modalities represent a mapping from the imaging space to the ”perception” space as identified by experts such as radiation oncologists[64]. Several segmentation algorithms are amenable to such generalization[145]. Among these algorithms are multiple thresholding, clustering such as *k*-means and fuzzy *c*-means (FCM) and active contours. In the case of multiple thresholding, CT volumes can be used to guide selection of PET thresholds [46] or using thresholds on the CT intensities to constrain the PET segmentation[145]. These conditions are typically developed empirically but could be optimized for a specific application. For clustering, the process is carried out by redefining the image intensities and

clustering centers as vectors (with elements being the intensities of the different modalities) in contrast to the typical scalars used in single modality images[144]. The formalism for FCM is given in Appendix I. However, both thresholding and clustering algorithms in their basic form suffer from loss of spatial connectivity, which is accounted for in active contour models by using a continuous geometrical form such as the level sets. The level set provides a continuous implicit representation of geometric models, which easily allows for adaptation of topological changes and its generalization to different image modalities. Assuming there are *N* imaging modalities, then using the concept of multi-valued level sets (MVLS)[146, 147] the different imaging modalities are represented by a weighted level set functional objective of the different modalities and the target boundary is defined at the zero level set[31] (Appendix I).

Finally, other approaches based on the Markov field combined with graph-cut methods[32], as well as random walk segmentation[33] or including topology[80], were developed and validated on clinical images, for multi-modal (PET, CT, MRI) images tumor segmentation, with promising results.

II.B.5. Vendor implementations

Here we provide a brief summary of the vendor implementations of PET-AS methods at the time when this report was written. Therefore, it may not describe the PET-AS methods provided by the vendors at the time of publication due to constant evolution of vendor software. Vendors also provide tools for manual segmentation that have been omitted for brevity. Since the algorithms implemented by vendors are not exactly known, the summary and classification provided below do carry a significant degree of uncertainty.

Gradient based edge detection tool is avaible by MIM Software Inc. (Cleveland, OH, see Section II.B.2 a) and Table 1,[63,148]). VelocityAI (Varian Medical Systems|Velocity Medical Solutions, Atlanta, GA) also point that their tool uses “rates of spatial change” in the segmentation process. PET-AS methods based on region growing tools (Section II.B.2 b) are available by Mirada XD (Mirada Medical, Oxford, UK) and RayStation (RaySearch Laboratories AB, Stockholm, Sweden).

Adaptive thresholding approaches (Setion II.B.1) are available by VelocityAI (the method by Daisne, et al.[149]) , GE Healthcare VCAR™ system (V 1.10) (GE Healthcare Inc., Rahway, NJ, the method by Sebastian, et al.[74],see Table 1), and ROVER (ABX GmbH, Radeberg, Germany, an iterative approach following Hofheinz et al[26, 61]).

Finally, practically all vendor implementations use some type of fixed or adaptive threshold based method (Setion II.B.1). For example, Varian’s Eclipse V.10 (Varian Medical Systems, Inc., Palo Alto, CA) as well as other vendor implementations including Philips Healthcare Pinnacle™ (Philips Healthcare, Andover, MA) and Raystation allow users to perform PET segmentation using thresholding in different units (Bq/ml or different SUV definitions), and percent from peak SUV.

III. COMPARISON OF THE PET-AS ALGORITHMS BASED ON CURRENT PUBLICATIONS

A comparison of PET-AS algorithms based on published reports is difficult and subject to controversy because each algorithm has been developed and validated (and often

optimized) on different datasets, often using a single type of scanner and/or processing software. However, some limited conclusions can be drawn. For instance, it is possible to compare the algorithms based on their level of validation as well as those algorithms that have been applied to the same datasets. Table 2 contains a survey of various papers in which several algorithms were compared, providing the type of datasets and methods used, the conclusions of the study, as well as some comments.

Most of the algorithms have been optimized/validated on phantom acquisitions of spheres, as this is a common tool in PET imaging to evaluate the sensitivities, noise properties and spatial resolution of PET scanners. On one hand, most algorithms usually give satisfactory results in these phantom acquisitions, even for varying levels of noise and contrast levels. However, homogeneous spheres on a homogeneous background are not realistic tumors. The number of algorithms that have been successfully applied to realistic simulated tumors or real clinical tumors with an acceptable surrogate of truth (e.g., histopathological measurements) is much smaller. Finally, algorithms that have been validated for robustness against several scanner models and their associated reconstruction algorithms are even less numerous since the datasets are not usually made publicly available.

It should also be emphasized that there are a few algorithms that have been applied to common (although not publically available) datasets. For instance, the gradient-based algorithm by Geets, et al.[37], the improved fuzzy *c*-means (FCM) by Belhassen and Zaidi[28, 150], the theory of possibility applied to Maximum intensity projections (MIP) by Dewalle-Vignon, et al.[65] and the Contourlet-based active contour model by Abdoli, et al.[78] have all been applied to a dataset of seven patients with 3D reconstruction of the surgical specimen in histology (from a dataset of nine patients originally obtained in a study by Daisne, et al.[151]), with 19±22%, 9±28%, 17±13% and 0.29±0.6% volume mean errors, respectively. Similarly, the improved fuzzy *c*-means by Belhassen, et al. [28], FLAB by Hatt, et al.[70,152] and the level sets and Jensen-Rényi divergence algorithm by Markel, et al.[34] were applied to the NSCLC tumors dataset with maximum diameters from MAASTRO (Maastricht Radiation Oncology)[139] (with ±6% error for FLAB, ±15% for the improved FCM and ±14.8% for the level sets approach, respectively). In addition, most of the advanced algorithms that have been proposed have been compared to some kind of fixed and/or adaptive thresholding using their respective test datasets and have, for the most part, demonstrated improvements in accuracy and robustness. In particular, it was observed that fixed and adaptive thresholding might lead to over 100% errors in cases of small and/or low-contrast objects and significant underestimation (-20 to -100%) in cases of larger volumes with more heterogeneous uptake distributions, whereas advanced methods were able to provide more satisfactory error rates (around or below 10 to 20% errors)[152, 153]. However, it is possible that simpler, e.g., adaptive threshold PET-AS-methods optimized for a specific body site, may perform comparably well or even better than some of the more advanced techniques[154].

In the largest comparison to date, Shepherd, et al. [115] segmented 7 VOIs in PET using variants of threshold-, gradient-, hybrid image-, region growing- and watershed-based algorithms, as well as more complex pipeline algorithms. Along with manual delineations, a total of 30 distinct segmentations were performed per VOI and grouped according to type and dependence upon complementary information from the user and from simultaneous CT. According to a statistical accuracy measure that accounts for uncertainties in ground truth, the most promising algorithms within the wider field of computer vision were a deformable contour model using energy minimization techniques, a fuzzy *c*-means (FCM) algorithm and

an algorithm that combines variants of region growing and the watershed transform. Another important finding was that user interaction proved in general to benefit segmentation accuracy, highlighting the need to incorporate expert human knowledge, and this in turn was made more effective by visualization of PET gradients or CT from PET-CT hybrid imaging.

There is little information to date concerning the comparison of the performance of an algorithm using datasets from different scanners and/or with the implementation of that algorithm in different software packages. In one study[155], an adaptive threshold segmentation algorithm was applied in three centers using two similar types of scanners from the same manufacturer. The authors demonstrated that significant differences were observed in the optimal threshold values depending on the center and imaging protocols, despite that both the scanner and reconstruction method were the same. In addition, significant differences were also observed, depending on the reconstruction settings (Fig. 4). They concluded that synchronization of imaging protocols can facilitate contouring activities between cooperating sites. In another investigation, dependence of the segmentation threshold providing the correct sphere volume on the reconstruction algorithm was also observed for small spheres[156].

In a German multicenter study[157], Schaefer, et al. evaluated the calibration of their adaptive threshold algorithm (contrast-oriented algorithm) for FDG PET-based delineation of tumor volumes in eleven centers, using three different scanner types from two vendors. They observed only minor differences in calibration parameters for scanners of the same type, provided that identical imaging protocols were used, whereas significant differences were found between scanner types and vendors. After calibrating the algorithm for all three scanners, the calculated SUV thresholds for auto-contouring did not differ significantly.

On the other hand, the FLAB algorithm by Hatt, et al. showed robustness to scanner type and performed well without pre-optimization, on four different scanners from three vendors (Philips GEMINI GXL and GEMINI TF, Siemens Biograph 16 and GE Discovery LS) using a large range of acquisition parameters such as voxel size, acquisition duration and sphere-to-background contrast.[43]

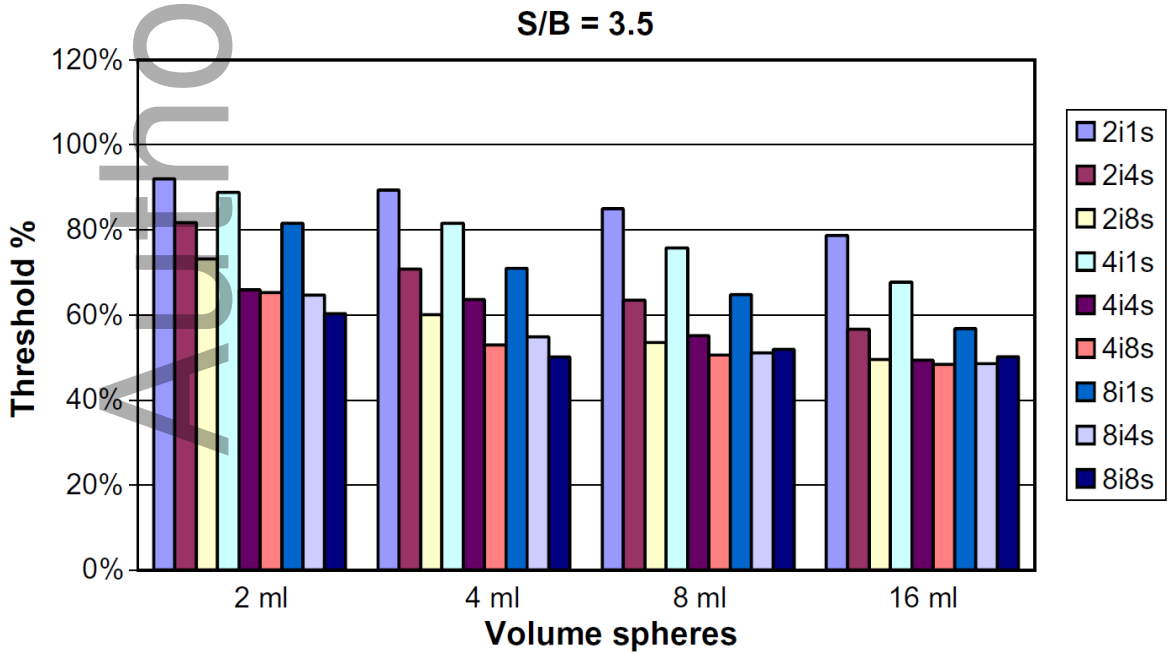


Fig. 4: Variation of the optimal threshold value (y axis) obtained according to different settings of the PET reconstruction with varying number of iterations and subsets (from 2 iterations 1 subset to 8 iterations 8 subsets, colored bars), and for spheres of different volumes (x axis) and a sphere-to-background ratio of 3.5, for one single scanner model. Reproduced with permission from Ollers, et al.[155]

While the natural incentive is to create algorithms which perform universally well across body sites and disease types, for at least one body site it was shown[154] that simpler (e.g., adaptive threshold) methods may perform comparably well if specifically optimized for these conditions. At present there is not a sufficient amount of published data to give specific recommendations for each clinical site. The emerging consensus[84] and decision tree[83] based methods however, provide a potential to provide adequate solution for each site if appropriately adapted and trained.

Given the above results, the validation of PET-AS algorithms, as described in current publications, does not provide sufficient information regarding which of the known approaches would be most accurate, applicable, or convenient for clinical use. In the following sections, we attempt to lay the basis for a framework that avoids the methodological weaknesses of the past and addresses the challenges inherent in segmentation in PET.

Table 2. A summary of segmentation method comparisons and reviews.

No.	Reference	Compared methods	Images or Phantoms used	Results and/or recommendations reported by the authors	Limitations and comments by TG211 or others as cited
			Comparison studies		
1	Nestle, et al. 2005[23]	Visual segmentation, 40% of SUV_{max} threshold, SUV>2.5 threshold, and an adaptive threshold	Patient scans	Large differences between volumes obtained with the 4 approaches	Only visual segmentation used as a surrogate of truth and only clinical data.
2	Schinagl, et al. 2007[158]	Visual segmentation, 40% and 50% of SUV_{max} threshold, and adaptive thresholding	78 Clinical PET/CT images of head and neck	The 5 methods led to very different volumes and shapes of the GTV. Fixed threshold at SUV of 2.5 led to the most disappointing results.	The GTV was defined manually on CT and used as a surrogate of truth for PET-derived delineation and only clinical data was used.
3	Geets, et al. 2007[37], Wanet, et al. 2011[56]	Fixed and adaptive thresholding, gradient-based segmentation	Phantom (spheres), simulated images, clinical images of lung and H&N cancers with	More accurate segmentation with gradient-based approach compared to threshold	Numerous issues associated with the 3D reconstruction of the surgical specimen used as gold standard.

			histopathology 3D measurements		
4	Greco, et al. 2008[159]	Manual segmentation, 50% SUV_{max}, SUV>2.5 threshold and iterative thresholding	12 Head and neck cancer patients.	Thresholding PET-AS algorithms are strongly threshold-dependent and may reduce target volumes significantly when compared to visual and physician-determined volumes.	Limitations: reference GTVs defined manually on CT and MRI.
5	Vees, et al. 2009[160]	Manual segmentation, SUV>2.5 threshold, 40% and 50% of SUV_{max} threshold, adaptive thresholding, gradient-based method and region growing.	18 Patients with high grade glioma	PET often detected tumors that are not visible on MRI and added substantial tumor extension outside the GTV defined by MRI in 33% of cases. The 2.5 SUV isocontour and "Gradient Find" "segmentation techniques performed poorly and should not be used for GTV delineation".	Ground-truth derived from manual segmentation on MRI only.
6	Belhassen, et al. 2009[161]	Three different implement-ations of the fuzzy C-means (FCM) clustering algorithm	Patient scans	Incorporating wavelet transform and spatial information through nonlinear anisotropic diffusion filter improved accuracy for heterogeneous cases	No comparison with other standard methods
7	Tylski, et al. 2010[42]	Four different threshold methods (% of max activity and three adaptive thresholding), and a model-	Spheres in an antropomor-phic torso phantom as well as non spherical simulated tumors	Large differences between volumes obtained with different segmentation algorithms. Model-based or background-	No clinical data, limited to threshold-based algorithms only, only volume error considered as a metric

		based thresholding		adjusted algorithms performed better than fixed thresholds.	
8	Hatt, et al. 2010[70], 2011[43, 152]	Fixed and adaptive thresholding, Fuzzy C-means, FLAB	IEC phantom (spheres); simulated images, clinical images with maximum diameter measurements in histopathology	Advanced algorithms are more accurate compared to threshold- based and are also more robust and repeatable.	For clinical images, only maximum diameters along one axis were available from histology.
9	Dewalle-Vignion, et al. 2011[65], 2012[162]	Manual segmentation, 42% of SUV_{max}, 2 different adaptive thresholding, fuzzy c-mean and an advanced method based on fuzzy set theory.	Phantom images, simulated images, clinical images with manual delineations	The advanced algorithm is more accurate and robust than threshold-based and closer to manual delineations by clinicians	Only manual delineation for surrogate of truth of clinical data. Comments: Interesting use of various metrics for assessment of image segmentation accuracy.
10	Werner-Wasik, et al. 2012[63]	Manual segmentation, fixed thresholds at 25% to 50% of SUV_{max} (by 5% increments) and gradient-based segmentation	IEC phantom (spheres) in multiple scanners, simulated images of lung tumors	A gradient-based algorithm is more "accurate and consistent" than manual and threshold segmentation.	Only volume error used as a metric of performance. "…manual verification using CT scan should be performed (…) GTV definition requires joint assessment by the radiologist, nuclear physician and radiation oncologist" [163]
11	Zaidi, et al. 2012[150]	Five thresholding methods, Standard and improved fuzzy c-means, level set technique,	Patient scans with histopathology 3D measurements (same as #5 above) N/A	The automated Fuzzy c-means algorithm provided was shown to be more accurate than 5 thresholding algorithms, the level set technique, the	Numerous issues associated with the 3D reconstruction of the surgical specimen used as gold standard. See Table 1

		stochastic EM.		stochastic EM approach and regular FCM. Adaptive threshold techniques need to be calibrated for each PET scanner and acquisition/pro-cessing protocol and should not be used without optimization.	
12	Shepherd, et al. 2012 [115]	30 methods from 13 different groups.	Tumor and lymph-node metastases in H&N cancer and physical phantom (irregular shapes). Simulated, experimental and clinical studies	Highest accuracy is obtained from optimal balance between interactivity and automation. Improvements are seen from visual guidance by PET gradient as well as by using CT.	Limitation: a small number of objects (n=7) were used for the evaluation.
13	Schaefer, et al. 2012[157]	One adaptive thresholding technique	Phantoms, same threshold algorithm, different scanners	The calibration of an adaptive threshold PET-AS algorithm is scanner and image analysis software-dependent.	Confirmation of previous findings about adaptive threshold segmentation.
14	Schinagl et al. 2013 [164]	Visual, SUV of 2.5, fixed threshold of 40% and 50%, and two adaptive threshold based methods using either the primary or the metastasis	Evaluation of the segmen-tation of metastatic lymph nodes against pathology in 12 head and neck cancer patients	SUV of 2.5 was unsatisfactory in 35% of cases; for the last four methods: i) using the node as a reference gave results comparable to visual ii) using the primary as a reference gave poor results;	Shows the limitations of threshold based methods.
15	Hofheinz, et al. 2013, [61]	Voxel-specific adaptive thresholding and standard lesion-specific adaptive threshold	30 simulated images based on real clinical datasets.	The voxel-specific adaptive threshold method was more accurate than the lesion-specific one in heterogeneous cases	Only simulated data were used.

16	Lapuyade-Lahorgue, et al. 2015 [81]	Improved generalized fuzzy c-means, fuzzy local information C-means and FLAB	34 simulated tumors and 9 clinical images with consensus of manual delineations. 3 acquisitions of phantoms for robustness of evaluation.	In both simulated and clinical images, the improved generalized FCM led to better results than another FCM implementation and FLAB, especially on complex and heterogeneous tumours, without any loss of robustness on data acquired in different scanners.	Only 9 clinical images used.
Reviews					
1	Boudraa, et al. 2006 [39]	N/A	Mostly clinical images	Extensive review of the fomalism of image segmentation algorithms used in nuclear medicine (not specific to PET and clinical oncology)	Only a few algorithms have been rigorously validated for accuracy, repeatability and robustness.
2	Lee, 2010[29]	N/A	Simulated, mostly clinical images	Discussed the main caveats of threshold-based techniques including the effect of phantom cold walls on threshold. A discussion of the available and desirable validation datasets and approaches is also included.	Extensive review of nuclear medicine image segmentation algorithms (not specific to PET and clinical oncology)
3	Zaidi and El Naqa, 2010 [54]	N/AN/A	Simulated, experimental and clinical studies	Despite being promising, advanced PET-AS algorithms are not used in the clinic.	
4	Hatt, et al. 2011[23]	N/A	N/A	Only a few algorithms have been rigorously validated for accuracy, repeatability and robustness.	See the three last columns of Table 1
5	Kirov A.S, Fanchon,	N/A	Clinical images with	Articles comparing PET-AS methods are	

	L, 2014,[165]		pathology derived ground truth.	summarized separately for lesions in five groups based on location in the body with a focus on the accuracy, usefulness and the role of the pathology-validated PET image sets.	
6	Foster, et al. 2014[36]	N/A	N/A	"although there is no PET image segmentation method that is optimal for all applications or can compensate for all of the difficulties inherent to PET images, development of trending image segmentation techniques which combine anatomical information and metabolic activities in the same hybrid frameworks (PET-CT, PET-CT, and MRI-PET-CT) is encouraging and open to further investigations."	Most exhaustive review of the state-of-the-art in 2014. Uses a similar classification of methods as in the present report.

IV. COMPONENTS OF AN EVALUATION STANDARD

A main conclusion of the work of this task group is that a common and standardized evaluation protocol or "benchmark" to assess the performance of PET–AS methods is needed. The design of such a protocol requires:

- Selection of evaluation endpoint and definition of performance criteria;
- Selection of a set of images;
- Selection of contour evaluation tools;

IV.A. Evaluation endpoints

For the purpose of radiation therapy the PET image is most often used to segment the so called gross tumor volume (GTV) which contains the macroscopically observable (demonstrable) disease[166]. Based on the GTV is later generated the clinical target volume

(CTV), which is supposed to include additional volume with a high probability of microscopic tumor extensions[166]. The planning target volume (PTV) encompasses the CTV and adds and additional margin for set-up error and organ motion[166]. The sub-target volume (sub-GTV) lies within the GTV and locates one or more metabolically distinct sub-volumes such as tumor growth, burden or hypoxia. Segmentation of the sub-GTV assumes availability of additional functional information, which can come from PET or from other imaging modalities[2]. Within this context, the smallest number of cells with uptake that is possible to image using PET has been assessed as 10^5 cells[167].

The endpoint for evaluating PET-AS algorithms can be selected at different levels of approximation of the tumor border. We consider the following three levels of approximation ordered from least to most accurate:

a) The PET avid tumor volume in the PET image as obtained through standard reconstruction, which typically leaves some of the physical artifacts (such as limited spatial resolution, motion), not routinely and/or fully corrected;
b) The PET avid tumor volume after optimal correction of more subtle artifacts such as resolution, motion and noise;
c) The spatial distribution of the biological quantity of clinical interest (e.g., the distribution of cells exhibiting a certain metabolic trait, e.g., proliferation).

The ideal endpoint for evaluating PET-AS algorithms if they are to facilitate reaching the clinical goal is (c). However, variations in the biological environment of the lesion (e.g., perfusion and inflammation) and other biological and physical uncertainty in PET images decrease the accuracy of numerical algorithms in aiming at the clinical endpoint (e.g., the GTV). The less ambitious endpoint (b) of contouring the volume based on the real tracer distribution is feasible, provided that important factors, such as PVE, motion and noise are accurately taken into account or corrected with state-of-the-art approaches (either within reconstruction or post-reconstruction).

Finally, most algorithms have been and can be evaluated against the activity as seen in the standard PET image (a). This case concerns standard acquisitions with routine clinical systems for which some of the physical artefacts (attenuation, scattered and random events, etc.) are corrected, but no correction is applied for others (e.g., spatial resolution, motion, statistical noise and post-filtering). This method is currently widely used. Nevertheless, for future standardized evaluation protocols, our task group recommends considering the three endpoints listed above[168]. Such future work should also consider segmenting radiotherapy targets using multi-spectral images from hybrid imaging studies[31, 33 ,34,114], dynamic imaging[75, 169] and/or multi-tracer PET images[170].

IV.B. Definition of performance criteria: accuracy, precision (reproducibility and repeatability) efficiency and robustness

In instrumental science any measurement tool can be characterized by its accuracy (degree of closeness to the true value) and precision (degree to which repeated measurements under unchanged conditions give the same results). Precision can be further stratified into repeatability and reproducibility. Reproducibility often implies that tools or operators are different, whereas repeatability relies on experimental conditions that are kept as identical as possible. For complex tools such as segmentation algorithms, stratification into reproducibility and repeatability is not necessary and precision suffices, provided all

parameters are identified, including those for which the operator has control, e.g., the region of an image used to characterize the background tracer uptake.

For a given segmentation algorithm we define accuracy as the correctness of retrieving the true 3D object spatial extent, shape and volume based on the reconstructed activity distribution in a PET image, irrespective of the correlation between this distribution and the underlying physiological process. This means that an image segmentation algorithm is not expected to differentiate specific from non-specific tracer uptake (e.g., inflammation and tumor in the case of FDG) if they are of the same intensity.

Within the context of this report, repeatability[1] is defined as the ability of a given algorithm to reach the same result when applied multiple times on a single image replicate (single acquisition), given potentially differing algorithmic initializations. In such a task, deterministic, fixed threshold approaches will always give the same result when applied to a given image. On the other hand, more advanced algorithms are susceptible to providing different results when applied with multiple runs on the same image because they could rely on more complex initializations or estimation processes, including random ones.

We define robustness as the ability of a given algorithm to generate consistent, segmented volumes under varying acquisition and image reconstruction conditions, including issues related to statistical counts and multiple replicates (multiple acquisition of the same object) due to noise. This robustness is determined as the variability of the segmentation results when a PET-AS algorithm is applied on images of the same object acquired using various scanners, and for each scanner, under various contrast and noise conditions, using different reconstruction and associated correction algorithms.

Finally, an important parameter of the algorithms is their efficiency, which may determine their practical viability[41]. Efficiency includes workflow and computational complexity required for completion of the segmentation task. Considering the computing power evolution and possibilities (parallel computing, graphical processing units, etc.) the main limiting factor is workflow and human interaction.

Below is laid out the vision of the task group for a future standard for PET-AS method evaluation. It has two main components: 1) Benchmark image set; 2) Performance evaluation criteria.

IV.C. Benchmark image sets

This section is dedicated to the selection of the benchmark images, which should cover a realistic range of parameters so that it ensures that the tested PET-AS algorithms can meet the challenges that may be encountered in various clinical cases. However, in order to allow for a practical and realistic evaluation and interpretation of the results, the number of images and datasets should be kept to a minimum. Therefore, the images that are likely to offer the most realistic, rigorous way to assess the performance of the various algorithms should be selected. A classification of the possible types of benchmark images is given in Table 3.

[1] The term *reproducibility* or repeatability is also used to denote the variability assessed using double baseline PET scan acquisitions (repeated acquisitions at a few days interval without treatment). This "physiological" reproducibility is a different topic than the repeatability/reproducibility of the PET-AS algorithm discussed here.

Table 3. Advantages and disadvantages of the different methods used for generation of test images.

	Experimental images			**Simulated images**	
	Simple Phantoms	**Realistic phantoms**	**Clinical images**	**Forward Projected Images**	**Monte Carlo (MC) Simulations**
Advantages:	• Exact representation of the scanner resolution, image noise and other image artifacts • Ground truth accurately known • Easy to generate and use	• Exact representation of the scanner resolution, image noise and other image artifacts • Capable to produce lesion shapes corresponding to actual tumors • Known ground truth	• Exact representation of the scanner resolution, image noise and other image artifacts • Real tumors	• Precise experimental control • Flexibility in phantom design • Precise knowledge of the reference object • Computationally cheap	• Precise experimental control • Realistic count distributions • Flexibility in phantom design • Precise knowledge of the reference object • Camera-specific information
Dis-advantages:	• The objects have simplistic and unrealistic shape and activity distribution • Most with few exceptions[171, 172] have cold walls	• Some uncertainties in the experiment can be hard to quantify • Difficult to generate inhomogeneous activity • Labor intensive	• Uncertainties in the knowledge of the reference object, even with histopathology reference	• Scatter count distributions and noise are usually less accurately modeled • Detailed physics and system information ignored	• Computationally expensive • Model requires extensive up front experience

The advantages and disadvantages of the different classes of datasets and of particular published image sets are discussed in more detail below. The various phantoms considered for a common PET-AS evaluation protocol are summarized in Appendix III.

IV.C.1 Physical phantoms

The main advantage of physical phantoms is that they contain the same exact degradations, namely resolution, noise, scatter, etc., as clinical PET scans, while also ensuring that the ground truth is both reproducible and known for repeated testing using prescribed conditions.

Most PET-AS algorithms are initially developed and optimized against simple phantoms containing uniform activity spheres or cylinders. Therefore, these phantoms are essential in evaluating segmentation accuracy for well-defined, simple-shaped objects. However, spherical targets oversimplify the segmentation problem and can erroneously favor an algorithm that would break down in the presence of a complex topology or heterogeneous tracer uptake distribution seen in real tumors. Testing the PET-AS algorithms against these images can nevertheless provide: 1) assurance that the algorithms compared are trustworthy for simple cases; 2) agreement limits for initial, basic evaluation; 3) opportunity for verifying algorithm operation over time (e.g., routine quality assurance) and 4) a convenient tool for testing the robustness of the algorithms under different experimental conditions (Fig. T.A2.1 in Appendix III) using for instance the National Electrical Manufacturers Association (NEMA)

image quality phantom[173], available in most PET centers. However, with few exceptions[171,172] these phantoms contain objects with cold walls in a homogeneous background[173]. Furthermore, most of the algorithms have already been optimized and assessed on these simplistic, physical phantom acquisitions at the development stage.

More realistic phantoms are of interest for more demanding evaluations with respect to the activity distribution endpoint. The contribution by Zito, et al.[174] regarding the use of phantoms containing zeolites (microporous, aluminosilicate minerals commonly used as commercial adsorbents which can absorb aqueous solutions of (18)F-FDG) is promising. This phantom allows tumor-like objects to be generated with any desired shape, size and contrast levels without cold walls. They also provides ground truth with sub-voxel resolution that is available from the associated co-registered CT images.

A limitation of these images (e.g. obtained from zeolite phantoms) is the lack of control and knowledge of the potential heterogeneity of the tracer uptake in the background and the "tumor." Several alternatives allow experimental modeling of non-uniform activity inside the lesions and in the background. These phantoms include structures generated by stacking paper sheets containing PET images printed with radioactive ink[175, 176] (see Fig.T.A1.6. in Appendix III), or 3D printers using radioactive ink[177]. Another option for generating non-uniform uptake distributions is the use of thin sheets to displace activity (see Fig.T.A1.7 in Appendix III)[178].

IV.C.2 Simulated images

Virtual or numerical phantoms associated with a PET-image generation process represent an inexpensive, precise way to test PET software and clinical methodologies[179]. One definition of a virtual or numerical PET simulation framework corresponds to any computer-generated object that is processed to produce a PET-like image. It should be clear that virtual phantoms are distinct from the resulting PET images and represent a reference source distribution from which the PET-like image is produced. In order to be useful the resulting image needs to be representative of what is observed in the images produced by a real PET camera.

Generating PET-like images for virtual phantoms can be done in several ways. Below we describe methods that range from simple to complex, as more realism is included in the simulation, and therefore, in the produced images.

Inserted Tumor PET-Like Images:

The simplest method is to insert an object with added noise, representing a tumor, directly into an existing PET image[180]. However, this method requires considerable effort to blend the noise and edge characteristics of the lesion into the image in order to avoid obvious edges from threshold or texture mismatches. This method is the least realistic of the various approaches for generating PET-like images. Because of these weaknesses and the difficulties in accurately matching the noise/spatial resolution properties of real PET images, it is not further discussed.

Forward Projected Tumors:

Alternatively, a more robust method is to consider a synthetic lesion that can be forward projected, have noise added and then inserted into the noiseless forward-projection of the existing PET image that is scaled appropriately to match the desired noise level of the tumor.

The projection data can then be reconstructed to preserve the basic characteristics of the original image. Care must be taken to ensure that the forward and backward projectors are matched, i.e., adjoint, that the original PET image is sufficiently oversampled and that the reconstruction process does not greatly alter the underlying PET image. Although this process is conceptually simple, its realism is limited by inaccuracies introduced in modeling the spatial variation of the PSF, the noise model and the effects of the reconstruction process.

Forward Projected Phantom Images:

Forward projected phantoms and tumors represent a middle ground between full Monte Carlo PET simulations of phantoms and directly inserting tumors into PET images. This has been implemented in an open source simulation tool[181] and used in several PET studies, as it is a standard means for evaluating image reconstruction methods[182, 183].

In this method, noiseless tumor and phantom images are forward projected and scaled to produce a similar number of total counts as would be seen in the equivalent projection data and then fused, which represents the reference images. Noise is then added to the resulting projection data via a Poisson distribution to create PET-like projection data. These data are then reconstructed to produce the PET-like images of the original virtual phantom. Additional realism can be included by blurring the images with a PSF (derived from physical parameters: positron range, annihilation photon non-collinearity, detector solid angle, block effects, etc.), adding attenuation, random and scatter counts and altering the fidelity of the projection matrix or the type of reconstruction. This process is described in Figure T.A1.8. in Appendix III. Motion can also be simulated by applying the appropriate motion-blurring kernel to the image prior to forward projecting the image into sinogram space. This method can be extended to insert realistic tumors into existing PET images[181].

Monte Carlo simulations:

The most realistic data can be obtained by simulating the entire positron emission, annihilation, interaction and detection processes with Monte Carlo (MC) simulations. The subsequent projection data can be reconstructed to produce very realistic images.

Using recent, state-of-the-art anthropomorphic phantoms such as the XCAT (4D NURBS-based Cardiac-Torso)[184] or Zubal phantoms[184, 185] and MC simulators such as SORTEO (Simulation Of Realistic Tridimensional Emitting Objects) [186], GATE (Geant4 Application for Tomography Emission)[187, 188] or SimSET (Simulation System for Emission Tomography)[189], combined with scanner system modeling (geometry, detectors, etc.)[190], can provide highly realistic simulations, including respiratory motion[153] with regular or irregular respiratory signals. Simulated tumors can be placed in various anatomical locations and generated with non-spherical shapes and complex uptake distributions, including realistic "activity gradients" (see Figure T.A1.9 in Appendix III). PET data are then simulated by assigning an uptake to each organ/tumor of the anatomical phantom. Parameters such as tumor-to-background or intra-tumor heterogeneities can be varied within any desired range. Similarly, it is possible to generate various noise realizations, as well as various SNR ratios, by selecting different parts of the overall simulated list mode data (lines of response) before reconstruction. It is therefore also possible to select lines of response corresponding to true coincidences only, or including the random and scattered data. Different scanner designs and reconstruction algorithms and/or parameters (number of iterations, post-filtering smoothing, voxel dimensions, etc.) can also be modeled if detailed information about the

scanner is available; hence this method allows the assessment of robustness and "universality" of the PET-AS algorithms.

Simulated data can provide a high level of realism without the disadvantages and inconveniences of real phantom acquisitions. It is possible to increase the number of activity levels to realistic numbers approximating ground truth to the voxel level achievable by some experimental approaches.[177, 178] This however can increase the complexity and time required for the design of the simulation.

IV. C.3. Clinical images

In patients, the "ground truth" is defined by the actual underlying extent of disease; however, the true biological margins are usually unknown. This is in contrast to phantoms, where the ground truth is clearly defined by the phantom design and therefore well-known. For clinical images the following surrogates of truth can be used: i) a consensus of several physicians or expert-drawn contours and/or ii) histopathological measurements of lesions resected within a reasonably short timeframe after the image acquisition and for which special precautions are taken as described below.

Consensus of several physician-drawn contours

If the clinical endpoint is selected as the decision basis in the absence of histopathology information, consensus of several physician-drawn contours is sometimes used as a surrogate of truth. When the segmentation contours can potentially be used in different clinical applications, images contoured by several experts or physicians from different specializations (e.g., the study by Bayne et al.[191] in which 2 radiologists, 2 radiation oncologists and 2 nuclear medicine physicians contoured 5 NSCLC patients), might reduce bias due to personal and specialty-based preferences at the price of a likely slightly higher inter-observer variability due to differences in training and habits. On the other hand, when considering a specific clinical application, such an approach may be less accurate than using consensus of contours drawn by several specialists in this specific application. Indeed, contouring by physicians from only one specialty, e.g., radiation oncologists[25], may provide more reliable estimates for an endpoint corresponding to the goals of this specific sub-specialty (target volume definition in this case).

The use of the consensus based methods discussed at the end of Section 2.B.2.d, which can account for a set of manual contours, may be expected to reduce errors under certain assumptions about the operators, as differences in performance or training can be taken into account (e.g., within the STAPLE framework).

Histopathological validation of PET image segmentation

This type of PET-AS validation can be carried out by using PET images of tumor specimens for which histopathological characterization is also available. In this case, PET-AS contours can be tested directly against the histopathology-derived contours. At present, these data serve as the most clinically relevant ground truth of tumor extent. However, there are several sources of errors that limit the accuracy of this surrogate of truth for PET-AS validation: (i) variable amount of deformation of the surgical specimen after excision, (ii) time difference between the PET scan and the specimen excision, (iii) uncertainty associated with manual delineation (usually by a single observer) of the tumor boundaries in digitized histopathology and (iv) imperfect co-registration of histopathology slices and PET volumes. While these errors can potentially limit the validity of the comparison, histopathological validation is an important part of thorough PET-AS evaluation. At present, there are several

datasets in which effort was made to minimize these errors[165]. Examples are: the lung tumor dataset from the MAASTRO (Maastricht Radiation Oncology) team with pathology-validated maximum diameters[139], the tumor datasets used in the study from the Jefferson Medical College (max. diameter)[192], the HNSCC (Head and Neck Squamous Cell Carcinoma),[151] the NSCLC[56] full 3D volumes reconstruction datasets from the Université Catholique de Louvain studies and the lobectomy-based dataset from The Netherlands Cancer Institute (NKI)[193]. Despite the challenges with respect to the accuracy of the reference contours, these pathology-validated images provide an important test for PET-AS methods. This justifies the need for further improvement of the current experimental approaches, as well as development of new techniques to improve the accuracy of histopathological validation[165, 194, 195].

It is important to note that because histopathological validation of PET image segmentation is carried out for a particular tissue and tracer pair, it cannot be implied that the results apply to alternative PET tracer/tissue combinations; hence, one should exercise care when using PET-AS algorithms to segment tracer/tissue pair images different from those for which they were validated. It should be noted that evaluating PET segmentation against anatomic or surgical delineation could be potentially misleading since biodistribution of a particular PET radiotracer may not conform to these structures. This is especially true for non-FDG tracers such as hypoxia probes where the entire tumor volume is not expected to display uptake.

IV.C.4. Blind study and updates

In order to facilitate the training and validation of PET-AS algorithms, it would be optimal to separate the images of the future standard into two groups: a) With ground truth given to the PET-AS developers for learning/training and b) Blind study (without ground truth) for testing. The rationale behind a) is that some algorithms, e.g., the learning algorithms may need to be trained, whereas b) will ensure more objective evaluation and validation. Simple geometrically shaped phantoms naturally fall in the first category, whereas clinical images are a natural candidate for the second group. Simulated images, or complex shape experimental phantoms, can be distributed among the two.

Since both experimental and numerical phantoms are currently in rapid development, it is important to make provisions for updating and expanding the set of images. The benchmark's goal can be better reached if it can facilitate and encourage the sharing of new acquired datasets by contributing users. As new data and PET-AS algorithms become available, the evaluation process can be organized so that the new, shared datasets become gradually included in the standard. For example, a rule may be considered according to which, a certain fraction of the images (e.g., ~ 60%) must have been used for evaluating at least ten algorithms.

IV.D. Figures of merit and evaluation criteria

Choosing the best set of Figures of Merit (FOM) depends on the complexity of the segmentation problem as well as on the evaluated endpoint. For example, when using spheres in a standard compartmental phantom, shape modifications and volume translations are unlikely to be observed. In this case, simple volumetric differences may be enough. In more realistic images, inaccuracies in shape or location are more likely and need to be detected with a more complex FOM. A statistical approach can further distinguish between

two types of errors with respect to assigning a voxel to a lesion or normal tissue: Type I - false positives and Type II - false negatives. The various FOMs are discussed in detail in Appendix IV. The FOMs listed in Table 4 were considered for use in the future standard.

Most of these FOMs have advantages and drawbacks, some of which are listed in Table 4. For example, optimizing sensitivity alone would favor methods that encompass and therefore overestimate the true volume. Similarly, optimizing positive predictive value alone would instead favor methods that underestimate the true volume. Other criteria are not strict enough (e.g., volume difference), computationally expensive (e.g., Hausdorff distance [196]), or unable to distinguish between the two error types (false positive and false negative, e.g., Jaccard and DSC). Therefore, we caution against using a single performance metric for segmentation evaluation and rather suggest reporting several FOMs such as the combination of sensitivity and positive predictive value, in order to convey complementary information.

Table 4. A comparison of various volume/contour agreement measures and their sensitivities to the properties of the segmented lesions. The important properties are whether they account for volume differences, shape discrepancies, false positive vs. false negative. The computational complexity is graded between easy (+) and complicated (+++), although none of the metrics are particularly slow to compute using modern toolkits and computers (Barycenter distance is the distance between the centers of mass of two sets).

Evaluation criteria	Location	Size	Shape	Type I/II	Complexity
Volume difference	no	yes	no	no	+
Barycenter distance	yes	no	no	no	++
Jaccard similarity coefficient	yes	yes	yes	no	++
Dice similarity coefficient (DSC)	yes	yes	yes	no	++
Hausdorff distance	yes	no	yes	no	+++
Sensitivity + Positive Predictive Value (PPV)	yes	yes	yes	yes	++

Notice that the Type I/II error distinction in sensitivity and PPV requires the knowledge of which of the two volumes is the actual ground truth, whereas other measures treat both volumes in the same way. In the absence of a ground truth volume (neither *A* nor *B* is preferred), then the Dice similarity coefficient can be used instead of Sensitivity + PPV.

As discussed in section IV.B, some image datasets, e.g., simulated and experimental images, may have more accurately defined ground truth than others, (e.g., clinical images accompanied with pathological results or manual contours). In the case of a less accurately defined ground truth, the inverse-ROC approach, used by Shepherd, et al.[115], can give a

reliable evaluation of the algorithms, provided a set of contours (*e.g.,* manual delineations) that encompass the ground truth contour exist.

Due to the complexity of the PET segmentation problem, more appropriate evaluation metrics will need to be derived from investigations of correlations between the assessment of these metrics and a clinical endpoint. In the case of radiotherapy treatment planning, an example would be the geometrical concordance of the delivered dose distributions to the PET segmentation contour and the treatment outcome. Tools which can account for such information have recently been proposed[197].

V. DISCUSSION OF SEGMENTATION LIMITATIONS, DEPENDENCIES AND IMPLEMENTATION

V.A Biological limitations of the segmentation concept

It has long been realized that cancer is an abnormal growth caused by unregulated cell proliferation. Cancerous tissue morphology is highly irregular and characterized by chaotic vascularization, resulting in a unique pattern of blood flow for every tumor, which modifies PET tracer availability and uptake in a way unique for each patient. Also, different parts of the same tumor can have very different micro-environmental status, including different levels of glucose metabolism. Other factors affecting intratumoral PET tracer distribution are the presence of necrosis and stromal tissue intertwined with cancer cells. As a result, the intratumoral pattern of FDG uptake is highly heterogeneous.

While it is possible to carry out in vitro studies to relate PET tracer binding/uptake to environmental parameters of the cells in culture, direct in vivo application of such data is highly speculative and lacks strong foundation due to the reasons listed above. The uniqueness and stochastic nature of the factors governing PET tracer uptake and its intratumoral distribution in each patient represents one of the biggest challenges for PET image segmentation. The complexity of the problem hampers the widespread adoption of auto-segmentation tools for routine clinical use.

Other factors can also potentially affect PET-based lesion segmentation. Tumors may lack a well-defined boundary separating them from the surrounding normal tissues. Microscopic cancer extensions can produce additional blurring of this idealized, macroscopic boundary. Furthermore, in addition to heterogeneities of tracer uptake in the lesion, surrounding normal structures are likely to be characterized by different levels of tracer uptake. Inflammation, if present, can result in further complications by significantly increasing FDG uptake. Correspondingly, the biological meaning of the segmented volume should be interpreted in the context of all these biological factors governing image formation in PET. Therefore, both PET image segmentation as well as interpretation of the segmentation result are very non-trivial tasks and should be approached with caution. However, for situations where tumor delineation is needed, e.g., radiation therapy treatment planning, the right choice of properly validated PET-AS methods used as a guidance tool by the physician can result in increased target definition accuracy and better treatment.

V.B. Dependence on segmentation task

There may be significant differences in terms of tumor segmentation algorithm parameterization and use, depending on the task. At the same time, it should be emphasized

that most published methods have been proposed either as a general PET segmentation approach, which can be used in any application (although rarely tested or validated for all), or as a method developed and validated for a specific clinical application (e.g., radiotherapy planning, without being tested in another setting).

In treatment planning the PET information can be used in two ways[198]:

1. Target volume delineation: The PET-based GTV should safely encompass the entire tumor volume without missing regions with low radiotracer accumulation. In order to avoid cancer under treatment, even equivocal voxels would usually be included. However, to avoid over irradiation of too large a volume of normal tissue, the GTV should not be larger than needed. To account for microscopic disease the radiation oncologist then draws the CTV by adding a margin to the GTV (section IV.A).

Uncertainties to the tumor contour for external radiotherapy may be generated based on the accuracy of the method as determined during the evaluation stage. The delineation uncertainty can be approximated as a shell or annular volume around the segmented volume. The thickness of the annular shell could, for example, be derived from the average thickness of the annular volume between the overlap and union volumes of the segmentation and reference surfaces determined during the evaluation stage. Other options are to use distance metrics between these surfaces, which can be based on the Hausdorff distance or similar methods[199].

2. Target substructure determination: In contrast, PET-based definition of tumor sub-volumes for so-called biologically conformal radiotherapy or dose painting[2] requires a different approach. In dose painting, radiation is shaped according to the PET uptake, theoretically delivering higher dose to the radiation resistant and/or tumor-rich parts of the tumor. In order to achieve this goal, one needs to rely on a detailed understanding of the underlying tumor biology and PET signal (e.g., PET tracer uptake and retention mechanisms), as well as how to determine the dose prescription function based on the PET signal. In that specific context, radiotracers different than FDG have been investigated, e.g., use of FMISO-PET might indicate hypoxic regions and the use of FLT-PET might indicate tumor proliferative regions where increased dose is needed. In such cases PET-AS methods would need to be able to define both the entire tumor volume as well as sub-volumes with different levels of activity. In some rare cases, multi-tracer datasets can be available and the images combined to define a biological target volume (BTV). Methods based on information fusion have been proposed to address this specific challenge[170, 200-202].

For treatment response assessment:

1. Segmentation can be used for the estimation of various uptake measurements (mean SUV, total SUV, heterogeneity of uptake using, e.g., histogram-derived first order features or more complex second and third order textural features), which may correlate better with the clinical outcome than less comprehensive metrics, such as maximum or peak SUV.
2. Automatic segmentation can be used for more consistent longitudinal tracking of treatment response to various cancer therapies. Repeatability and reproducibility of segmentation in this case could be more important than absolute accuracy, especially within the context of the known relatively high test-retest variability of PET scan imaging[203, 204].

The PET avid volume and/or on tumor-to-background ratio may change as a result of therapy. Therefore for PET-AS methods which are dependent on these parameters (e.g. some adaptive threshold methods[23, 52]) use of the same method for segmentation of PET images before and after therapy without proper adjustment of parameters may result in incorrect and inconsistent segmentation. This may then affect the accuracy of the metrics derived from the segmented volume. In general using FDG for adaptive radiation therapy may be problematic due to change of the SNR as a result from reduction of the tumor uptake and/or inflammation. This means that, for example, a threshold set to 42% of peak activity may provide erroneous results if the tumor/background ratio changes substantially or if the PET avid volume decreases under a certain value [48]. This volume was found to be about 1.5 mL for older PET scanners but will be partial volume and therefore scanner dependent. Similarly, if the PET avid volume has an irregular shape with both wide and thin parts, the threshold may have to be adapted to the effective size of these parts of the volume.

The time saved by using automatic segmentation is also important; lack of time in daily practice is one of the major limitations preventing investigators from using ROI-based methods for treatment response assessment in cases where volume (or volume-derived) information is important. As a result, in current practice, SUV_{max} and SUV_{peak}, which are less dependent on accurate edge and volume definition, are more widely used for response assessment. Automatic segmentation provides consistency and time efficiency in longitudinal studies. However, consistency is harder to achieve for PET measures dependent on the segmented volume (e.g., SUV_{mean}, SUV_{total}), compared to measures that do not depend on it, but simply follow the voxel(s) with highest activity concentration anywhere within the GTV (e.g., SUV_{max}, SUV_{peak}).

In addition, even for a single clinical goal (e.g., radiation treatment planning), the PET-AS methods may meet different requirements for different disease types and body sites. This may profoundly affect the method evaluation process. For example, this may result in favoring relatively simple, e.g., adaptive threshold methods, optimized for each lesion type versus more complex advanced methods, which may do equally well in different parts of the body. During the development of a future evaluation standard, this possibility may be investigated by sorting the performance results of the PET-AS methods between body sites and tumor types.

V.C. Dependence on scanner, image acquisition and reconstruction protocol

One major consideration in PET image analysis is the lack of standardization of clinical imaging protocols resulting from hardware and software variability, as well as the variation of procedures between clinical centers (injected dose, delay between injection and acquisition, acquisition duration, etc.). Thus, every post-acquisition, post-reconstruction analysis and extraction of relevant parameters from PET images depend on the actual qualitative and quantitative characteristics of the analyzed PET image (e.g., resolution and noise), which are strongly influenced by the acquisition protocol. For this reason users are cautioned to always evaluate and validate published PET-AS methods for their specific clinical application and scanning protocol before clinical use.

Recently, there have been several efforts to propose ways for the standardization of imaging procedures. These efforts have sought to minimize the impact of acquisition protocols on the resulting visual quality and quantitative accuracy and consistency of PET

images[4, 205, 206 207]. One of the main reasons is to help improve consistency in multi-center trials that combine images acquired from different clinical centers, scanners and imaging protocols.

These efforts are to be encouraged. By reducing the existing variability in PET images encountered in clinical practice, they will contribute toward improved data consistency, which will facilitate the use of PET-AS algorithms across different centers and thus allow the use of advanced quantitative tools for treatment assessment. This will also contribute to reducing the dependence of PET defined tumor volumes on the specific instrumentation and protocols in a given clinical center.

The quality of an image is defined by several parameters, which may have different importance for different tasks[208-211]. Segmentation differs from typical diagnostic tasks in that it seeks to identify the boundary locations and therefore uses a much larger parameter space. Based on this, there is an expectation that this problem is more ill-posed and requires less noisy data to reduce errors.

This can be achieved by modifying the injected activity, uptake period, acquisition and image reconstruction. Increasing the injected activity may improve the noise equivalent count rate. Typically, the injected activity is unaltered from that used for diagnostic imaging and it is possible that the risk from therapy is large enough that it outweighs the risks associated with injecting a larger amount of activity. For the case of FDG-PET, a 1-hour post-injection delay is used; however, the contrast ratio of uptake to background continues to increase with time. This 1 h selection is due to tradeoffs between workflow, consistency, diagnostic efficacy, etc. Increased dwell times over the tumor regions and/or additional spot scans can also be used to improve the images.

Increasing the number of counts in the data using these approaches would allow achieving higher resolution image by increasing the number of iterations while preserving the noise level[212, 213]. Beyond this, some penalized image reconstruction methods with edge preserving prior models have been developed[214-218]. These may produce images with edges that are more easily segmented. These tradeoffs may be considered for future protocol optimizations together with the risks associated with higher doses related to therapy.

V.D. Dependence on tracer type and physical isotope

Current investigations are dominated by FDG and ^{18}F-based tracers. This is understandable because FDG remains the most widely used radiotracer in oncologic imaging. However, there is a growing interest in non-FDG tracers, including radiolabeled amino acids such as L-methyl-^{11}C-methionine (MET) or O-(2-^{18}F-fluoroethyl)-L-tyrosine (FET) for brain tumor delineation, proliferation markers such as ^{18}F-3′-fluoro-3′-deoxy-L-thymidine (FLT) or hypoxia tracers such as ^{18}F-fluoromisonidazole (F-MISO). At least some of these agents show a lower intensity of uptake in tumor lesions than FDG (e.g., FLT and FMISO), and thus physicians may apply different criteria for what constitutes significant radiotracer uptake (for instance in comparison to background reference regions or blood activity). In fact, little attention has been given to the question of how the use of these alternate radiotracers can affect the accuracy of the various segmentation algorithms. Most of the segmentation approaches have been designed for FDG-PET. Also most fixed and adaptive threshold-based methods are optimized for a specific range of tumor-to-background ratios. However, some methods have been used successfully on different radiotracers[219, 220]. For some tracers

a lower target-to-background ratio may lead to significant problems in the use of threshold-based algorithms. Tracers other than FDG may be of great interest for dose painting and contouring of tumor sub-volumes. However, for isotopes other than ^{18}F, differing physical parameters, such as positron range and emission of cascade gamma rays, may degrade image quality and must be taken into account. This emphasizes the need for more robust algorithms that can deal with varying contrast and noise levels in reconstructed images. Histological validation of such tracer accumulation is necessary to determine sensitivity, specificity and detection limits before these agents can be considered for dose modulation.

It should be noted that multi-tracer datasets have been acquired in research protocols and clinical trials in order to investigate the complementary value of different tracers. Since the acquisition and investigation of multi-tracer data is currently in its dawn and their segmentation is a very specific and challenging task outside common clinical practice we are limiting their discussion only to this paragraph. Several novel methods have been developed to segment such data, usually with the goal of deriving a single biological target volume (BTV) from multi-tracer images. The use of information fusion has been suggested to achieve this as early as 2011[200-202] and some recent fusion-based methods have been evaluated with promising results[170].

New, more sophisticated pattern recognition/machine learning algorithms are also on the horizon; these may make use of more subtle image characteristics, including noise distribution, underlying PSF, and nominal biological distribution. Such algorithms will require training sets of expert identified and segmented data and will only be valid for the type of data they were trained to process (see section II.B.2). Therefore, while at present most segmentation schemes are radiotracer/isotope agnostic, this may rapidly change, as more sophisticated image-processing techniques become available.

V.E. Effect of motion

Motion can have an important impact on the apparent size, shape and contrast of lesions in PET images, especially in the thoracic area, due to respiratory motion. There has been a significant advancement of respiratory motion correction algorithms based on breath hold, external or internal gating, deformation corrections and post-frame summing, blur deconvolution and others[221-238]. Of these the data driven gating approaches of PET images promise to yield comparable results with less discomfort for the patients and to be more efficient than hardware driven approaches[239-241]. The development of synergistic algorithms, which account for motion simultaneously with segmentation, is also expected[242-244].

In the context of stereotactic body radiotherapy (SBRT) it has been shown that it may be useful to derive respiratory correlated target volumes from gated (4D) PET/CT scans in addition to 4D-CTs [243, 245, 246]. However, the current common practice still is to segment the PET volume integrated over the scan time. If PET-AS algorithms are evaluated on clinical images against the activity as seen in the uncorrected PET image, endpoint a) as described in section IV.A, the potential effect of breathing motion in these images is disregarded. If the CT images are also used in the segmentation process, the uncorrected PET images should not be used for cases potentially affected by motion due to possible misalignment between the CT and PET.

V.F. Guidelines for acceptance and implementation for PET auto-segmentation algorithms

Vendors have adapted and further developed some of the PET-AS methods when implementing them in commercial software. However, the number of published algorithms is much larger than the number of those implemented (see section II.B.5). The algorithms implemented by vendors, while being adaptations of published algorithms, may have modifications and enhancements that have a "black box" quality if vendors are reluctant to disclose proprietary techniques. Therefore, the vendors may have specific recommendations on how to test their PET-AS algorithms, which the user should address first.

An additional factor to consider is the variability in implementations in the various commercial software visualization and analysis platforms. As it has been recently demonstrated, even for very simple metrics such as SUV_{max}, considerable variability has been shown to exist across various vendors and software, likely due to implementation errors, as well as different interpretation of, or assumptions about the data[247]. The developed digital reference object is a very useful tool that will allow verification and validation of the vendor's implementation. Indeed, similar observations were made regarding contours and volumes that were substantially modified when transferred from one station to another[248]. It is, indeed, not uncommon to transfer segmentation results such as contours from one station to another (e.g., a nuclear medicine-dedicated analysis station to a radiotherapy planning station) and the user should verify their consistency. In that respect, considerable standardization efforts are needed to ensure that adopted PET-AS methods will be correctly implemented and the results are compatible across the various platforms of different vendors.

Following vendor suggested acceptance testing, this task group envisions a three phase procedure (Table 5) for the implementation of segmentation algorithms that reflect the different level of closeness to reality of the PET images (see section IV). The images to be used in the three stages would contain lesions represented by; (1) spherical/cylindrical objects, (2) irregularly shaped objects and (3) human datasets. Since each of these image types may present a different evaluation endpoint (IV.A) and specific challenges that depend on how it was generated, this will allow a more thorough evaluation of PET-AS methods.

Most current implementation tests typically stop with the first phase, incorrectly assuming that the PET-AS algorithm would be sufficiently accurate for realistic clinical images. For the first phase, phantoms with spherical inserts (diameter: 1 cm – 4 cm) imaged at varying object-to-background ratios (e.g., 2:1 to 10:1) can be used. In addition, iodinated contrast can be used to aid in segmenting the ground truth volumes and in excluding the wall of the objects in the CT images. Simple shape (e.g., spherical) objects in uniform background, preferably without cold wall,[171, 172, 174] are also most convenient for robustness evaluation across scanners and reconstruction schemes.

For the second phase, a combination of physical phantoms capable of constructing irregularly shaped objects[115, 174] and non-uniform activity distributions[178], as well as numerically simulated phantoms that contain irregular shaped objects and/or non-uniform uptake, can be selected among the family of phantoms discussed in IV.C.1 and IV.C.2. Finally, for the third phase for which we suggest using clinical images, the main limitation is insufficient knowledge of the ground truth. As discussed in IV.C.3, ground truth surrogates such as pathology findings of excised specimens and/or statistical consensus from several manually drawn contours (preferably by different experts) can be used. Since both ground

truth surrogates have a fair degree of uncertainty, the benchmark dataset should ideally comprise both of these image types.

The evaluation metrics for assessment of the segmentation accuracy are described in IV.D. While several of these tools can be used, a combined metric, e.g., including sensitivity, positive predictive value and Hausdorff distance is expected to provide a more reliable method assessment (see. IV.D.3). However, further investigations are needed to generate a combined evaluation metric that is not affected by biases of the metrics or correlations between them. We suggest that the results of the evaluation stage be used to estimate the contouring uncertainty as discussed in section V.B.

A standard, which will provide access to the selected benchmark datasets and various performance metrics, is currently under construction by members of the task group [168, 249, 250]. As pointed out in the last column of Table 5, the ultimate evaluation of segmentation will be analyzing the outcome of treatments using the respective segmentation approach.

Table. 5 Stages of evaluation of PET auto-segmentation (PET-AS) methods. DSC (Dice Similarity Coefficient), PPV (Positive Predictive Values), HD (Hausdorff Distance).

Step	1. Vendor acceptance	2. Basic evaluation	3. Phase two evaluation:	4. Phase three evaluation	5. Impact evaluation
Objective	Proper functioning of software	Accuracy of clinic-specific images; robustness of image properties	Accuracy, repeatability and robustness of realistic shapes and variable uptake;	Accuracy, repeatability and robustness of clinical images from the intended application	Evaluation of clinical impact
Datasets	Vendor recommendation	Simple objects in uniform background; repeated acquisitions	Irregular shape and/or non-uniform uptake lesions in experimental or digital phantoms without cold wall; multiple realizations	Clinical images	Clinical images, treatment plans and follow-up records
Ground truth	Vendor recommendation	CT defined voxel level accuracy.	High resolution CT or digital ground-truth defined voxel level accuracy.	Digitized histopathology and/or consensus of several manual delineations	Treatment outcome data
Metrics	Vendor recommendation	Volume errors, DSC	DSC, Sensitivity, PPV, HD	DSC, Sensitivity, PPV, HD, Statistical evaluation of clinical endpoint (prognostic/predictive value)	Statistical multi-parameter treatment outcome analysis

Ideally, a segmentation algorithm would be portable across different scanners with their individual and sometimes proprietary reconstruction schemes and parameters. Since this

may not always be realistic for many PET-AS algorithms, the implementation should be appropriately tagged as being optimized for specific scanner types and protocols.

The minimum requirements for an algorithm depend on the intended application goal: diagnostic, therapy planning, or treatment/prognostic assessment. For diagnosis, the most important aspect of the PET-AS method is its ability to identify the tumor (not necessarily exact extents/boundaries) for a large range of tumor sizes on either original or PVE corrected images, in order to provide the most accurate volume and the associated activity.

For radiotherapy planning, the minimum requirements include the PET-AS's accuracy in the delineation of the gross tumor volume and the ability to identify sub-volumes (for dose boosting/painting/redistribution applications), as is its ability to achieve a high sensitivity (to be sure to include the entire target) with minimal loss of specificity (to reduce irradiation of healthy tissues and organs at risk).

In the case of response assessment, the main requirement is that the portion of the tumor image that maximizes the predictive power of the particular parameter (biomarker) used, is correctly segmented. As a result, for this type of segmentation task, the link to the physical aspects of the tumor and imaging system are difficult to convincingly establish, and the need for the clinical impact evaluation step (Step 4, Table 5) is especially important[251].

Also in most cases, for follow-up and therapy assessment applications, the PET-AS algorithm will have to be applied to serial scans independently, although developments dedicated to consider simultaneously sequential scans are also being developed[200, 201, 252-255]. Therefore, its robustness versus different contrast, heterogeneity and tumor size is extremely important in order to provide non-biased results regarding the evolution of tumors during therapy. AAPM Task Group 174 (Utilization of 18F-Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) in Radiation Therapy) is working towards standardizing the methodology used for sequential scanning (or even inter-patient scanning for clinical trial patients), so as to allow segmentation techniques to be used for fair comparison between the pre- and intra/post-treatment PET scans.

Type of disease and body site dependence of the performance of PET-AS methods should also be expected. This means that the user should evaluate the chosen PET-AS algorithm for the intended body site. Furthermore, within the context of multi-centric studies, it is important that the chosen algorithm be validated for robustness against the varying noise and texture properties associated with different scanner models and reconstruction algorithms and their associated parameters (voxel sizes, etc.). Alternatively, the algorithm should be easy to adapt/optimize to the characteristics of each individual center/scanner. Scanning the same phantom at the involved institutions and comparing PET-AS method performance is suggested. The limitations of the selected phantom need to be well-understood as discussed in IV.C.

V.G. The complementary role of manual and auto-segmentation for PET

In order to satisfy the requirements laid out in the previous sections, PET-AS algorithms need to accurately account for the physical and technical sources of bias and uncertainty in the PET images. In addition, the ideal PET-AS algorithm should be able to account for anatomical, physiological and other clinical information not present in a PET image, which can alter the location of a contour. Although some of the algorithms listed in Table 1 promise to answer most of the physical requirements, accounting for clinical information not present in

the PET image is beyond the capabilities of most available PET-AS algorithms. As a result, there is a need for active physician involvement in the segmentation process. Therefore, at present and in the near future, automatically generated contours can be used only as a starting point for GTV delineation by the physician, who may decide to change them based on his/her knowledge. It is likely that human supervision will remain necessary, both before and after the process of automatic contouring, although this rule may change in the future[256]. A recent work has presented a method for head-and-neck, in which user interaction is kept minimal but exploited nonetheless so that the user can provide simple cues to guide the segmentation algorithm in an efficient and intuitive manner [257].

Before auto-contouring:

Because automatic contouring algorithms cannot distinguish between malignant and benign tissue tracer uptake, the selection of the lesion, i.e., the diagnostic decision to regard a certain region of elevated tracer uptake as malignant, must be done by a knowledgeable physician. This step includes all forms of diagnostic decision making, considering clinical information not present in the image, topography, pattern and anatomical location of the suspected uptake, as well as the probability for malignant spread.

After auto-contouring:

The review and editing of the final contour is required for consistency with known diagnostic information, including findings by other imaging modalities, endoscopy results and clinical knowledge. The contours drawn on the same lesion may differ if the goal is therapeutic (need to include all malignant tissue) compared to the case when the goal is diagnostic (need to mark structures containing tumor with a high probability).

To ensure a smooth workflow in daily practice, the contouring software should facilitate both automatic contouring and user interactions for lesion selection and contour editing or algorithm guidance (by providing better initialization, for example[258]). It is also necessary to enable co-viewing or fusing of the PET scan with other imaging modalities to include all diagnostic information in the contouring process. In this context, beyond the application of well-designed and thoroughly evaluated algorithms for automatic contouring, the use of multi-modality imaging and collaboration between radiation oncologists and/or oncologists, and imaging specialists (e.g., diagnostic radiologist and/or nuclear medicine expert) are necessary to ensure better understanding of planning images.

VI. CONCLUSIONS

Given the large number of published PET-AS algorithms, their different level of validation and because most of these published algorithms are not yet implemented in commercially available software, recommending a single PET-AS method is challenging and premature. Furthermore, even if such a recommendation could be made it may become obsolete considering the rapid development of the field. Instead, we have provided basis for understanding the logic and the limitations of the main classes of approaches and a framework for their rigorous evaluation and comparison, which we believe will be of greater value for future developments.

As reviewed in this report, there is accumulating evidence in the literature pointing to the higher accuracy and robustness of the approaches based on more advanced image segmentation and analysis paradigms, when supplemented with manual and visual verification, compared to simple threshold-based approaches. These advantages, however,

come at the expense of the ease of implementation and understanding of the simpler algorithms. At the same time it is possible that simpler (e.g., adaptive threshold) methods may perform comparably well, if not better, for a certain body site/disease type if specifically optimized for these conditions. Recent algorithms which employ some type of consensus or automatic selection between several PET-AS methods have a potential to overcome the limitations of the individual methods when appropriately trained. In either case, accuracy evaluation is required for each different PET scanner and scanning and image reconstruction protocol. For the simpler, less robust approaches, adaptation to scanning conditions, tumor type and tumor location by optimization of parameters is necessary. The results from the method evaluation stage can be used to estimate the contouring uncertainty. All PET-AS contours should be critically verified by a physician.

Clearly, further research for solving the dilemma of PET image segmentation is needed, and one potential solution for going forward is the creation of a standardized protocol (i.e., a benchmark) for consistent evaluation and comparison of the PET-AS methods. This task group suggests the following considerations for generating such a standard:

1) The evaluation endpoints need to be clearly separated based on algorithmic accuracy and clinical relevance. In order of increasing clinical relevance the reference choices are: a) the unmodified PET images; b) the tracer distribution corrected for artifacts; c) the underlying histopathology.
2) At present, the benchmark needs to consist of several image datasets of different types: experimental (phantoms), numerically simulated and clinical, in order to compensate for the deficiencies of each of them. Also, a complete set of images should include images from all body sites, since algorithm performance may depend on local tracer uptake specifics.
3) The performance of the methods needs to be evaluated using different metrics, which include volume overlap measures, classification evaluation tools as well as voxel-to-voxel distance metrics.

These considerations are the core of the guidelines for PET-AS algorithm evaluation presented in more details in section V.F.

A standard that conforms to these requirements will provide a more objective comparison of the algorithms by mitigating the large variability of image sets and metrics used for evaluation. At present, a benchmark following these recommendations is under development within the task group. Different PET-AS methods are currently being tested within this framework to evaluate the benchmark design and components[168, 249,250]. A publicly available tool such as this should aid users in evaluating current algorithms to increase confidence in selecting the most adequate PET-AS method to use for a particular application under physician supervision and to provide reference criteria to evaluate future methods.

Acknowledgments:

We want to thank D. Nelson (MiM Software, Cleveland, OH), for his contribution as a consultant (industry observer) and the very helpful staff of the American Association of Physicists in Medicine for providing support to the task group. The work on this report was funded in part by the American Association of Physicists in Medicine and the contribution

1640 from few of us (CRS, HS and ASK) was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748.

Appendix I

PET-AS Formalism examples

A) Fixed and Adaptive Threshold algorithms

Thresholding could be expressed as follows:

$$\hat{I} \in \Upsilon_T(I_i) = \begin{cases} 1, & I_i \geq T, \\ 0, & I_i < T, \end{cases} \quad I_i \in I_{VOI} \tag{A1}$$

where $\Upsilon_T(\cdot)$ is the indicator function for threshold *T*, $\hat{\mathrm{I}}$ is the segmented subset of the voxels within a Volume of Interest (VOI) in image *I*, and $\mathrm{I_i}$ *i*s the uptake value (generally normalized to SUV) at voxel i.

The particular thresholds for performing segmentation are often found by minimizing the difference between known volumes, $\mathrm{V_{known}}$, (typically a phantom study) and the volumes defined by applying different thresholds, $\mathrm{V(T)}$. This often is described as,

$$T^* = \mathrm{argmin}_T \left(V_{known} - V(T)\right)^2 \tag{A2}$$

which could be solved by a least-squares estimation technique. Additionally, it is possible to add some topological constraints to $\hat{\mathrm{I}}$ to ensure its connectedness and/or that it is simply connected, to avoid islands or holes within the segmentation ROI. In case of NSCLC the optimal threshold for a specific scanner and protocol[46] was related to volume via,

$$T(V) = I_{max}\left(59 \log_{10} V\left(\hat{I}\right) - 18\right), \quad I_i \in \hat{I} \tag{A3}$$

where $\mathrm{I_{max}}$ is the maximum uptake in the segmented subset and $\mathrm{V}\left(\hat{\mathrm{I}}\right)$ is the segmented volume. The functional forms of various threshold segmentation schemes are given in Table A 1.

Table A1: Functional forms of various threshold segmentation schemes.

Comments	Threshold Estimator
Drever, et al.'s single-parameter FTS fit[259]: It is most notable for its use of the histogram's mode for more stable estimation of the background.	$T = a\left(I_{max} - I_{bkg}\right) + I_{bkg}$
Nestle, et al.'s single-parameter FTS fit[23]: This fit uses the mean of voxels greater than 70% of the lesion's maximum. The use of the mean instead of the maximum uptake reduces the variability.	$T = a\, I_{mean,70\%max} + I_{bkg}$

Daisne, et al.'s show a two-parameter FTS fit model [149]: The scaling parameter, $\mathrm{I_{max}}$, can be recast as a mean-value or volume-based measure for an ATS algorithm[260].	$T = a + b\,\frac{I_{bkg}}{I_{max}}$
Schaefer, et al.'s two-parameter FTS fit[30]: This fit is extended from Nestle's scheme above[23].	$T = \frac{a\,I_{mean,70\%max} + b\,I_{bkg}}{I_{max}}$
Erdi, et al.'s two-parameter FTS fit[48]: uses a fixed threshold of 42%. It was noted that it worked well for large lung tumors, however the authors go on to say that its use should be limited to homogeneous uptake distributions.	$T = a\,e^{-b\,V(T)}$
Black, et al.'s two-parameter ATS fit[53]: The use of the mean SUV to make the algorithm more stable to noise requires a threshold for its calculation.	$T = a + b\,I_{mean}\big(V(T)\big)$
Biehl, et al.'s two-parameter ATS fit[46]: The volume is the GTV defined by CT. This algorithm is shown to work for a wide range of tumor volumes in NSCLC.	$T = I_{max}\big(a + b\,ln(CT_{GTV})\big)$
Jentzen, et al.'s three-parameter ATS fit[51]: The parameters were fitted from phantom data. The inverse volume requires a threshold.	$T = \frac{a}{V(T)} + b\,\frac{I_{bkg}}{I_{max}} + c$
Nehmeh, et al.'s four-parameter ATS fit[52]: The fit used Monte Carlo simulation results to avoid cold wall effects.	$T = I_{max}\left(a + b\,V^{c}(T)e^{d/V(T)}\right)$
Burger, et al.'s Background Subtracted Lesion (BSL) [261]: Not meant as segmentation but rather a volume estimation scheme, an equivalent volume threshold can be found (Li et al.[262]) Note that this method tends to overestimate the volume by including spill-out.	Procedure: T, such that the volume from a threshold matches the volume used in the BSL measurement.

B) Segmentation of multimodality imaging:

In the case of Fuzzy C-Means (FCM) multimodality segmentation, a fuzzy membership function and the cluster center c_k^n are updated according to

$$u_{ik}^n = \frac{\|x_i - c_k^n\|}{\sum_{k=1}^{K}\|x_i - c_k^n\|^{-2}},\ c_k^{n+1} = \frac{\sum_{i=1}^{N}\left(u_{ik}^n\right)^b x_i}{\sum_{i=1}^{N}\left(u_{ik}^n\right)^b} \qquad \text{(A4)}$$

where u_{ik}^n is the fuzzy membership probability that image pixel x_i belongs to cluster k at iteration n, and c_k^n is the updated cluster center intensity.

In the case of Multi Valued Level Sets (MVLS), the objective functional for *N* imaging modalities could be presented as:

$$\inf_c J(C, c^+, c^-) \propto \frac{1}{N}\sum_m \Big(\lambda_m^+ \int_\Omega |I_m(x) - c_m^+|^2 H\big(\phi(x)\big)\,dx + \lambda_m^- \int_\Omega |I_m(x) - c_m^-|^2 \Big(1 - H\big(\phi(x)\big)\Big)\,dx \Big) \quad ,$$

(A5)

where $I_m(x)$ is the intensity from imaging modality m at image location x, ϕ is the level set function. $c_m^+(c_m^-)$ corresponds to the pixel intensity mean values inside (outside) of the contour C_m. $H(\cdot)$ is the Heaviside function and $(\lambda_m^+, \lambda_m^-)$ are user-defined parameter pairs providing relative importance weights for each of the imaging modalities *m.* The target boundary is defined at the zero level $(\phi(C) = 0)$ and the integrals are over the space Ω of each image type.

Appendix II. Uptake Normalization and Threshold Parameter Estimation

Uptake normalization:

Preprocessing the uptake data is important for inter-patient comparisons and for defining a segmentation scheme. The most common uptake preprocessing is the conversion to SUV. The use of it or something similar is essential to making the selection of a threshold activity independent and applicable across patients and institutions[47]. SUV itself comes in many flavors, with normalization being carried out with respect to total body mass, lean body mass, body surface area, etc. Beyond SUV, it has been advocated normalizing patient data to the aortic arch (RTOG 1106)[263, 264] or to mean liver uptake. For inter-patient comparisons these normalizations are likely sufficient, but for segmentation, the tumors themselves may need additional and individual normalization.

For segmentation, further normalizing the intensity within the images or VOI allows for greater consistency between different image sets. Although several segmentation algorithms based on contour detection or region determination through contrast measurements do not require nor benefit from any SUV conversion, such SUV normalization is often used for segmentation. Many algorithms[23, 30, 53, 259, 260, 265] rely on subtracting the background activity from the images. In such cases the segmentation effectively uses a function of this form:

$$\xi_i = \xi\left(I_i, I_{bkg}, I_{ref}\right) = \frac{\left(I_i - I_{bkg}\right)}{\left(I_{ref} - I_{bkg}\right)}\ , \qquad \text{(A6)}$$

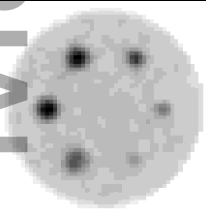
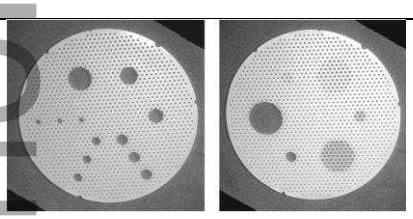
where $\mathrm{I_{ref}}$ is a reference voxel value and $\mathrm{I_{bkg}}$ is the background value, often $\mathrm{I_{ref}} = \mathrm{I_{max}}$. As a further simplification, the image can be normalized solely to $\mathrm{I_{ref}}$ under the assumption that $\mathrm{I_{bkg}}$ is small and does not vary much between images. The maximum contrast results in $\xi_{\max} = 1$ and all voxels for which $\xi_i \geq T$ within the VOI are included in the segmented volume. While alternatives to the equation above exist, the various values that compose it are often similar. As a result, some discussion regarding their choice is useful.

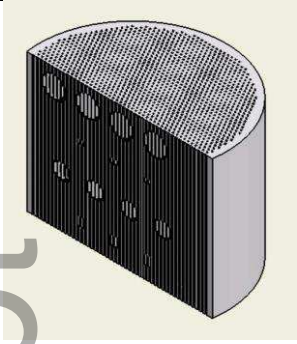
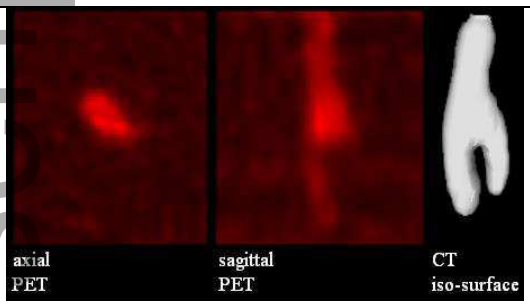
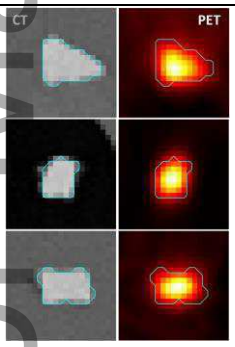
In choosing the voxel values used to define the equation above, some care is necessary to ensure that they are relatively insensitive to the segmentation region and image noise[266]. In the case of the background value, $\mathrm{I_{bkg}}$ is often taken to be the mean intensity over a large region, where the mean is taken from voxels that are far enough from the edge of the object to avoid PVE. Alternatively, when using a histogram approach the mode (the most frequent value) of the voxels' intensity distribution can be chosen instead of the mean[265, 267]. The mode has the advantage of being less susceptible to PVE near the edge of the lesion. In either case, both the mean and the mode are typically well-defined and relatively insensitive to image noise.

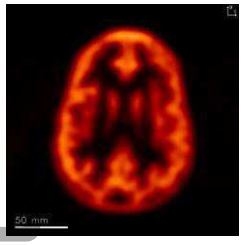
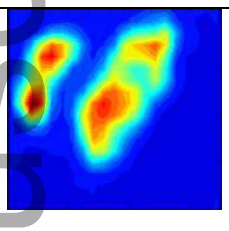
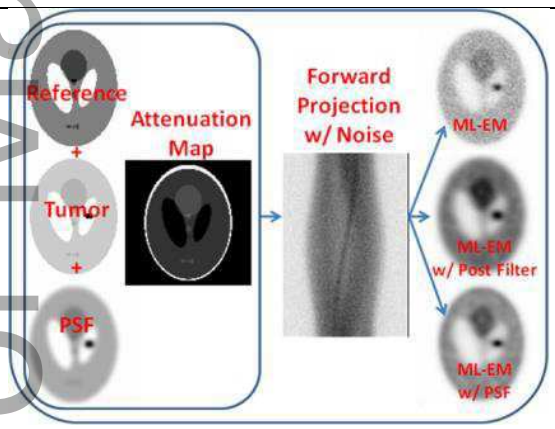
On the other hand, for the reference uptake value, I_{ref}, the choice of the maximum intensity voxel in the object tends to be sensitive to the image noise thus using I_{max} is problematic[23, 30, 53, 260]. Virtually all phantom-based threshold models assume that the activity is uniform in the lesion. Yet the maximum intensity voxel of a region is very sensitive to the size of the object; large objects may exhibit a larger variation of their maximum SUV than small ones. In patient data, this is less clear due to tumor heterogeneity, but it has prompted the use of alternative definitions to maximum SUV for characterizing tumor uptake, such as peak SUV, a grouping of the 10 highest uptake voxels, or similar[268]. One approach described in Nestle, et al.[23] and later expanded on by Schaefer, et al.[30] is to define the reference value as the mean of a region defined by a percent threshold of the maximum voxel (in both papers 70% max SUV). This approach helps reduce the noise associated with a single voxel and provides some stability to the measurement. Using the mean of the segmented volume makes the threshold a function of the segmentation boundary and requires an iterative solution.

APPENDIX III. PET Phantoms

Table A2. A summary of the existing phantoms that are considered as potential candidates to provide data for a future PET-AS evaluation protocol.

Name or type and Reference	Example figure and brief description	Advantages	Disadvantages
	Experimental phantoms		
IEC, NEMA NU 2-2001[173]	 Fig. T.A2.1. Image of a transaxial slice through the center of the spheres	Known ground truth, Variable sphere size, Widely available	Overly simple unrealistic lesion shapes; Uniform background; cold walls
"Porous phantom," Di Filippo, et al. 2004[172]	 Fig. T.A2.2. Perforated discs used for the construction of the hot spheres: multi resolution (left), multi contrast (right). Reprinted with permission from Med. Phys. J..	Adds variable contrast, eliminates cold walls	Simple shapes, Uniform uptake Tedious manufacturing

"Swiss cheese," Hunt, *et al.* 2007[171]	 Fig. T.A2.3. A cut view of the phantom, produced by rapid prototyping. Reprinted with permission from Med. Phys. J..	Similar to the "Porous phantom," above Easier manufacturing	Simple shapes, Uniform uptake
Tumor phantom, Shepherd, et al. 2012[115]	 Fig. T.A2.4. PET images and CT image based iso-surface of the tumor model phantom.	Irregular shapes including branching Based on clinical images Ground truth in PET image space from hybrid CT thresholding	Cold walls of glass compartments Does not recreate heterogeneity (requires internal medium of spatially varying absorbance)
Molecular sieves, Zito, et al. 2012[174]	 Fig. T.A2.5. CT and PET images of 3 zeolites with superimposed ground truths.	Known ground truth; irregular shapes	Uniform lesion uptake; uniform background

Using Radioactive Ink in 2D and 3D printers, Larsson, et al. 2000[175], El-Ali, et al. 2003[176], Miller & Hutchins, 2008[269,177], Berthon et al, 2015[270]	 Fig. T.A2.6. PET image of the human brain phantom produced by incorporating a radioactive dye in rapid prototyping.[269] (Reprinted with permission.)	Can match the irregular shape and non-uniform uptake of real lesions	A specially adapted and calibrated printer is needed
NonU phantom, Kirov, et al. 2011[178]	 Fig. T.A2.7. Relative activity levels of the reference activity distribution obtained by displacing activity by thin foils with irregular cutouts.	Can match the irregular shape and non-uniform uptake of real lesions and background.	Expensive Need to improve accuracy of ground truth
Simulated phantoms			
Simulated PET images from forward projected reference images, 2015 [181]	 Fig. T.A2.8. A schematic showing the generation of PET-like images from reference activity, tumor and attenuation distribution images. These reference images are forward projected, scaled, blurred and noise is added to simulate realistic PET data that is subsequently reconstructed.	Realistic tumor and normal tissues and organs simulations, irregular shapes and heterogen-eous activity distributions Known ground truth	Because geometric projection, rather than photon transport is used, many of the physical aspects of the real image acquisition process are ignored
Monte Carlo simulations Le Maitre, *et al.* 2009 [271]		Realistic tumor and normal tissues and organs simulations,	Computationally expensive; requires extensive up front experience.

Papadimitro-ulas, et al. 2013 [272]	 Fig. T.A2.9.	irregular shapes and heteroge-neous activity distributions; Known ground truth	

APPENDIX IV: Evaluation Criteria for Segmentation Tools

Agreement between sets of voxels

Let *A* and *B* be two volumes lying in space *S* composed of voxels. This space can be the 3D image matrix or the field of view of a scanner. Volumes *A* and *B* are subsets of this space. *A* and *B* are therefore sets of voxels.

The measured volume of *A* is equal to $v|A|$, where *v* is the voxel size expressed in volume units and $|A|$ denotes the cardinality of *A*. The voxel size does not need to be specified and the volume of *A* can be expressed by its cardinality without loss of generality.

The agreement between *A* and *B* basically depends on the cardinality of their intersection $|A \cap B|$. The disagreement is reflected by the two set differences $|A \backslash B|$ and $|B \backslash A|$, i.e., elements of A that are not elements of B and vice versa, respectively. There are therefore two types of errors. These are also absolute errors. The simplest normalization factor is $|A \cup B|^{-1}$. In this case, we have

$$\frac{|A \cap B|}{|A \cup B|} + \frac{|A \backslash B|}{|A \cup B|} + \frac{|B \backslash A|}{|A \cup B|} = 1 \qquad \text{(A8)}$$

The first term is known as the Jaccard similarity coefficient, which varies between 0 and 1. For example, let us assume that A and B are different volumes but $|A| = |B|$ and $|A \cap B| = |A|/2$. Given this then we have $|A \cup B| = 3/2|A|$ and the Jaccard coefficient is equal to 1/3, whereas the overlap actually represents 50% of *A*. This distortion of the intuitive perception of the overlap is addressed by the Dice Similarity Coefficient (DSC), which is defined as

$$\mathrm{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}. \qquad \text{(A9)}$$

The normalization factor is the inverse of the average volume. In the same example as above, the DSC is equal to 1/2 and concurs with the intuition. It can easily be verified that the DSC varies between 0 and 1.

At this point, both *A* and *B* have been considered on the same footing. Let us now define *A* as representing some ground truth and that *B* is defined as an observation of *A* with some inaccuracies. In this case, $|A \backslash B|$ and $|B \backslash A|$ are the numbers of false negatives (FN) and false positives (FP), respectively. This, and the above information, can be written in a confusion matrix as

$$\begin{bmatrix} |A \cap B| & |A \backslash B| \\ |B \backslash A| & |S \backslash (A \cup B)| \end{bmatrix}, \qquad \text{(A10)}$$

where $|S \backslash (A \cup B)|$ is the number of true negatives (TN), which is obviously of little interest, as it depends primarily on the unimportant volume of space *S*, contrary to true positives (TP) in $|A \cap B|$. The most natural normalization factor here is $|A|$. The ratio $|A \cap B|/|A|$ is closely related to the DSC, provided we have $|B| = |A|$. We have here a single equality given by

$$\frac{|A \cap B|}{|A|} + \frac{|A \backslash B|}{|A|} = \frac{|A|}{|A|} = 1 \qquad \text{(A11)}$$

because

$$\frac{|A\cap B|}{|A|}+\frac{|B\backslash A|}{|A|}=\frac{|B|}{|A|} \qquad \text{(A12)}$$

can obviously be larger than 1 as soon as $|B|>|A|$. This issue can be addressed by using the specificity and sensitivity, defined by

$$\text{spec.}=\frac{|TN|}{|TN|+|FP|} \text{ and } \text{sens.}=\frac{|TP|}{|TP|+|FN|}\ . \qquad \text{(A13)}$$

Because the number of true negatives depends on the space volume, the specificity makes little sense and only the sensitivity conveys useful information. The specificity can be replaced with the positive predictive value (PPV) (see Fig.A1 for visual illustration of sensitivity and PPV), defined as

$$\text{PPV}=\frac{|A\cap B|}{|B|}=\frac{|TP|}{|TP|+|FP|}. \qquad \text{(A14)}$$

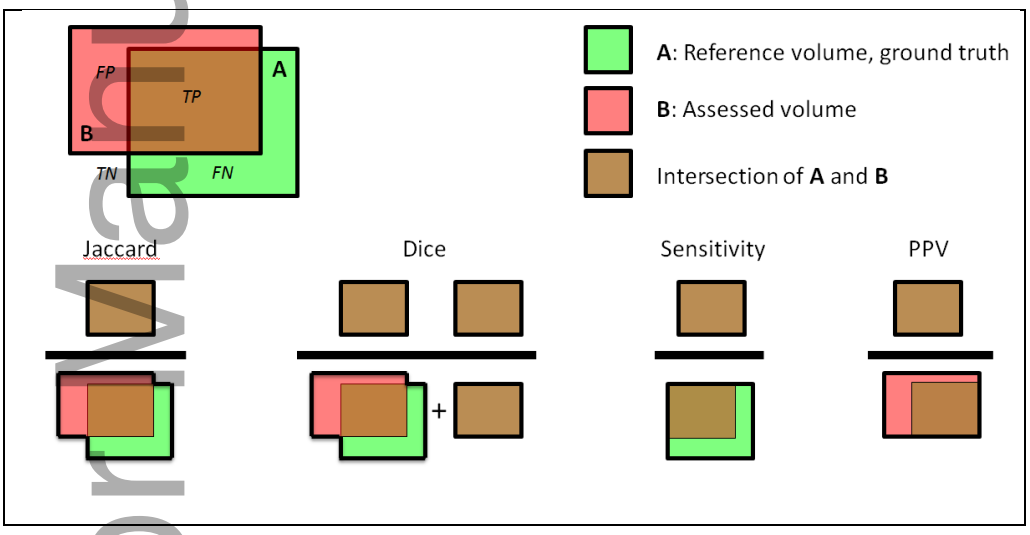


Fig. A1. A graphical illustration of the Jaccard and the Dice similarity coefficients, and of the sensitivity and the positive predictive value (PPV).

All quantities described above assume that set operations can be computed. If only the cardinalities $|A|$and $|B|$ are known, then only the volume difference $|A|-|B|$can be found. The normalization factor can be either $|A|$ or $(|A|+|B|)/2$. The volume difference has two critical shortcomings. First, there is no possibility of distinguishing Type I and Type II errors, apart from the difference sign. Second, the volume difference is overly optimistic: it can be optimal ($|A|-|B|=0$) with actually no overlap ($|A|=|B|$ but $|A\cap B|=0$). The overlap can be approximated with the distance between the centroids (or barycenters) of *A* and *B*, for instance.

Hausdorff distance

If set *A* is rewritten as $\{a_i\}$ and set *B* as $\{b_j\}$, then $\delta(a_i,b_j)$ can denote the distance between voxels a_i and b_j. This distance can be the Euclidean distance from the center of a_i

to the center of b_j. Starting from this voxel-to-voxel distance, the Hausdorff distance is defined as [273]:

$$\mathrm{HD}(A,B) = \max\left\{\max_i \min_j \delta\left(a_i, b_j\right), \max_j \min_i \delta\left(a_i, b_j\right)\right\}. \qquad \text{(A15)}$$

The first term calculates the maximum of the distances from each element of *A* to the closest element of *B*. The second term performs the symmetric computation, with respect to *B* instead of *A*. The maximum of these two quantities is the Hausdorff distance and they can be considered separately to extract information about errors of Types I and II.

The main shortcoming of the Hausdorff distance is its (relative) computational complexity. Notice also that the maximum and minimum operators involved in the definition are very sensitive to image noise. A straightforward variant of the Hausdorff distances addresses this issue by replacing the max operators with averages. This leads to a modified Hausdorff distance [274]:

$$\mathrm{MHD}(A,B) = \frac{1}{|A|}\sum_i \min_j \delta\left(a_i, b_j\right) + \frac{1}{|B|}\sum_j \min_i \delta\left(a_i, b_j\right). \qquad \text{(A16)}$$

The Hausdorff distance is good at reflecting translations between A and B, as well as shape discrepancies. Its interpretation, in terms of volumetric changes, is less obvious. In that sense, it is complementary to the volume difference and to overlap indexes.

The Hausdorff distance can be computed on contours and surfaces as well, instead of sets of voxels. However, in this case the implementation is more specific and requires the user to make some specific choices and/or to adjust additional parameters.

It can be noted that all quantities described above depend on the image matrix or voxel grid. The finer the grid is, the closer the estimated quantities will be to their actual value.

REFERENCES

[1] M. MacManus, U. Nestle, K.E. Rosenzweig, I. Carrio, C. Messa, O. Belohlavek, M. Danna, T. Inoue, E. Deniaud-Alexandre, S. Schipani, N. Watanabe, M. Dondi, B. Jeremic, "Use of PET and PET/CT for radiation therapy planning: IAEA expert report 2006-2007," Radiother Oncol **91**, 85-94 (2009).

[2] C.C. Ling, J. Humm, S. Larson, H. Amols, Z. Fuks, S. Leibel, J.A. Koutcher, "Towards multidimensional radiotherapy (MD-CRT): biological imaging and biological conformality," Int. J. Radiat. Oncol. Biol. Phys. **47**, 551-560 (2000).

[3] S.C. Huang, "Anatomy of SUV. Standardized uptake value," Nuclear medicine and biology **27**, 643-646 (2000).

[4] R. Boellaard, "Standards for PET image acquisition and quantitative data analysis," J Nucl Med **50 Suppl 1**, 11S-20S (2009).

[5] S.S. Gambhir, J. Czernin, J. Schwimmer, D.H. Silverman, R.E. Coleman, M.E. Phelps, "A tabulated summary of the FDG PET literature," J Nucl Med **42**, 1S-93S (2001).

[6] D.E. Heron, R.S. Andrade, S. Beriwal, R.P. Smith, "PET-CT in radiation oncology: the impact on diagnosis, treatment planning, and assessment of treatment response," Am J Clin Oncol **31**, 352-362 (2008).

[7] H. Zaidi, H. Vees, M. Wissmeyer, "Molecular PET/CT imaging-guided radiation therapy treatment planning," Acad Radiol **16**, 1108-1133 (2009).

[8] U. Nestle, W. Weber, M. Hentschel, A.L. Grosu, "Biological imaging in radiation therapy: role of positron emission tomography," Phys Med Biol **54**, R1-R25 (2009).

[9] M.P. Mac Manus, R.J. Hicks, "The role of positron emission tomography/computed tomography in radiation therapy planning for patients with lung cancer," Semin Nucl Med **42**, 308-319 (2012).

[10] M.P. Mac Manus, S. Everitt, M. Bayne, D. Ball, N. Plumridge, D. Binns, A. Herschtal, D. Cruickshank, M. Bressel, R.J. Hicks, "The use of fused PET/CT images for patient selection and radical radiotherapy target volume definition in patients with non-small cell lung cancer: Results of a prospective study with mature survival data," Radiother Oncol **106**, 292-298 (2013).

[11] V. Gregoire, K. Haustermans, X. Geets, S. Roels, M. Lonneux, "PET-based treatment planning in radiotherapy: a new standard?," J Nucl Med **48 Suppl 1**, 68S-77S (2007).

[12] S. Chua, J. Dickson, A.M. Groves, "PET imaging for prediction of response to therapy and outcome in oesophageal carcinoma," Eur J Nucl Med Mol Imaging (2011).

[13] T. Cazaentre, F. Morschhauser, M. Vermandel, N. Betrouni, T. Prangere, M. Steinling, D. Huglo, "Pre-therapy 18F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma," Eur J Nucl Med Mol Imaging **37**, 494-504 (2010).

[14] H.Y. Lee, S.H. Hyun, K.S. Lee, B.T. Kim, J. Kim, Y.M. Shim, M.J. Ahn, T.S. Kim, C.A. Yi, M.J. Chung, "Volume-based parameter of 18)F-FDG PET/CT in malignant pleural mesothelioma: prediction of therapeutic response and prognostic implications," Ann Surg Oncol **17**, 2787-2794 (2010).

[15] I. El Naqa, P. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, W. Thorstad, J.O. Deasy, "Exploring feature-based approaches in PET images for predicting cancer treatment outcomes," Pattern Recognit **42**, 1162-1171 (2009).

[16] M. Vaidya, K.M. Creach, J. Frye, F. Dehdashti, J.D. Bradley, I. El Naqa, "Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer," Radiother Oncol **102**, 239-245 (2012).

[17] M. Soret, S.L. Bacharach, I. Buvat, "Partial-volume effect in PET tumor imaging," J Nucl Med **48**, 932-945 (2007).

[18] A.M. Alessio, P.E. Kinahan, P.M. Cheng, H. Vesselle, J.S. Karp, "PET/CT scanner instrumentation, challenges, and solutions," Radiologic clinics of North America **42**, 1017-1032, vii (2004).
[19] C. Lartizien, P.E. Kinahan, R. Swensson, C. Comtat, M. Lin, V. Villemagne, R. Trebossen, "Evaluating image reconstruction methods for tumor detection in 3-dimensional whole-body PET oncology imaging," J Nucl Med **44**, 276-290 (2003).
[20] D. Visvikis, D. Griffiths, D.C. Costa, J. Bomanji, P.J. Ell, "Clinical evaluation of 2D versus 3D whole-body PET image quality using a dedicated BGO PET scanner," Eur J Nucl Med Mol Imaging **32**, 1050-1056 (2005).
[21] O. Mawlawi, T. Pan, H.A. Macapinlac, "PET/CT imaging techniques, considerations, and artifacts," J Thorac Imaging **21**, 99-110 (2006).
[22] A.S. Kirov, C.R. Schmidtlein, H. Kang, N. Lee, "Rationale, Instrumental Accuracy, and Challenges of PET Quantification for Tumor Segmentation in Radiation Treatment Planning," in *Positron Emission Tomography-Current Clinical and Research Aspects,* edited by C.-H. Hsieh (InTech, 2012).
[23] U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welsch, D. Hellwig, C. Rube, C.M. Kirsch, "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 **46**, 1342-1348 (2005).
[24] S.A. Terezakis, M.A. Hunt, A. Kowalski, P. McCann, C.R. Schmidtlein, A. Reiner, M. Gonen, A.S. Kirov, A.M. Gonzales, H. Schoder, J. Yahalom, "[^{18}F]FDG-positron emission tomography coregistration with computed tomography scans for radiation treatment planning of lymphoma and hematologic malignancies," Int J Radiat Oncol Biol Phys **81**, 615-622 (2011).
[25] R.J. Steenbakkers, J.C. Duppen, I. Fitton, K.E. Deurloo, L.J. Zijp, E.F. Comans, A.L. Uitterhoeve, P.T. Rodrigus, G.W. Kramer, J. Bussink, K. De Jaeger, J.S. Belderbos, P.J. Nowak, M. van Herk, C.R. Rasch, "Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis," Int J Radiat Oncol Biol Phys **64**, 435-448 (2006).
[26] F. Hofheinz, C. Potzsch, L. Oehme, B. Beuthien-Baumann, J. Steinbach, J. Kotzerke, J. van den Hoff, "Automatic volume delineation in oncological PET. Evaluation of a dedicated software tool and comparison with manual delineation in clinical data sets," Nuklearmedizin **51**, 9-16 (2012).
[27] H. Zaidi, "Quantitative analysis in Nuclear Medicine Imaging," (Springer, New York, 2006).
[28] S. Belhassen, H. Zaidi, "A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET," Med Phys **37**, 1309-1324 (2010).
[29] J.A. Lee, "Segmentation of positron emission tomography images: some recommendations for target delineation in radiation oncology," Radiother Oncol **96**, 302-307 (2010).
[30] A. Schaefer, S. Kremp, D. Hellwig, C. Rube, C.M. Kirsch, U. Nestle, "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 **35**, 1989-1999 (2008).
[31] I. El Naqa, D. Yang, A. Apte, D. Khullar, S. Mutic, J. Zheng, J.D. Bradley, P. Grigsby, J.O. Deasy, "Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning," Med Phys **34**, 4738-4749 (2007).
[32] Q. Song, J. Bai, D. Han, S. Bhatia, W. Sun, W. Rockey, J.E. Bayouth, J.M. Buatti, X. Wu, "Optimal co-segmentation of tumor in PET-CT images with context information," IEEE Trans Med Imaging **32**, 1685-1697 (2013).
[33] U. Bagci, J.K. Udupa, N. Mendhiratta, B. Foster, Z. Xu, J. Yao, X. Chen, D.J. Mollura, "Joint segmentation of anatomical and functional images: Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images," Med Image Anal **17**, 929-945 (2013).
[34] D. Markel, H. Zaidi, I. El Naqa, "Novel multimodality segmentation using level sets and Jensen-Rényi divergence. ," Med Phys **40**, 121908 (2013).

[35] M. Hatt, N. Boussion, C. Cheze-Le Rest, D. Visvikis, O. Pradier, "Metabolically active volumes automatic delineation methodologies in PET imaging: Review and perspectives," Cancer Radiother **16**, 70-81 (2012).
[36] B. Foster, U. Bagci, A. Mansoor, Z.Y. Xu, D.J. Mollura, "A review on segmentation of positron emission tomography images," Comput Biol Med **50**, 76-96 (2014).
[37] X. Geets, J.A. Lee, A. Bol, M. Lonneux, V. Gregoire, "A gradient-based method for segmenting FDG-PET images: methodology and validation," Eur J Nucl Med Mol Imaging **34**, 1427-1438 (2007).
[38] E. De Bernardi, E. Faggiano, F. Zito, P. Gerundini, G. Baselli, "Lesion quantification in oncological positron emission tomography: a maximum likelihood partial volume correction strategy," Med Phys **36**, 3040-3049 (2009).
[39] A.O. Boudraa, H. Zaidi, "Image Segmentation Techniques In Nuclear Medicine Imaging,," in *Quantitative analysis in Nuclear Medicine Imaging,* edited by H. Zaidi (Springer, New York, 2006), pp. 308-357.
[40] H. Yu, C. Caldwell, K. Mah, D. Mozeg, "Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning," IEEE Trans Med Imaging **28**, 374-383 (2009).
[41] J.K. Udupa, V.R. Leblanc, Y. Zhuge, C. Imielinska, H. Schmidt, L.M. Currie, B.E. Hirsch, J. Woodburn, "A framework for evaluating image segmentation algorithms," Comput Med Imaging Graph **30**, 75-87 (2006).
[42] P. Tylski, S. Stute, N. Grotus, K. Doyeux, S. Hapdey, I. Gardin, B. Vanderlinden, I. Buvat, "Comparative assessment of methods for estimating tumor volume and standardized uptake value in (18)F-FDG PET," J Nucl Med **51**, 268-276 (2010).
[43] M. Hatt, C. Cheze Le Rest, N. Albarghach, O. Pradier, D. Visvikis, "PET functional volume delineation: a robustness and repeatability study," Eur J Nucl Med Mol Imaging **38**, 663-672 (2011).
[44] B. Berthon, C. Marshall, A. Edwards, M. Evans, E. Spezi, "Influence of cold walls on PET image quantification and volume segmentation: a phantom study," Med Phys **40**, 082505 (2013).
[45] B. Berthon, C. Marshall, M. Evans, E. Spezi, "Evaluation of advanced automatic PET segmentation methods using nonspherical thin-wall inserts," Med Phys **41**, 022502 (2014).
[46] K.J. Biehl, F.M. Kong, F. Dehdashti, J.Y. Jin, S. Mutic, I. El Naqa, B.A. Siegel, J.D. Bradley, "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 **47**, 1808-1812 (2006).
[47] J.A. van Dalen, A.L. Hoffmann, V. Dicken, W.V. Vogel, B. Wiering, T.J. Ruers, N. Karssemeijer, W.J. Oyen, "A novel iterative method for lesion delineation and volumetric quantification with FDG PET," Nucl Med Commun **28**, 485-493 (2007).
[48] Y.E. Erdi, O. Mawlawi, S.M. Larson, M. Imbriaco, H. Yeung, R. Finn, J.L. Humm, "Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding," Cancer **80**, 2505-2509 (1997).
[49] A.C. Paulino, P.A. Johnstone, "FDG-PET in radiotherapy treatment planning: Pandora's box?," Int J Radiat Oncol Biol Phys **59**, 4-5 (2004).
[50] K.J. Biehl, F. Kong, F. Dehdashti, J. Jin, S. Mutic, I. El Naqa, B. Siegel, J. Bradley, "FDG-PET definition of gross tumor volume for radiotherapy of non-small-cell lung cancer: Is a single SUV threshold approach appropriate?," J Nucl Med **47**, 1808-1812 (2006).
[51] W. Jentzen, L. Freudenberg, E.G. Eising, M. Heinze, W. Brandau, A. Bockisch, "Segmentation of PET volumes by iterative image thresholding," J Nucl Med **48**, 108-114 (2007).
[52] S.A. Nehmeh, H. El-Zeftawy, C. Greco, J. Schwartz, Y.E. Erdi, A. Kirov, C.R. Schmidtlein, A.B. Gyau, S.M. Larson, J.L. Humm, "An iterative technique to segment PET lesions using a Monte Carlo based mathematical model," Med Phys **36**, 4803-4809 (2009).

[53] Q.C. Black, I.S. Grills, L.L. Kestin, C.Y. Wong, J.W. Wong, A.A. Martinez, D. Yan, "Defining a radiotherapy target with positron emission tomography," Int J Radiat Oncol Biol Phys **60**, 1272-1282 (2004).
[54] H. Zaidi, I. El Naqa, "PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques," Eur J Nucl Med Mol Imaging **37**, 2165-2187 (2010).
[55] H. Li, W.L. Thorstad, K.J. Biehl, R. Laforest, Y. Su, K.I. Shoghi, E.D. Donnelly, D.A. Low, W. Lu, "A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours," Med Phys **35**, 3711-3721 (2008).
[56] M. Wanet, J.A. Lee, B. Weynand, M. De Bast, A. Poncelet, V. Lacroix, E. Coche, V. Gregoire, X. Geets, "Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: A comparison with threshold-based approaches, CT and surgical specimens," Radiother Oncol **98**, 117-125 (2011).
[57] R. Adams, L. Bischof, "Seeded region growing," IEEE Trans Pattern Anal Mach Intell **16**, 641-647 (1994).
[58] T. Pavlidis, Y.T. Liow, "Integrating region growing and edge-detection," IEEE Trans. Pattern Anal. Mach. Intell. **12**, 225-233 (1990).
[59] L. Ibanez, W. Schroeder, L. Ng, J. Cates, *The ITK Software Guide*. (Kitware, 2005).
[60] E. Day, J. Betler, D. Parda, B. Reitz, A. Kirichenko, S. Mohammadi, M. Miften, "A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients," Med Phys **36**, 4349-4358 (2009).
[61] F. Hofheinz, J. Langner, J. Petr, B. Beuthien-Baumann, J. Steinbach, J. Kotzerke, J. van den Hoff, "An automatic method for accurate volume delineation of heterogeneous tumors in PET," Med Phys **40**, 082503 (2013).
[62] P. Tylski, G. Bonniaud, E. Decenciere, J. Stawiaski, J. Coulot, D. Lefkopoulos, M. Ricard, "18F-FDG PET images segmentation using morphological watershed: a phantom study," 2006 IEEE Nuclear Science Symposium Conference p. 2063.
[63] M. Werner-Wasik, A.D. Nelson, W. Choi, Y. Arai, P.F. Faulhaber, P. Kang, F.D. Almeida, Y. Xiao, N. Ohri, K.D. Brockway, J.W. Piper, A.S. Nelson, "What is the best way to contour lung tumors on PET scans? Multiobserver validation of a gradient-based method using a NSCLC digital PET phantom," Int J Radiat Oncol Biol Phys **82**, 1164-1171 (2012).
[64] I. El Naqa, "Radiotherapy Informatics: Targeted Control," Enterprise Imaging & Therapeutic Radiology Management **18**, 39-42 (2008).
[65] A.S. Dewalle-Vignion, N. Betrouni, R. Lopes, D. Huglo, S. Stute, M. Vermandel, "A New Method for Volume Segmentation of PET Images, Based on Possibility Theory," IEEE Trans Med Imaging **30**, 409-423 (2011).
[66] M. Aristophanous, B.C. Penney, M.K. Martel, C.A. Pelizzari, "A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography," Med Phys **34**, 4223-4235 (2007).
[67] D.W. Montgomery, A. Amira, H. Zaidi, "Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model," Med Phys **34**, 722-736 (2007).
[68] M. Hatt, F. Lamare, N. Boussion, A. Turzo, C. Collet, F. Salzenstein, C. Roux, P. Jarritt, K. Carson, C. Cheze-Le Rest, D. Visvikis, "Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET," Phys Med Biol **52**, 3467-3491 (2007).
[69] M. Hatt, C. Cheze le Rest, A. Turzo, C. Roux, D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," IEEE Trans Med Imaging **28**, 881-893 (2009).
[70] M. Hatt, C. Cheze le Rest, P. Descourt, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, O. Pradier, D. Visvikis, "Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications," Int J Radiat Oncol Biol Phys **77**, 301-308 (2010).

[71] D. Markel, C. Caldwell, H. Alasti, H. Soliman, Y. Ung, J. Lee, A. Sun, "Automatic Segmentation of Lung Carcinoma Using 3D Texture Features in 18-FDG PET/CT," International journal of molecular imaging **2013**, 980769 (2013).
[72] M.S. Sharif, M. Abbod, A. Amira, H. Zaidi, "Artificial Neural Network-Based System for PET Volume Segmentation," Int J Biomed Imaging **ID 105610** , 11 pages (2010).
[73] M.S. Sharif, M. Abbod, A. Amira, H. Zaidi, "Artificial neural network-statistical approach for PET volume analysis and classification," Advances in Fuzzy Systems **ID 327861**, 10 pages (2012).
[74] T.B. Sebastian, R.M. Manjeshwar, T.J. Akhurst, J.V. Miller, "Objective PET lesion segmentation using a spherical mean shift algorithm," Lect Notes Comput Sc **4191**, 782-789 (2006).
[75] M.H. Janssen, H.J. Aerts, M.C. Ollers, G. Bosmans, J.A. Lee, J. Buijsen, D. De Ruysscher, P. Lambin, G. Lammering, A.L. Dekker, "Tumor delineation based on time-activity curve differences assessed with dynamic fluorodeoxyglucose positron emission tomography-computed tomography in rectal cancer patients," Int J Radiat Oncol Biol Phys **73**, 456-465 (2009).
[76] E. De Bernardi, F. Fiorani Gallotta, C. Gianoli, F. Zito, P. Gerundini, G. Baselli, "ML segmentation strategies for object interference compensation in FDG-PET lesion quantification," Methods of information in medicine **49**, 537-541 (2010).
[77] D.P. Onoma, S. Ruan, S. Thureau, L. Nkhali, R. Modzelewski, G.A. Monnehan, P. Vera, I. Gardin, "Segmentation of heterogeneous or small FDG PET positive tissue based on a 3D-locally adaptive random walk algorithm," Comput Med Imaging Graph **38**, 753-763 (2014).
[78] M. Abdoli, R.A. Dierckx, H. Zaidi, "Contourlet-based active contour model for PET image segmentation," Med Phys **40**, 082507 (2013).
[79] W. Mu, Z. Chen, W. Shen, F. Yang, Y. Liang, R. Dai, N. Wu, J. Tian, "A Segmentation Algorithm for Quantitative Analysis of Heterogeneous Tumors of the Cervix with 18F-FDG PET/CT," IEEE Trans Biomed Eng (2015).
[80] H. Cui, X. Wang, J. Zhou, S. Eberl, Y. Yin, D. Feng, M. Fulham, "Topology polymorphism graph for lung tumor segmentation in PET-CT images," Phys Med Biol **60**, 4893-4914 (2015).
[81] J. Lapuyade-Lahorgue, D. Visvikis, O. Pradier, C. Cheze Le Rest, M. Hatt, "SPEQTACLE: An automated generalized fuzzy C-means algorithm for tumor delineation in PET," Med Phys **42**, 5720-5734 (2015).
[82] S. Devic, H. Mohammed, N. Tomic, S. Aldelaijan, F. De Blois, J. Seuntjens, S. Lehnert, S. Faria, "FDG-PET-based differential uptake volume histograms: a possible approach towards definition of biological target volumes," Br J Radiol **89**, 20150388 (2016).
[83] B. Berthon, C. Marshall, M. Evans, E. Spezi, "ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography," Phys Med Biol **61**, 4855-4869 (2016).
[84] A. Schaefer, M. Vermandel, C. Baillet, A.S. Dewalle-Vignion, R. Modzelewski, P. Vera, L. Massoptier, C. Parcq, D. Gibon, T. Fechter, U. Nemer, I. Gardin, U. Nestle, "Impact of consensus contours from multiple PET segmentation methods on the accuracy of functional volume delineation," Eur J Nucl Med Mol Imaging **43**, 911-924 (2016).
[85] W. Pieczynski, "Modèles de Markov en traitement d'images," Trait Signal **20**, 255-278 (2003).
[86] Y. Delignon, A. Marzouki, W. Pieczynski, "Estimation of generalized mixtures and its application in image segmentation," IEEE Trans Image Process **6**, 1364-1375 (1997).
[87] H. Caillol, W. Pieczynski, A. Hillion, "Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation," IEEE Trans Image Process **6**, 425-440 (1997).
[88] F. Salzenstein, C. Collet, S. Lecam, M. Hatt, "Non-stationary fuzzy Markov chain," Pattern Recognition Letters **28**, 2201-2008 (2007).

[89] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern classification*, 2nd ed. (Wiley, New York, 2001).
[90] A.K. Jain, M.N. Murty, P.J. Flynn, "Data clustering: a review," ACM Comput. Surv. **31**, 264-323 (1999).
[91] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," Nature **521**, 436-444 (2015).
[92] M.R. Avendi, A. Kheradvar, H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," Med Image Anal **30**, 108-119 (2016).
[93] K.H. Cha, L. Hadjiiski, R.K. Samala, H.P. Chan, E.M. Caoili, R.H. Cohan, "Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets," Med Phys **43**, 1882 (2016).
[94] A. Kerhet, C. Small, H. Quon, T. Riauka, L. Schrader, R. Greiner, D. Yee, A. McEwan, W. Roa, "Application of machine learning methodology for PET-based definition of lung cancer," Current oncology **17**, 41-47 (2010).
[95] C. Lian, S. Ruan, T. Denoeux, F. Jardin, P. Vera, "Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction," Med Image Anal **32**, 257-268 (2016).
[96] M. Hatt, F. Tixier, L. Pierce, P.E. Kinahan, C.C. Le Rest, D. Visvikis, "Characterization of PET/CT images using texture analysis: the past, the present... any future?," Eur J Nucl Med Mol Imaging **44**, 151-165 (2017).
[97] A. Materka, M. Strzelecki, "Texture Analysis Methods – A Review," in *COST B11* (Technical University of Lodz, Institute of Electronics, Brussels, 1998).
[98] A.K. Jain, F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters," Pattern Recogn **24**, 1167-1186 (1991).
[99] S. Arivazhagan, L. Ganesan, "Texture classification using wavelet transform," Pattern Recognition Letters **24**, 1513-1521 (2003).
[100] G.P. Stachowiak, P. Podsiadlo, G.W. Stachowiak, "A comparison of texture feature extraction methods for machine condition monitoring and failure analysis," Tribol Lett **20**, 133-147 (2005).
[101] R.M. Haralick, K. Shanmugam, "Textural features for image classification," IEEE Trans. Syst., Man, Cybern.,**3**, 610-621 (1973).
[102] M. Amadasun, R. King, "Textural features corresponding to textural properties," IEEE Trans. Syst., Man, Cybern., **19**, 1264-1274 (1989).
[103] M. Galloway, "Texture analysis using grey level run lengths," Comput Vision Graph **4**, 172-179 (1975).
[104] S. Mohamed, A. Youssef, E. El-Saadany, M.M. Salama, "Artificial life feature selection techniques for prostrate cancer diagnosis using TRUS images," Lect Notes Comput Sc, 903-913 (2005).
[105] B.J. Woods, B.D. Clymer, T. Kurc, J.T. Heverhagen, R. Stevens, A. Orsdemir, O. Bulan, M.V. Knopp, "Malignant-lesion segmentation using 4D co-occurrence texture analysis applied to dynamic contrast-enhanced magnetic resonance breast image data," J Magn Reson Imaging **25**, 495-501 (2007).
[106] M.F. McNitt-Gray, E.M. Har, N. Wyckoff, J.W. Sayre, J.G. Goldin, D.R. Aberle, "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results," Med Phys **26**, 880-888 (1999).
[107] A.C. Silva, A.C. Paiva, P.C.P. Carvalho, M. Gattass, "Semivariogram and SGLDM methods comparison for the diagnosis of solitary lung nodule," Pattern Recognition and Image Analysis, Pt 2, Proceedings **3523**, 479-486 (2005).
[108] R. Uppaluri, E.A. Hoffman, M. Sonka, P.G. Hartley, G.W. Hunninghake, G. McLennan, "Computer recognition of regional lung disease patterns," Am J Resp Crit Care **160**, 648-654 (1999).
[109] H.U. Kauczor, K. Heitmann, C.P. Heussel, D. Marwede, T. Uthmann, M. Thelen, "Automatic detection and quantification of ground-glass opacities on high-resolution CT using multiple neural networks: Comparison with a density mask," Am J Roentgenol **175**, 1329-1334 (2000).

[110] F. Chabat, G.Z. Yang, D.M. Hansell, "Obstructive lung diseases: Texture classification for differentiation at CT," Radiology **228**, 871-877 (2003).
[111] B. Ganeshan, S. Abaleke, R.C.D. Young, C.R. Chatwin, K.A. Miles, "Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage," Cancer Imaging **10**, 137-143 (2010).
[112] B.J. Pichler, A. Kolb, T. Nagele, H.P. Schlemmer, "PET/MRI: paving the way for the next generation of clinical multimodality imaging applications," J Nucl Med **51**, 333-336 (2010).
[113] H. Zaidi, A. Del Guerra, "An outlook on future design of hybrid PET/MRI systems," Med Phys **38**, 5667-5689 (2011).
[114] H. Yu, C. Caldwell, K. Mah, I. Poon, J. Balogh, R. MacKenzie, N. Khaouam, R. Tirona, "Automated Radiation Targeting in Head-and-Neck Cancer Using Region-Based Texture Analysis of PET and CT Images," Int J Radiat Oncol Biol Phys **75**, 618-625 (2009).
[115] T. Shepherd, M. Teras, R.R. Beichel, R. Boellaard, M. Bruynooghe, V. Dicken, M.J. Gooding, P.J. Julyan, J.A. Lee, S. Lefevre, "Comparative study with new accuracy metrics for target volume contouring in PET image guided radiation therapy," IEEE Trans Med Imaging **31**, 2006-2024 (2012).
[116] S.K. Warfield, K.H. Zou, W.M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," IEEE Trans Med Imaging **23**, 903-921 (2004).
[117] R.J. McGurk, J. Bowsher, J.A. Lee, S.K. Das, "Combining multiple FDG-PET radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods," Med Phys **40**, 042501 (2013).
[118] D.L. Pham, C. Xu, J.L. Prince, "Current methods in medical image segmentation," Annu Rev Biomed Eng **2**, 315-337 (2000).
[119] V. Bettinardi, I. Castiglioni, E. De Bernardi, M.C. Gilardi, "PET quantification: strategies for partial volume correction," Clinical and Translational Imaging, **2** 199-218 (2014).
[120] L.B. Lucy, "An iterative technique for the rectification of observed distributions," Astron J **79**, 745 (1974).
[121] W.H. Richardson, "Bayesian-Based Iterative Method of Image Restoration," Journal of the Optical Society of America, vol. 62, issue 1, 55 (1972).
[122] A.S. Kirov, J.Z. Piao, C.R. Schmidtlein, "Partial volume effect correction in PET using regularized iterative deconvolution with variance control based on local topology," Phys Med Biol **53**, 2577-2591 (2008).
[123] N. Boussion, C. Cheze Le Rest, M. Hatt, D. Visvikis, "Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging," Eur J Nucl Med Mol Imaging **36**, 1064-1075 (2009).
[124] D.L. Barbee, R.T. Flynn, J.E. Holden, R.J. Nickles, R. Jeraj, "A method for partial volume correction of PET-imaged tumor heterogeneity using expectation maximization with a spatially varying point spread function," Phys Med Biol **55**, 221-236 (2010).
[125] A.M. Alessio, C.W. Stearns, S. Tong, S.G. Ross, S. Kohlmyer, A. Ganin, P.E. Kinahan, "Application and evaluation of a measured spatially variant system model for PET image reconstruction," IEEE Trans Med Imaging **29**, 938-949 (2010).
[126] B.W. Jakoby, Y. Bercier, C.C. Watson, B. Bendriem, D.W. Townsend, "Performance Characteristics of a New LSO PET/CT Scanner With Extended Axial Field-of-View and PSF Reconstruction," IEEE Transactions on Nuclear Science **56**, 633-639 (2009).
[127] E. De Bernardi, M. Mazzoli, F. Zito, G. Baselli, "Resolution recovery in PET during AWOSEM reconstruction: a performance evaluation study," IEEE Trans Nucl Sci **54**, 1626-1638 (2007).
[128] B.K. Teo, Y. Seo, S.L. Bacharach, J.A. Carrasquillo, S.K. Libutti, H. Shukla, B.H. Hasegawa, R.A. Hawkins, B.L. Franc, "Partial-volume correction in PET: validation of

an iterative postreconstruction method with phantom and patient data," J Nucl Med **48**, 802-810 (2007).
129 N. Boussion, M. Hatt, D. Visvikis, "Partial volume correction in PET based on functional volumes," J Nucl Med **49**, 388P (2008).
130 C.H. Chen, R.F. Muzic, Jr., A.D. Nelson, L.P. Adler, "Simultaneous recovery of size and radioactivity concentration of small spheroids with PET data," J Nucl Med **40**, 118-130 (1999).
131 E. De Bernardi, C. Soffientini, F. Zito, G. Baselli, "Joint Segmentation and Quantification of Oncological Lesions in PET/CT: Preliminary Evaluation on a Zeolite Phantom," IEEE NSS MIC 2012, October 29 - November 3, 2012, Anaheim, California, pp. 3306-3310 (2012).
132 A.D. King, "Multimodality imaging of head and neck cancer," Cancer Imaging **7 Spec No A**, S37-46 (2007).
133 M.T. Munley, L.B. Marks, C. Scarfone, G.S. Sibley, E.F. Patz Jr, T.G. Turkington, R.J. Jaszczak, D.R. Gilland, M.S. Anscher, R.E. Coleman, "Multimodality nuclear medicine imaging in three-dimensional radiation treatment planning for lung cancer: challenges and prospects," Lung Cancer **23**, 105-114 (1999).
134 R. Chen, J.J. Parry, W.J. Akers, M.Y. Berezin, I.M. El Naqa, S. Achilefu, W.B. Edwards, B.E. Rogers, "Multimodality imaging of gene transfer with a receptor-based reporter gene," J Nucl Med **51**, 1456-1463 (2010).
135 E.M. DeFeo, C.-L. Wu, W.S. McDougal, L.L. Cheng, "A decade in prostate cancer: from NMR to metabolomics," Nat Rev Urol **8**, 301-311 (2011).
136 A.R. Hsu, W. Cai, A. Veeravagu, K.A. Mohamedali, K. Chen, S. Kim, H. Vogel, L.C. Hou, V. Tse, M.G. Rosenblum, X. Chen, "Multimodality molecular imaging of glioblastoma growth inhibition with vasculature-targeting fusion toxin VEGF121/rGel," J Nucl Med **48**, 445-454 (2007).
137 W.L. Smith, C. Lewis, G. Bauman, G. Rodrigues, D. D'Souza, R. Ash, D. Ho, V. Venkatesan, D. Downey, A. Fenster, "Prostate volume contouring: a 3D analysis of segmentation using 3DTRUS, CT, and MR," Int J Radiat Oncol Biol Phys **67**, 1238-1247 (2007).
138 J. Buijsen, J. van den Bogaard, H. van der Weide, S. Engelsman, R. van Stiphout, M. Janssen, G. Beets, R. Beets-Tan, P. Lambin, G. Lammering, "FDG-PET-CT reduces the interobserver variability in rectal tumor delineation," Radiother Oncol **102**, 371-376 (2012).
139 A. van Baardwijk, G. Bosmans, L. Boersma, J. Buijsen, S. Wanders, M. Hochstenbag, R.J. van Suylen, A. Dekker, C. Dehing-Oberije, R. Houben, S.M. Bentzen, M. van Kroonenburgh, P. Lambin, D. De Ruysscher, "PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes," Int J Radiat Oncol Biol Phys **68**, 771-778 (2007).
140 H. Metwally, F. Courbon, I. David, T. Filleron, A. Blouet, M. Rives, F. Izar, S. Zerdoud, G. Plat, J. Vial, A. Robert, A. Laprie, "Coregistration of prechemotherapy PET-CT for planning pediatric Hodgkin's disease radiotherapy significantly diminishes interobserver variability of clinical target volume definition," Int J Radiat Oncol Biol Phys **80**, 793-799 (2011).
141 C.M. Anderson, W. Sun, J.M. Buatti, J.E. Maley, B. Policeni, S.L. Mott, J.E. Bayouth, "Interobserver and intermodality variability in GTV delineation on simulation CT, FDG-PET, and MR Images of Head and Neck Cancer," Jacobs journal of radiation oncology **1**, 006 (2014).
142 Y. Zheng, X. Sun, J. Wang, L. Zhang, X. Di, Y. Xu, "FDG-PET/CT imaging for tumor staging and definition of tumor volumes in radiation treatment planning in non-small cell lung cancer," Oncology letters **7**, 1015-1020 (2014).
143 A. Sebbahi, A. Herment, A. de Cesare, E. Mousseaux, "Multimodality cardiovascular image segmentation using a deformable contour model," Comput Med Imag Grap **21**, 79-89 (1997).

[144] J. Zheng, I. El Naqa, F.E. Rowold, T.K. Pilgram, P.K. Woodard, J.E. Saffitz, D. Tang, "Quantitative assessment of coronary artery plaque vulnerability by high-resolution magnetic resonance imaging and computational biomechanics: a pilot study ex vivo," Magn Reson Med **54**, 1360-1368 (2005).

[145] D. Yang, J. Zheng, A. Nofal , Y. Wu, J. Deasy, I. El Naqa, "Techniques and software tool for 3D multimodality medical image segmentation," Journal of radiation oncology informatics **1**, 1-21 (2009).

[146] T.F. Chan, B.Y. Sandberg, L.A. Vese, "Active Contours without Edges for Vector-Valued Images," Journal of Visual Communication and Image Representation **11**, 130-141 (2000).

[147] J. Shah, "Curve evolution and segmentation functionals: application to color images," in *International Conference on Image Processing Proceedings*, *Vol. 1*, pp. 461-464. (1996)

[148] S.E. Fogh, A. Farach, C. Intenzo, R. Axelrod, P. McCue, A. Harper, A. Nelson, M. Werner-Wasik, "Pathologic Correlation of PET-CT Based Auto Contouring for Radiation Planning in Lung Cancer," Int J Radiat Oncol Biol Phys **78**, S202-S203 (2010).

[149] J.F. Daisne, M. Sibomana, A. Bol, T. Doumont, M. Lonneux, V. Gregoire, "Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms," Radiother Oncol **69**, 247-250 (2003).

[150] H. Zaidi, M. Abdoli, C.L. Fuentes, I.M. El Naqa, "Comparative methods for PET image segmentation in pharyngolaryngeal squamous cell carcinoma," Eur J Nucl Med Mol Imaging **39**, 881-891 (2012).

[151] J.F. Daisne, T. Duprez, B. Weynand, M. Lonneux, M. Hamoir, H. Reychler, V. Gregoire, "Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen," Radiology **233**, 93-100 (2004).

[152] M. Hatt, C. Cheze-le Rest, A. van Baardwijk, P. Lambin, O. Pradier, D. Visvikis, "Impact of tumor size and tracer uptake heterogeneity in (18)F-FDG PET and CT non-small cell lung cancer tumor delineation," J Nucl Med **52**, 1690-1697 (2011).

[153] M. Hatt, A.L. Maitre, D. Wallach, H. Fayad, D. Visvikis, "Comparison of different methods of incorporating respiratory motion for lung cancer tumor volume delineation on PET images: a simulation study," Phys Med Biol **57**, 7409-7430 (2012).

[154] B. Berthon, C. Marshall, M. Evans, E. Spezi, "Implementation and optimization of automatic 18F-FDG PET segmentation methods," Eur J Nucl Med Mol Imaging **39(Suppl 2)**, S385 (2013).

[155] M. Ollers, G. Bosmans, A. van Baardwijk, A. Dekker, P. Lambin, J. Teule, W. Thimister, A. Rhamy, D. De Ruysscher, "The integration of PET-CT scans from different hospitals into radiotherapy treatment planning," Radiother Oncol **87**, 142-146 (2008).

[156] B. Knausl, A. Hirtl, G. Dobrozemsky, H. Bergmann, K. Kletter, R. Dudczak, D. Georg, "PET based volume segmentation with emphasis on the iterative TrueX algorithm," Zeitschrift fur medizinische Physik **22**, 29-39 (2012).

[157] A. Schaefer, U. Nestle, S. Kremp, D. Hellwig, A. Grgic, H.G. Buchholz, W. Mischke, C. Gromoll, P. Dennert, M. Plotkin, S. Senftleben, D. Thorwarth, M. Tosch, A. Wahl, H. Wengenmair, C. Rube, C.M. Kirsch, "Multi-centre calibration of an adaptive thresholding method for PET-based delineation of tumour volumes in radiotherapy planning of lung cancer," Nuklearmed-Nucl Med **51**, 101-110 (2012).

[158] D.A. Schinagl, W.V. Vogel, A.L. Hoffmann, J.A. van Dalen, W.J. Oyen, J.H. Kaanders, "Comparison of five segmentation tools for 18F-fluoro-deoxy-glucose-positron emission tomography-based target volume definition in head and neck cancer," Int J Radiat Oncol Biol Phys **69**, 1282-1289 (2007).

[159] C. Greco, S.A. Nehmeh, H. Schoder, M. Gonen, B. Raphael, H.E. Stambuk, J.L. Humm, S.M. Larson, N.Y. Lee, "Evaluation of different methods of 18F-FDG-PET

target volume delineation in the radiotherapy of head and neck cancer," Am J Clin Oncol **31**, 439-445 (2008).
160 H. Vees, S. Senthamizhchelvan, R. Miralbell, D.C. Weber, O. Ratib, H. Zaidi, "Assessment of various strategies for 18F-FET PET-guided delineation of target volumes in high-grade glioma patients," Eur J Nucl Med Mol Imaging **36**, 182-193 (2009).
161 S. Belhassen, C.S. Llina Fuentes, A. Dekker, D. De Ruysscher, O. Ratib, H. Zaidi "Comparative methods for 18F-FDG PET-based delineation of target volumes in non-small-cell lung cancer," J Nucl Med, **50** , 27P (2009).
162 A.S. Dewalle-Vignion, N. Yeni, G. Petyt, L. Verscheure, D. Huglo, A. Beron, S. Adib, G. Lion, M. Vermandel, "Evaluation of PET volume segmentation methods: comparisons with expert manual delineations," Nuclear Medicine Communications **33**, 34-42 (2012).
163 A. Lacout, P.Y. Marcy, J. Giron, J. Thariat, "Gradient-PET based delineation may be improved with combined post contrast high resolution CT scan: in regard to Werner-Wasik M et al. (Int J Radiat Oncol Biol Phys 2011 Apr 28)," Int J Radiat Oncol Biol Phys **82**, 496; author reply 496-497 (2012).
164 D.A. Schinagl, P.N. Span, F.J. van den Hoogen, M.A. Merkx, P.J. Slootweg, W.J. Oyen, J.H. Kaanders, "Pathology-based validation of FDG PET segmentation tools for volume assessment of lymph node metastases from head and neck cancer," Eur J Nucl Med Mol Imaging **40**, 1828-1835 (2013).
165 A.S. Kirov, L. Fanchon, "Pathology-validated PET image data sets and their role for PET segmentation," Clinical and Translational Imaging **2**, 253-267 (2014).
166 T.R. Mackie, V. Gregoire, "International Commission on Radiation Units and Measurements (ICRU) Report 83. Prescribing, Recording, and Reporting Photon-Beam Intensity-Modulated Radiation Therapy (IMRT)," **Vol. 10(1)** (2010).
167 B.M. Fischer, M.W.B. Olsen, C.D. Ley, T.L. Klausen, J. Mortensen, L. Hojgaard, P.E.G. Kristjansen, "How few cancer cells can be detected by positron emission tomography? A frequent question addressed by an in vitro study," Eur J Nucl Med Mol Imaging **33**, 697-702 (2006).
168 B. Berthon, E. Spezi, P. Galavis, A.T. Shepherd, A. Apte, M. Hatt, H. Fayad, E. De Bernardi, C. Soffientini, C. R. Schmidtlein, I. El Naqa, R. Jeraj, W. Lu, S. Das, H. Zaidi, O. Mawlawi, D. Visvikis, J. A. Lee, A.S. Kirov, "Towards a standard for the evaluation of PET Auto-Segmentation methods: requirements and implementation," submitted to Med. Phys. J (2016).
169 T. Shepherd, R. Owenius, "Gaussian Process Models of Dynamic PET for Functional Volume Definition in Radiation Oncology," IEEE Trans Med Imaging **31**, 1542-1556 (2012).
170 B. Lelandais, S. Ruan, T. Denoeux, P. Vera, I. Gardin, "Fusion of multi-tracer PET images for dose painting," Med Image Anal **18**, 1247-1259 (2014).
171 D.C. Hunt, H. Easton, C.B. Caldwell, "Design and construction of a quality control phantom for SPECT and PET imaging," Med Phys **36**, 5404-5411 (2009).
172 F.P. DiFilippo, J.P. Price, D.N. Kelsch, R.F. Muzic, Jr., "Porous phantoms for PET and SPECT performance evaluation and quality assurance," Med Phys **31**, 1183-1194 (2004).
173 "NEMA NU 2-2001, Performance Measurements of Positron Emission Tomographs," (National Electrical Manufacturers Association, Rosslyn, VA, 2001).
174 F. Zito, E. De Bernardi, C. Soffientini, C. Canzi, R. Casati, P. Gerundini, G. Baselli, "The use of zeolites to generate PET phantoms for the validation of quantification strategies in oncology," Med Phys, **39**, 5353-5361 (2012).
175 S.A. Larsson, C. Jonsson, M. Pagani, L. Johansson, H. Jacobsson, "A novel phantom design for emission tomography enabling scatter- and attenuation-"free" single-photon emission tomography imaging," Eur J Nucl Med **27**, 131-139 (2000).
176 H. El-Ali, M. Ljungberg, S.E. Strand, J. Palmer, L. Malmgren, J. Nilsson, "Calibration of a radioactive ink-based stack phantom and its applications in nuclear medicine," Cancer Biother Radiopharm **18**, 201-207 (2003).

[177] M. Miller, G. Hutchins, "3D Anatomically accurate phantoms for PET and SPECT imaging," J. Nucl. Med. **49**, 65P (2008).
[178] A.S. Kirov, E. Sculley, C.R. Schmidtlein, W. Siman, B. Kandel, R. Zdenek, R. Schwar, G. Ayzenberg, E. Yorke, H. Schöder, J.L. Humm, H. Amols, "A New Phantom Allowing Realistic Non-Uniform Activity Distributions for PET Quantification, Abstract presented at the 2011 Joint AAPM/COMP meeting," Med. Phys. **38**, 3387 (2011).
[179] H. Zaidi, X.G. Xu, "Computational anthropomorphic models of the human anatomy: The path to realistic Monte Carlo modeling in medical imaging," Annu Rev Biomed Eng **9**, 471-500 (2007).
[180] W. Wang, J.C. Georgi, S.A. Nehmeh, M. Narayanan, T. Paulus, M. Bal, J. O'Donoghue, P.B. Zanzonico, C.R. Schmidtlein, N.Y. Lee, J.L. Humm, "Evaluation of a compartmental model for estimating tumor hypoxia via FMISO dynamic PET imaging," Phys Med Biol **54**, 3083-3099 (2009).
[181] B. Berthon, I. Häggström, A. Apte, B.J. Beattie, A.S. Kirov, J. Humm, C. Marshall, E. Spezi, A. Larsson, C.R. Schmidtlein, "PETSTEP: Generation of Synthetic PET Lesions for Fast Evaluation of Segmentation Methods," Phys Med **31**, 969-980 (2015).
[182] L.A. Shepp, Y. Vardi, "Maximum likelihood reconstruction for emission tomography," IEEE Trans Med Imaging **1**, 113-122 (1982).
[183] E. Asma, S. Ahn, S.G. Ross, A. Chen, R.M. Manjeshwar, "Accurate and consistent lesion quantitation with clinically acceptable penalized likelihood images," Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE, pp. 4062-4066 (2012).
[184] W.P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, B.M. Tsui, "4D XCAT phantom for multimodality imaging research," Med Phys **37**, 4902-4915 (2010).
[185] I.G. Zubal, C.R. Harrell, E.O. Smith, Z. Rattner, G. Gindi, P.B. Hoffer, "Computerized three-dimensional segmented human anatomy," Med Phys **21**, 299-302 (1994).
[186] A. McLennan, A. Reilhac, M. Brady, "SORTEO: Monte Carlo-based simulator with list-mode capabilities," 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 3751-3754 (2009).
[187] S. Jan, D. Benoit, E. Becheva, T. Carlier, F. Cassol, P. Descourt, T. Frisson, L. Grevillot, L. Guigues, L. Maigne, C. Morel, Y. Perrot, N. Rehfeld, D. Sarrut, D.R. Schaart, S. Stute, U. Pietrzyk, D. Visvikis, N. Zahra, I. Buvat, "GATE V6: a major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy," Phys Med Biol **56**, 881-901 (2011).
[188] S. Jan, G. Santin, D. Strul, S. Staelens, K. Assie, D. Autret, S. Avner, R. Barbier, M. Bardies, P.M. Bloomfield, D. Brasse, V. Breton, P. Bruyndonckx, I. Buvat, A.F. Chatziioannou, Y. Choi, Y.H. Chung, C. Comtat, D. Donnarieix, L. Ferrer, S.J. Glick, C.J. Groiselle, D. Guez, P.F. Honore, S. Kerhoas-Cavata, A.S. Kirov, V. Kohli, M. Koole, M. Krieguer, D.J. van der Laan, F. Lamare, G. Largeron, C. Lartizien, D. Lazaro, M.C. Maas, L. Maigne, F. Mayet, F. Melot, C. Merheb, E. Pennacchio, J. Perez, U. Pietrzyk, F.R. Rannou, M. Rey, D.R. Schaart, C.R. Schmidtlein, L. Simon, T.Y. Song, J.M. Vieira, D. Visvikis, R. Van de Walle, E. Wieers, C. Morel, "GATE: a simulation toolkit for PET and SPECT," Phys Med Biol **49**, 4543-4561 (2004).
[189] R. Harrison, S. Gillispie, R. Schmitz, T. Lewellen, "Modeling block detectors in SimSET," J Nucl Med **49 Suppl 1**, 410P (2008).
[190] F. Lamare, A. Turzo, Y. Bizais, C.C. Le Rest, D. Visvikis, "Validation of a Monte Carlo simulation of the Philips Allegro/GEMINI PET systems using GATE," Phys Med Biol **51**, 943-962 (2006).
[191] M. Bayne, R.J. Hicks, S. Everitt, N. Fimmell, D. Ball, J. Reynolds, E. Lau, A. Pitman, R. Ware, M. MacManus, "Reproducibility of "intelligent" contouring of gross tumor volume in non-small-cell lung cancer on PET/CT images using a standardized visual method," Int J Radiat Oncol Biol Phys **77**, 1151-1157 (2010).
[192] S.E. Fogh, J. Kannarkatt, A. Farach, P. McCue, R. Axelrod, A. Nelson, M. Werner-Wasik, "Pathologic correlation of PET-CT based auto contouring for radiation treatment planning in lung cancer," J Thorac Oncol **4**, S528-S529 (2009).

[193] J. van Loon, C. Siedschlag, J. Stroom, H. Blauwgeers, R.J. van Suylen, J. Knegjens, M. Rossi, A. van Baardwijk, L. Boersma, H. Klomp, W. Vogel, S. Burgers, K. Gilhuijs, "Microscopic Disease Extension in Three Dimensions for Non-Small-Cell Lung Cancer: Development of a Prediction Model Using Pathology-Validated Positron Emission Tomography and Computed Tomography Features," Int J Radiat Oncol Biol Phys **82**, 448-456 (2012).

[194] M. Axente, J. He, C.P. Bass, G. Sundaresan, J. Zweit, J.F. Williamson, A. Pugachev, "An alternative approach to histopathological validation of PET imaging for radiation therapy image-guidance: a proof of concept," Radiother Oncol **110**, 309-316 (2014).

[195] L.M. Fanchon, S. Dogan, A.L. Moreira, S.A. Carlin, C.R. Schmidtlein, E. Yorke, A.P. Apte, I.A. Burger, J.C. Durack, J.P. Erinjeri, M. Maybody, H. Schoder, R.H. Siegelbaum, C.T. Sofocleous, J.O. Deasy, S.B. Solomon, J.L. Humm, A.S. Kirov, "Feasibility of in situ, high-resolution correlation of tracer uptake with histopathology by quantitative autoradiography of biopsy specimens obtained under 18F-FDG PET/CT guidance," J Nucl Med **56**, 538-544 (2015).

[196] M.-P. Dubuisson, A.K. Jain, "A modified Hausdorff distance for object matching," Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on, vol. 1, 566-568 (1994).

[197] Kim H, J.I. Monroe, S. Lo, M. Yao, P.M. Harari, M. Machtay, J.W. Sohn, "Quantitative evaluation of image segmentation incorporating medical consideration functions," Med Phys **42**, 3013-3023 (2015).

[198] V. Gregoire, R. Jeraj, J.A. Lee, B. O'Sullivan, "Radiotherapy for head and neck tumours in 2012 and beyond: conformal, tailored, and adaptive?," Lancet Oncol **13**, e292-300 (2012).

[199] A. Skretting, J.F. Evensen, A.M. Londalen, T.V. Bogsrud, O.K. Glomset, K. Eilertsen, "A gel tumour phantom for assessment of the accuracy of manual and automatic delineation of gross tumour volume from FDG-PET/CT," Acta Oncol **52**, 636-644 (2013).

[200] S. David, D. Visvikis, C. Roux, M. Hatt, "Multi-observation PET image analysis for patient follow-up quantitation and therapy assessment," Phys Med Biol **56**, 5771-5788 (2011).

[201] S. David, D. Visvikis, G. Quellec, C.C. Le Rest, P. Fernandez, M. Allard, C. Roux, M. Hatt, "Image Change Detection Using Paradoxical Theory for Patient Follow-Up Quantitation and Therapy Assessment," IEEE Trans Med Imaging **31**, 1743-1753 (2012).

[202] B. Lelandais, I. Gardin, L. Mouchard, P. Vera, S. Ruan, "Segmentation of biological target volumes on multi-tracer PET images based on information fusion for achieving dose painting in radiotherapy," Med Image Comput Comput Assist Interv -MICCAI **15**, 545-552 (2012).

[203] V. Frings, A.J. de Langen, E.F. Smit, F.H. van Velden, O.S. Hoekstra, H. van Tinteren, R. Boellaard, "Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer," J Nucl Med **51**, 1870-1877 (2010).

[204] M. Hatt, C. Cheze-Le Rest, E.O. Aboagye, L.M. Kenny, L. Rosso, F.E. Turkheimer, N.M. Albarghach, J.P. Metges, O. Pradier, D. Visvikis, "Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements," J Nucl Med **51**, 1368-1376 (2010).

[205] A.J. Buckler, R. Boellaard, "Standardization of Quantitative Imaging: The Time Is Right, and F-18-FDG PET/CT Is a Good Place to Start," J Nucl Med **52**, 171-172 (2011).

[206] R.L. Wahl, H. Jacene, Y. Kasamon, M.A. Lodge, "From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors," J Nucl Med **50**, 122s-150s (2009).

[207] C.R. MacFarlane, R. American College of, "ACR accreditation of nuclear medicine and PET imaging departments," J Nucl Med Technol **34**, 18-24 (2006).

[208] H.H. Barrett, "Objective assessment of image quality: effects of quantum noise and object variability," J Opt Soc Am A **7**, 1266-1278 (1990).
[209] H.H. Barrett, C.K. Abbey, E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions," J Opt Soc Am A **15**, 1520-1535 (1998).
[210] H.H. Barrett, J.L. Denny, R.F. Wagner, K.J. Myers, "Objective assessment of image quality. II. Fisher information, Fourier crosstalk, and figures of merit for task performance," J Opt Soc Am A **12**, 834-852 (1995).
[211] H.H. Barrett, M.A. Kupinski, S. Mueller, H.J. Halpern, J.C. Morris, R. Dwyer, "Objective assessment of image quality VI: imaging in radiation therapy," Phys Med Biol **58**, 8197-8213 (2013).
[212] J.A. Fessler, W.L. Rogers, "Spatial resolution properties of penalized-likelihood image reconstruction: Space-invariant tomographs," IEEE T Image Process **5**, 1346-1358 (1996).
[213] H.H. Barrett, D.W. Wilson, B.M.W. Tsui, "Noise Properties of the Em Algorithm .1. Theory," Phys Med Biol **39**, 833-846 (1994).
[214] D.F. Yu, J.A. Fessler, "Edge-preserving tomographic reconstruction with nonlocal regularization," IEEE Trans Med Imaging **21**, 159-173 (2002).
[215] J.A. Fessler, E.P. Ficaro, N.H. Clinthorne, K. Lange, "Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction," IEEE Trans Med Imaging **16**, 166-175 (1997).
[216] A. Krol, S. Li, L. Shen, Y. Xu, "Preconditioned Alternating Projection Algorithms for Maximum a Posteriori ECT Reconstruction," Inverse problems **28**, 115005 (2012).
[217] E. Rapisarda, L. Presotto, E. De Bernardi, M.C. Gilardi, V. Bettinardi, "Optimized Bayes variational regularization prior for 3D PET images," Comput Med Imag Grap **38**, 445-457 (2014).
[218] S. Ahn, S.G. Ross, E. Asma, J. Miao, X. Jin, L.S. Cheng, S.D. Wollenweber, R.M. Manjeshwar, "Quantitative comparison of OSEM and penalized likelihood image reconstruction using relative difference penalties for clinical PET," Phys Med Biol **60**, 5733-5751 (2015).
[219] A.I. Arens, E.G. Troost, B.A. Hoeben, W. Grootjans, J.A. Lee, V. Gregoire, M. Hatt, D. Visvikis, J. Bussink, W.J. Oyen, J.H. Kaanders, E.P. Visser, "Semiautomatic methods for segmentation of the proliferative tumour volume on sequential FLT PET/CT images in head and neck carcinomas and their relation to clinical outcome," Eur J Nucl Med Mol Imaging **41**, 915-924 (2014).
[220] B. Henriques de Figueiredo, C. Zacharatou, S. Galland-Girodet, J. Benech, H. De Clermont-Gallerande, F. Lamare, M. Hatt, L. Digue, E. De Mones del Pujol, P. Fernandez, "Hypoxia imaging with [18F]-FMISO-PET for guided dose escalation with intensity-modulated radiotherapy in head-and-neck cancers," Strahlenther Onkol **191**, 217-224 (2015).
[221] D.A. Low, M. Nystrom, E. Kalinin, P. Parikh, J.F. Dempsey, J.D. Bradley, S. Mutic, S.H. Wahab, T. Islam, G. Christensen, D.G. Politte, B.R. Whiting, "A method for the reconstruction of four-dimensional synchronized CT scans acquired during free breathing," Med Phys **30**, 1254-1263 (2003).
[222] N. Wink, C. Panknin, T.D. Solberg, "Phase versus amplitude sorting of 4D-CT data," J Appl Clin Med Phys **7**, 77-85 (2006).
[223] J.R. Olsen, W. Lu, J.P. Hubenschmidt, M.M. Nystrom, P. Klahr, J.D. Bradley, D.A. Low, P.J. Parikh, "Effect of novel amplitude/phase binning algorithm on commercial four-dimensional computed tomography quality," International journal of radiation oncology, biology, physics **70**, 243-252 (2008).
[224] S.A. Nehmeh, Y.E. Erdi, K.E. Rosenzweig, H. Schoder, S.M. Larson, O.D. Squire, J.L. Humm, "Reduction of respiratory motion artifacts in PET imaging of lung cancer by respiratory correlated dynamic PET: Methodology and comparison with respiratory gated PET," Journal of Nuclear Medicine **44**, 1644-1648 (2003).
[225] F. Qiao, T. Pan, J.W. Clark Jr., O.R. Mawlawi, "A motion-incorporated reconstruction method for gated PET studies," Physics in medicine and biology, 3769 (2006).

[226] C. Pai-Chun Melinda, M. Osama, A.N. Sadek, E.E. Yusuf, A.B. Peter, L. Dershan, M. Radhe, P. Tinsu, "Design of respiration averaged CT for attenuation correction of the PET data from PET/CT," Med Phys **34**, 2039-2047 (2007).
[227] K. Berlinger, O. Sauer, L. Vences, M. Roth, "A simple method for labeling CT images with respiratory states," Med Phys **33**, 3144-3148 (2006).
[228] F. Qiao, T. Pan, J.J.W. Clark, O. Mawlawi, "Joint model of motion and anatomy for PET image reconstruction," Med Phys **34**, 4626-4639 (2007).
[229] M. Dawood, F. Buther, N. Lang, O. Schober, K.P. Schafers, "Respiratory gating in positron emission tomography: A quantitative comparison of different gating schemes," Med Phys **34**, 3067-3076 (2007).
[230] P.P. Bruyant, C.C.L. Rest, A. Turzo, P. Jarritt, K. Carson, D. Visvikis, "A method for synchronizing an external respiratory signal with a list-mode PET acquisition," Med Phys **34**, 4472-4475 (2007).
[231] S.A. Nehmeh, Y.E. Erdi, G.S. Meirelles, O. Squire, S.M. Larson, J.L. Humm, H. Schoder, "Deep-inspiration breath-hold PET/CT of the thorax," J Nucl Med **48**, 22-26 (2007).
[232] W. Sureshbabu, O. Mawlawi, "PET/CT imaging artifacts," J Nucl Med Technol **33**, 156-161; quiz 163-154 (2005).
[233] G. Chang, T. Chang, T. Pan, J.W. Clark, Jr., O.R. Mawlawi, "Implementation of an automated respiratory amplitude gating technique for PET/CT: clinical evaluation," J Nuc Med **51**, 16-24 (2010).
[234] F. Buther, I. Ernst, M. Dawood, P. Kraxner, M. Schafers, O. Schober, K.P. Schafers, "Detection of respiratory tumour motion using intrinsic list mode-driven gating in positron emission tomography," Eur J Nucl Med Mol Imaging **37**, 2315-2327 (2010).
[235] P.J. Schleyer, M.J. O'Doherty, S.F. Barrington, P.K. Marsden, "Retrospective data-driven respiratory gating for PET/CT," Phys Med Biol **54**, 1935-1950 (2009).
[236] A.L. Kesner, C. Kuntner, "A new fast and fully automated software based algorithm for extracting respiratory signal from raw PET data and its comparison to other methods," Med Phys **37**, 5550-5559 (2010).
[237] I. El Naqa, D.A. Low, J.D. Bradley, M. Vicic, J.O. Deasy, "Deblurring of breathing motion artifacts in thoracic PET images by deconvolution methods," Med Phys **33**, 3587-3600 (2006).
[238] P.K. Yalavarthy, D. Low, C. Noel, W. Zhouping, Y. Deshan, A. Apte, J. Bradley, J. Deasy, I. El Naqa, "Current role of PET in oncology: Potentials and challenges in the management of non-small cell lung cancer," 2008 42nd Asilomar Conference on Signals, Systems and Computers, 1067-1071 (2008).
[239] F. Buther, T. Vehren, K.P. Schafers, M. Schafers, "Impact of Data-driven Respiratory Gating in Clinical PET," Radiology **281**, 229–238, (2016).
[240] A.L. Kesner, J.H. Chung, K.E. Lind, J.J. Kwak, D. Lynch, D. Burckhardt, P.J. Koo, "Validation of Software Gating: A Practical Technology for Respiratory Motion Correction in PET," Radiology, 152105 (2016).
[241] A.L. Kesner, P.J. Schleyer, F. Buther, M.A. Walter, K.P. Schafers, P.J. Koo, "On transcending the impasse of respiratory motion correction applications in routine clinical imaging - a consideration of a fully automated data driven motion control framework," EJNMMI Physics **1**, 8 (2014).
[242] M. Aristophanous, J.T. Yap, J.H. Killoran, A.B. Chen, R.I. Berbeco, "Four-dimensional positron emission tomography: implications for dose painting of high-uptake regions," Int J Radiat Oncol Biol Phys **80**, 900-908 (2011).
[243] M. Aristophanous, R.I. Berbeco, J.H. Killoran, J.T. Yap, D.J. Sher, A.M. Allen, E. Larson, A.B. Chen, "Clinical utility of 4D FDG-PET/CT scans in radiation treatment planning," Int J Radiat Oncol Biol Phys **82**, e99-105 (2012).
[244] J.M. Lamb, C. Robinson, J. Bradley, R. Laforest, F. Dehdashti, B.M. White, S. Wuenschel, D.A. Low, "Generating lung tumor internal target volumes from 4D-PET maximum intensity projections," Med Phys **38**, 5732-5737 (2011).
[245] L. Guerra, S. Meregalli, A. Zorz, R. Niespolo, E. De Ponti, F. Elisei, S. Morzenti, S. Brenna, A. Crespi, G. Gardani, C. Messa, "Comparative evaluation of CT-based and

respiratory-gated PET/CT-based planning target volume (PTV) in the definition of radiation treatment planning in lung cancer: preliminary results," Eur J Nucl Med Mol Imaging **41**, 702-710 (2014).
246 A. Chirindel, S. Adebahr, D. Schuster, T. Schimek-Jasch, D.H. Schanne, U. Nemer, M. Mix, P. Meyer, A.L. Grosu, T. Brunner, U. Nestle, "Impact of 4D-(18)FDG-PET/CT imaging on target volume delineation in SBRT patients with central versus peripheral lung tumors. Multi-reader comparative study," Radiother Oncol **115**, 335-341 (2015).
247 L.A. Pierce, 2nd, B.F. Elston, D.A. Clunie, D. Nelson, P.E. Kinahan, "A Digital Reference Object to Analyze Calculation Accuracy of PET Standardized Uptake Value," Radiology, 141262 (2015).
248 N. Withofs, C. Bernard, C. Van der Rest, P. Martinive, M. Hatt, S. Jodogne, D. Visvikis, J.A. Lee, P.A. Coucke, R. Hustinx, "FDG PET/CT for rectal carcinoma radiotherapy treatment planning: comparison of functional volume delineation algorithms and clinical challenges," J Appl Clin Med Phys **15**, 4696 (2014).
249 T. Shepherd, B. Berthon, P. Galavis, E. Spezi, A. Apte, J. Lee, D. Visvikis, M. Hatt, E. de Bernardi, S. Das, I. El Naqa, U. Nestle, C.R. Schmidtlein, H. Zaidi, A. Kirov, "Design of a Benchmark Platform for Evaluating PET-based Contouring Accuracy in Oncology Applications," Eur. J. Nucl Med Mol Imaging, **39**, S264 (2012).
250 B. Berthon, E. Spezi, C.R. Schmidtlein, A. Apte, P. Galavis, H. Zaidi, E. De Bernardi, J.A. Lee, A. Kirov, "Development of a software platform for evaluating automatic PET segmentation methods," Radiotherapy and Oncology **111**, S166 (2013).
251 M. Mamede, E.L.P. Abreu, M.R. Oliva, V. Nose, H. Mamon, V.H. Gerbaudo, "FDG-PET/CT tumor segmentation-derived indices of metabolic activity to assess response to neoadjuvant therapy and progression-free survival in esophageal cancer: correlation with histopathology results," Am J Clin Oncol **30**, 377-388 (2007).
252 H. Necib, C. Garcia, A. Wagner, B. Vanderlinden, P. Emonts, A. Hendlisz, P. Flamen, I. Buvat, "Detection and characterization of tumor changes in 18F-FDG PET patient monitoring using parametric imaging," J Nucl Med **52**, 354-361 (2011).
253 H.M. Mi, C. Petitjean, P. Vera, S. Ruan, "Joint tumor growth prediction and tumor segmentation on therapeutic follow-up PET images," Medical Image Analysis **23**, 84-91 (2015).
254 H.M. Mi, C. Petitjean, B. Dubray, P. Vera, S. Ruan, "Prediction of Lung Tumor Evolution During Radiotherapy in Individual Patients With PET," IEEE Trans Med Imaging **33**, 995-1003 (2014).
255 F. Sampedro, S. Escalera, A. Domenech, I. Carrio, "A computational framework for cancer response assessment based on oncological PET-CT scans," Comput Biol Med **55**, 92-99 (2014).
256 P. Obara, H. Liu, K. Wroblewski, C.P. Zhang, P. Hou, Y. Jiang, P. Chen, Y. Pu, "Quantification of metabolic tumor activity and burden in patients with non-small-cell lung cancer: Is manual adjustment of semiautomatic gradient-based measurements necessary?," Nucl Med Commun **36**, 782-789 (2015).
257 R.R. Beichel, M. Van Tol, E.J. Ulrich, C. Bauer, T. Chang, K.A. Plichta, B.J. Smith, J.J. Sunderland, M.M. Graham, M. Sonka, J.M. Buatti, "Semiautomated segmentation of head and neck cancers in 18F-FDG PET scans: A just-enough-interaction approach," Med Phys **43**, 2948 (2016).
258 A.I.J. Arens, E.G.C. Troost, B.A.W. Hoeben, W. Grootjans, J.A. Lee, V. Gregoire, M. Hatt, D. Visvikis, J. Bussink, W.J.G. Oyen, J.H.A.M. Kaanders, E.P. Visser, "Semiautomatic methods for segmentation of the proliferative tumour volume on sequential FLT PET/CT images in head and neck carcinomas and their relation to clinical outcome," Eur J Nucl Med Mol Imaging **41**, 915-924 (2014).
259 L. Drever, D.M. Robinson, A. McEwan, W. Roa, "A local contrast based approach to threshold segmentation for PET target volume delineation," Med Phys **33**, 1583-1594 (2006).
260 S. Vauclin, K. Doyeux, S. Hapdey, A. Edet-Sanson, P. Vera, I. Gardin, "Development of a generic thresholding algorithm for the delineation of 18FDG-PET-positive tissue:

application to the comparison of three thresholding models," Phys Med Biol **54**, 6901-6916 (2009).
261 I.A. Burger, H.A. Vargas, A. Apte, B.J. Beattie, J.L. Humm, M. Gonen, S.M. Larson, C. Ross Schmidtlein, "PET quantification with a histogram derived total activity metric: superior quantitative consistency compared to total lesion glycolysis with absolute or relative SUV thresholds in phantoms and lung cancer patients," Nuclear medicine and biology **41**, 410-418 (2014).
262 G. Li, C.R. Schmidtlein, I.A. Burger, C.A. Ridge, S.B. Solomon, J.L. Humm, "Assessing and accounting for the impact of respiratory motion on FDG uptake and viable volume for liver lesions in free-breathing PET using respiration-suspended PET images as reference," Med Phys **41, 091905** (2014).
263 F. Kong, M. Machtay, J. Bradley, R. Ten Haken, Y. Xiao, M. Matuszak, "RTOG 1106/ACRIN 6697: Randomized phase II trial of individualized adaptive radiotherapy using during treatment FDG-PET," CT and modern technology in locally advanced non-small lung cancer (NSCLC) (2012).
264 F.M. Kong, K.A. Frey, L.E. Quint, R.K. Ten Haken, J.A. Hayman, M. Kessler, I.J. Chetty, D. Normolle, A. Eisbruch, T.S. Lawrence, "A pilot study of [18F]fluorodeoxyglucose positron emission tomography scans during and after radiation-based therapy in patients with non small-cell lung cancer," J Clin Oncol **25**, 3116-3123 (2007).
265 L. Drever, W. Roa, A. McEwan, D. Robinson, "Iterative threshold segmentation for PET target volume delineation," Med Phys **34**, 1253-1265 (2007).
266 N.C. Krak, R. Boellaard, O.S. Hoekstra, J.W. Twisk, C.J. Hoekstra, A.A. Lammertsma, "Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial," Eur J Nucl Med Mol Imaging **32**, 294-301 (2005).
267 I.A. Burger, H.A. Vargas, B.J. Beattie, D.A. Goldman, J. Zheng, S.M. Larson, J.L. Humm, C.R. Schmidtlein, "How to assess background activity: introducing a histogram-based analysis as a first step for accurate one-step PET quantification," Nucl Med Commun **35**, 316-324 (2014).
268 M. Vanderhoek, S.B. Perlman, R. Jeraj, "Impact of the definition of peak standardized uptake value on quantification of treatment response," J Nucl Med **53**, 4-11 (2012).
269 M. Miller, G. Hutchins, "3D Anatomically accurate phantoms for PET and SPECT imaging," IEEE Nuclear Science Symposium and Medical Imaging Conference Record, Proceedings paper, M26-8, p. 4252-4256, **49** ((2007).
270 B. Berthon, C. Marshall, R. Holmes, E. Spezi, "A novel phantom technique for evaluating the performance of PET auto-segmentation methods in delineating heterogeneous and irregular lesions," EJNMMI Physics **2**, 13 (2015).
271 A. Le Maitre, W.P. Segars, S. Marache, A. Reilhac, J.M. Hatt, S. Tomei, C. Lartizien, D. Visvikis, "Incorporating Patient-Specific Variability in the Simulation of RealisticWhole-Body 18F-FDG Distributions for Oncology Applications," Proceedings of the IEEE **97**, 2026-2038 (2009).
272 P. Papadimitroulas, G. Loudos, A. Le Maitre, M. Hatt, F. Tixier, N. Efthimiou, G.C. Nikiforidis, D. Visvikis, G.C. Kagadis, "Investigation of realistic PET simulations incorporating tumor patient's specificity using anthropomorphic models: Creation of an oncology database," Med Phys **40** (2013).
273 J.R. Munkres, *Topology (2nd ed.)*. (Prentice Hall, 1999).
274 N. Aspert, D. Santa-Cruz, T. Ebrahimi, "Mesh: Measuring errors between surfaces using the Hausdorff distance," IEEE International Conference on Multimedia and Expo (ICME), **1**, pp. 705-708 (2002).

TABLES

Table 1. Summary of the main characteristics of some representative advanced PET-AS algorithms (not an exhaustive list) and their respective evaluation.

Reference(s)	Image segmentation paradigm(s) used	User interaction [1]	Pre- and post-processing steps	Aimed application[2]	Validation data and ground truth [3]	Accuracy evaluation on realistic tumors[5]	Robustness evaluation[6]	Repeatability evaluation
Tylski, et al. 2006 [62]	Watershed	Std + multiple markers placement	None	Global	PA(1): Vol. and CiTu images	No	No	No
Werner-Wasik, et al. *2012* [63]	Gradient-based	Std + initialization using drawn diameters	Unknown	Global	PA(5): Diam. 31 MCST: Vol.	Yes	Yes	Yes
Geets, et al. 2007[37]	Gradient-based	Std + initialization	Denoising and deconvolution steps	Global	PS(1) and PA(1): Vol. + Diam. 7 CiTuH: Complete	No	No	No
El Naqa, et al. 2008[64]	AT + active contours	Std + several parameters to set	None	Global	PA(1): Vol. 1 CiTu - Ø	Yes	No	Yes
El Naqa, et al. 2007[31]	Multi-modal (PET/CT) active contours	Std + initialization of the contour shape, selection of weights	Normalization and registering of PET and CT images, deconvolution of PET images.	GTV definition on PET/CT	PA(1): Vol. 2 CiTu: MC(1), FT	Yes	No	No
Dewalle-Vignion, et al. 2011[65]	Possibility theory applied to MIP projections	Std	None	Global	PA(1): Vol. 5 MCST: Vox. 7 CiTuH: Complete	Yes	No	No
Belhassen and Zaidi 2010[28]	Improved Fuzzy C-Means (FCM)	A posteriori interpretation of resulting classes in the segmentation of entire image	Denoising, wavelet decompositions	Global	3 AST: Vox. 21 CiTuH: Diam. 7 CiTuH: Complete	Yes	No	No
Aristophanous, et al. 2007[66]	Gaussian mixture modeling without spatial constraints	Std + initialization of the model and selection of the number of classes	None	Pulmonary tumors	7 CiTu: Ø	No	No	Yes
Montgomery, et al. 2007[67]	Multi scale Markov field segmentation	A posteriori interpretation of resulting segmentation on the entire image	Wavelet decompositions	Global	PA(1) : Vol. 3 CiTu: Ø	No	No	No
Hatt, et al. 2007[68]	Fuzzy Hidden Markov Chains	Std	None	Global	PS(1) and PA(2): Vox.	No	No	No
Hatt, et al. 2009[69], 2010 [70],2011[43]	Fuzzy locally adaptive Bayesian	Std	None	Global	PS(1) and PA(4): Vox. 20 MCST: Vox. 18 CiTuH: Diam.	Yes	Yes	Yes
Day, et al. 2009[60]	Region growing based on mean and SD of the region	Std + optimization on each scanner	None	Rectal tumors	18 CiTu: MC(1)	No	No	No
Yu, et al. 2009[40], Markel, et al. 2013[71]	Decision tree built based on learning of PET and CT textural features	A posteriori interpretation of resulting segmentation on the entire image	Learning for building the decision tree	GTV definition of H&N and lung tumors	10 CiTu: MC(3) 31 CiTu: MC(3)	No	No	No

Sharif, et al. 2010[72], 2012[73]	Neural network	A posteriori interpretation of resulting segmentation on the entire image.	Learning for building the neural network	Global	PA(1): Vol. 3 AST: Vox. 1 CiTuH: Diam.	No	No	Yes
Sebastian, et al. 2006[74]	Spherical Mean shift	Std	Resampling in a different spatial domain	Global	280 AST: Vox.	No	No	No
Janssen, et al. 2009[75]	Voxels classification based on time-activity curve	Std + Initialization and choice of the number of classes	Only on dynamic imaging, denoising and deconvolution steps	Rectal tumors in dynamic imaging	PA(1): Vol. + Diam. 21 CiTu: MC(1)	No	No	No
De Bernardi, et al. 2010[76]	Combined with PVE (image reconstruction)	Std + initialization	PSF model of the scanner and access to raw data required	Global	PA(1): Vol.	No	Yes	No
Bagci, et al. 2013[33]	Multimodal random walk	Std	Multimodal images registration	Global for PET/CT or PET/MR	77 CiTu: MC(3) PA(1)	Yes	No	No
Onoma, et al. 2014 [77]	Improved random walk	Std	None	Global	PA(1), 4 AST: Vox, 14 CiTu: MC(2)	Yes	No	No
Song, et al. 2013[32]	Markov field + graph cut	Std	None	Global	3 CiTu: MC(3)	Yes	No	No
Hofheinz, et al. 2013[61]	Locally adaptive thresholding	Std + one parameter to determine on phantom acquisitions	None	Global	30 AST: Vox.	Yes	No	No
Abdoli, et al. 2013[78]	Active contour	Std + several parameters to optimize	Wavelet decompositions	Global	1 AST: Vox. 9 CiTuH: Complete. 3 CiTuH: Complete 2 CiTuH: Complete	No	Yes	No
Mu, et al. 2015 [79]	Level set combined with PET/CT Fuzzy C-Means	Std	None	Specific to cervix	7 AST: Vox, 27 CiTu: MC(2)	Yes	No	No
Cui, et al. 2015 [80]	Graph cut improved with topology modeling	Std + one free parameter previously optimized	PET/CT registration	Specific to lung tumors and PET/CT	20 PA(1), 40 CiTu(2)	Yes	No	No
Lapuyade-Lahorgue, et al. 2015 [81]	Generalized fuzzy C-means with automated norm estimation	Std	None	Global	PA(4): Vol. 34 MCST: Vox. 9 CiTu: MC(3).	Yes	Yes	Yes

Devic et al, 2016 [82]	Differential uptake volume histograms for identifying biological target sub-volumes	Selection of three ROIs encompassing PET avid area; iterative decomposition of differential uptake histograms into multiple Gaussian functions	None	Isolation of glucose phenotype driven biological sub-volumes specific to NSCLC,	None	No	No	No
Berthon et al, 2016 [83]	Decision tree based learning using nine different segmentation approaches (region growing, thresholds, FCM, etc.) with the goal of selecting the most appropriate method given the image characteristics	Std	Learning on 100 simulated cases to train/build the decision tree	Global	85 NSTuP: Vox.	Yes	No	No
Schaefer et al, 2016 [84]	Consensus between contours from 3 segmentation methods (contrast-oriented, possibility theory, adaptive thresholding) based on majority vote or STAPLE	Std	None	Global	22 CiTuH; Complete 10 CiTu: MC(4) 10 CiTu: MC(1)	Yes	No	Yes

Table legend

[1] std = « standard » interaction (i.e. the metabolic volume of interest is first manually isolated in a region of interest that is used as an input to the algorithm.)

[2] global = not application specific

[3] PA(x) = Phantom (spheres) Acquisitions on x different scanners; PS(x) = Phantom (spheres) Simulations on x different scanners; AST = Analytically Simulated Tumors; MCST = Monte Carlo Simulated Tumors; NSTuP = Non spherical tumors simulated in phantoms (thin-wall inserts, printed phantoms, etc.). CiTu = Clinical Tumors; CiTuH = Clinical Tumors with Histopathology

Vol. = only volume; Vox. = voxel-by-voxel; Diam = histopathology maximum diameter; Complete = 3D histopathology reconstruction; MC(x) = manual contouring by x experts; FT = fixed threshold, AT = adaptive threshold

[4] Highly heterogeneous, complex shapes, low contrasts and rigorous ground truth.

[5] Requires multiple acquisitions on different systems and a large number of parameters.

[6] With respect to different scanners and protocols

Table 2. A summary of segmentation method comparisons and reviews.

No.	Reference	Compared methods	Images or Phantoms used	Results and/or recommendations reported by the authors	Limitations and comments by TG211 or others as cited
Comparison studies					
1	Nestle, et	Visual	Patient scans	Large differences	Only visual

	al. 2005[23]	segmentation, 40% of SUV_{max} threshold, SUV>2.5 threshold, and an adaptive threshold		between volumes obtained with the 4 approaches	segmentation used as a surrogate of truth and only clinical data.
2	Schinagl, et al. 2007[158]	Visual segmentation, 40% and 50% of SUV_{max} threshold, and adaptive thresholding	78 Clinical PET/CT images of head and neck	The 5 methods led to very different volumes and shapes of the GTV. Fixed threshold at SUV of 2.5 led to the most disappointing results.	The GTV was defined manually on CT and used as a surrogate of truth for PET-derived delineation and only clinical data was used.
3	Geets, et al. 2007[37], Wanet, et al. 2011[56]	Fixed and adaptive thresholding, gradient-based segmentation	Phantom (spheres), simulated images, clinical images of lung and H&N cancers with histopathology 3D measurements	More accurate segmentation with gradient-based approach compared to threshold	Numerous issues associated with the 3D reconstruction of the surgical specimen used as gold standard.
4	Greco, et al. 2008[159]	Manual segmentation, 50% SUV_{max}, SUV>2.5 threshold and iterative thresholding	12 Head and neck cancer patients.	Thresholding PET-AS algorithms are strongly threshold-dependent and may reduce target volumes significantly when compared to visual and physician-determined volumes.	Limitations: reference GTVs defined manually on CT and MRI.
5	Vees, et al. 2009[160]	Manual segmentation, SUV>2.5 threshold, 40% and 50% of SUV_{max} threshold, adaptive thresholding, gradient-based method	18 Patients with high grade glioma	PET often detected tumors that are not visible on MRI and added substantial tumor extension outside the GTV defined by MRI in 33% of cases. The 2.5 SUV isocontour and "Gradient Find"	Ground-truth derived from manual segmentation on MRI only.

		and region growing.		"segmentation techniques performed poorly and should not be used for GTV delineation".	
6	Belhassen, et al. 2009[161]	Three different implement-ations of the fuzzy C-means (FCM) clustering algorithm	Patient scans	Incorporating wavelet transform and spatial information through nonlinear anisotropic diffusion filter improved accuracy for heterogeneous cases	No comparison with other standard methods
7	Tylski, et al. 2010[42]	Four different threshold methods (% of max activity and three adaptive thresholding), and a model-based thresholding	Spheres in an antropomor-phic torso phantom as well as non spherical simulated tumors	Large differences between volumes obtained with different segmentation algorithms. Model-based or background-adjusted algorithms performed better than fixed thresholds.	No clinical data, limited to threshold-based algorithms only, only volume error considered as a metric
8	Hatt, et al. 2010[70], 2011[43, 152]	Fixed and adaptive thresholding, Fuzzy C-means, FLAB	IEC phantom (spheres); simulated images, clinical images with maximum diameter measurements in histopathology	Advanced algorithms are more accurate compared to threshold- based and are also more robust and repeatable.	For clinical images, only maximum diameters along one axis were available from histology.
9	Dewalle-Vignion, et al. 2011[65], 2012[162]	Manual segmentation, 42% of SUV_{max}, 2 different adaptive thresholding, fuzzy c-mean and an advanced method based on fuzzy set	Phantom images, simulated images, clinical images with manual delineations	The advanced algorithm is more accurate and robust than threshold-based and closer to manual delineations by clinicians	Only manual delineation for surrogate of truth of clinical data. Comments: Interesting use of various metrics for assessment of image segmentation accuracy.

		theory.			
10	Werner-Wasik, et al. 2012[63]	Manual segmentation, fixed thresholds at 25% to 50% of SUV_{max} (by 5% increments) and gradient-based segmentation	IEC phantom (spheres) in multiple scanners, simulated images of lung tumors	A gradient-based algorithm is more "accurate and consistent" than manual and threshold segmentation.	Only volume error used as a metric of performance. "…manual verification using CT scan should be performed (…) GTV definition requires joint assessment by the radiologist, nuclear physician and radiation oncologist" [163]
11	Zaidi, et al. 2012[150]	Five thresholding methods, Standard and improved fuzzy c-means, level set technique, stochastic EM.	Patient scans with histopathology 3D measurements (same as #5 above) N/A	The automated Fuzzy c-means algorithm provided was shown to be more accurate than 5 thresholding algorithms, the level set technique, the stochastic EM approach and regular FCM. Adaptive threshold techniques need to be calibrated for each PET scanner and acquisition/pro-cessing protocol and should not be used without optimization.	Numerous issues associated with the 3D reconstruction of the surgical specimen used as gold standard. See Table 1
12	Shepherd, et al. 2012 [115]	30 methods from 13 different groups.	Tumor and lymph-node metastases in H&N cancer and physical phantom (irregular shapes). Simulated, experimental and clinical studies	Highest accuracy is obtained from optimal balance between interactivity and automation. Improvements are seen from visual guidance by PET gradient as well as by using CT.	Limitation: a small number of objects (n=7) were used for the evaluation.
13	Schaefer, et al. 2012[157]	One adaptive thresholding	Phantoms, same threshold	The calibration of an adaptive threshold PET-AS algorithm is	Confirmation of previous findings about adaptive

		technique	algorithm, different scanners	scanner and image analysis software-dependent.	threshold segmentation.
14	Schinagl et al. 2013 [164]	Visual, SUV of 2.5, fixed threshold of 40% and 50%, and two adaptive threshold based methods using either the primary or the metastasis	Evaluation of the segmen-tation of metastatic lymph nodes against pathology in 12 head and neck cancer patients	SUV of 2.5 was unsatisfactory in 35% of cases; for the last four methods: i) using the node as a reference gave results comparable to visual ii) using the primary as a reference gave poor results;	Shows the limitations of threshold based methods.
15	Hofheinz, et al. 2013, [61]	Voxel-specific adaptive thresholding and standard lesion-specific adaptive threshold	30 simulated images based on real clinical datasets.	The voxel-specific adaptive threshold method was more accurate than the lesion-specific one in heterogeneous cases	Only simulated data were used.
16	Lapuyade-Lahorgue, et al. 2015 [81]	Improved generalized fuzzy c-means, fuzzy local information C-means and FLAB	34 simulated tumors and 9 clinical images with consensus of manual delineations. 3 acquisitions of phantoms for robustness of evaluation.	In both simulated and clinical images, the improved generalized FCM led to better results than another FCM implementation and FLAB, especially on complex and heterogeneous tumours, without any loss of robustness on data acquired in different scanners.	Only 9 clinical images used.
			Reviews		
1	Boudraa, et al. 2006 [39]	N/A	Mostly clinical images	Extensive review of the fomalism of image segmentation algorithms used in nuclear medicine (not specific to PET and clinical oncology)	Only a few algorithms have been rigorously validated for accuracy, repeatability and robustness.
2	Lee, 2010[29]	N/A	Simulated, mostly clinical	Discussed the main caveats of threshold-	Extensive review of nuclear

			images	based techniques including the effect of phantom cold walls on threshold. A discussion of the available and desirable validation datasets and approaches is also included.	medicine image segmentation algorithms (not specific to PET and clinical oncology)
3	Zaidi and El Naqa, 2010 [54]	N/AN/A	Simulated, experimental and clinical studies	Despite being promising, advanced PET-AS algorithms are not used in the clinic.	
4	Hatt, et al. 2011[23]	N/A	N/A	Only a few algorithms have been rigorously validated for accuracy, repeatability and robustness.	See the three last columns of Table 1
5	Kirov A.S, Fanchon, L, 2014,[165]	N/A	Clinical images with pathology derived ground truth.	Articles comparing PET-AS methods are summarized separately for lesions in five groups based on location in the body with a focus on the accuracy, usefulness and the role of the pathology-validated PET image sets.	
6	Foster, et al. 2014[36]	N/A	N/A	"although there is no PET image segmentation method that is optimal for all applications or can compensate for all of the difficulties inherent to PET images, development of trending image segmentation techniques which combine anatomical information and metabolic activities in the same hybrid	Most exhaustive review of the state-of-the-art in 2014. Uses a similar classification of methods as in the present report.

				frameworks (PET-CT, PET-CT, and MRI-PET-CT) is encouraging and open to further investigations.”	

2725

Table 3. Advantages and disadvantages of the different methods used for generation of test images.

	Experimental images			**Simulated images**	
	Simple Phantoms	**Realistic phantoms**	**Clinical images**	**Forward Projected Images**	**Monte Carlo (MC) Simulations**
Advantages:	• Exact representation of the scanner resolution, image noise and other image artifacts • Ground truth accurately known • Easy to generate and use	• Exact representation of the scanner resolution, image noise and other image artifacts • Capable to produce lesion shapes corresponding to actual tumors • Known ground truth	• Exact representation of the scanner resolution, image noise and other image artifacts • Real tumors	• Precise experimental control • Flexibility in phantom design • Precise knowledge of the reference object • Computationally cheap	• Precise experimental control • Realistic count distributions • Flexibility in phantom design • Precise knowledge of the reference object • Camera-specific information
Dis-advantages:	• The objects have simplistic and unrealistic shape and activity distribution • Most with few exceptions[171, 172] have cold walls	• Some uncertainties in the experiment can be hard to quantify • Difficult to generate inhomogeneous activity • Labor intensive	• Uncertainties in the knowledge of the reference object, even with histopathology reference	• Scatter count distributions and noise are usually less accurately modeled • Detailed physics and system information ignored	• Computationally expensive • Model requires extensive up front experience

2730

Table 4. A comparison of various volume/contour agreement measures and their sensitivities to the properties of the segmented lesions. The important properties are whether they account for volume differences, shape discrepancies, false positive vs. false negative. The computational complexity is graded between easy (+) and complicated (+++), although none of the metrics are particularly slow to compute using modern toolkits and computers (Barycenter distance is the distance between the centers of mass of two sets).

Evaluation criteria	Location	Size	Shape	Type I/II	Complexity
Volume difference	no	yes	no	no	+
Barycenter distance	yes	no	no	no	++
Jaccard similarity coefficient	yes	yes	yes	no	++
Dice similarity coefficient (DSC)	yes	yes	yes	no	++
Hausdorff distance	yes	no	yes	no	+++
Sensitivity + Positive Predictive Value (PPV)	yes	yes	yes	yes	++

Table. 5 Stages of evaluation of PET auto-segmentation (PET-AS) methods. DSC (Dice Similarity Coefficient), PPV (Positive Predictive Values), HD (Hausdorff Distance).

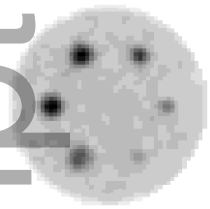
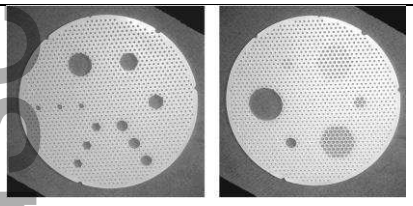
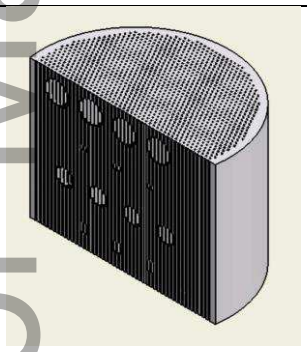
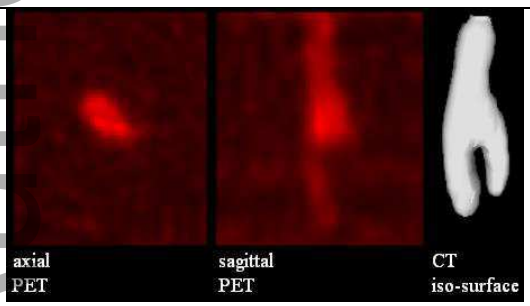
Step	1. Vendor acceptance	2. Basic evaluation	3. Phase two evaluation:	4. Phase three evaluation	5. Impact evaluation
Objective	Proper functioning of software	Accuracy of clinic-specific images; robustness of image properties	Accuracy, repeatability and robustness of realistic shapes and variable uptake;	Accuracy, repeatability and robustness of clinical images from the intended application	Evaluation of clinical impact
Datasets	Vendor recommendation	Simple objects in uniform background; repeated acquisitions	Irregular shape and/or non-uniform uptake lesions in experimental or digital phantoms without cold wall; multiple realizations	Clinical images	Clinical images, treatment plans and follow-up records
Ground truth	Vendor recommendation	CT defined voxel level accuracy.	High resolution CT or digital ground-truth defined voxel level accuracy.	Digitized histopathology and/or consensus of several manual delineations	Treatment outcome data
Metrics	Vendor recommendation	Volume errors, DSC	DSC, Sensitivity, PPV, HD	DSC, Sensitivity, PPV, HD, Statistical evaluation of clinical endpoint (prognostic/predictive value)	Statistical multi-parameter treatment outcome analysis

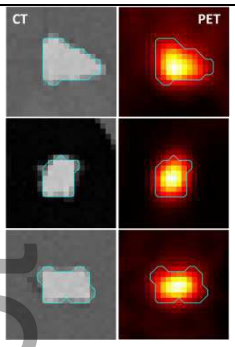
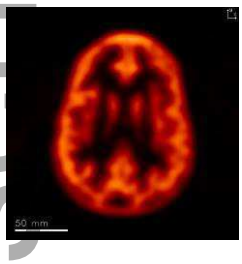
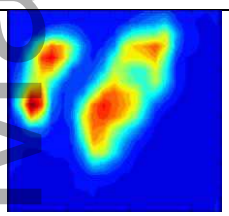
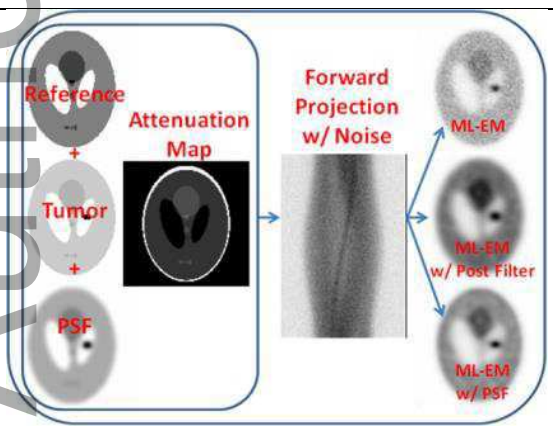
Table A1: Functional forms of various threshold segmentation schemes.

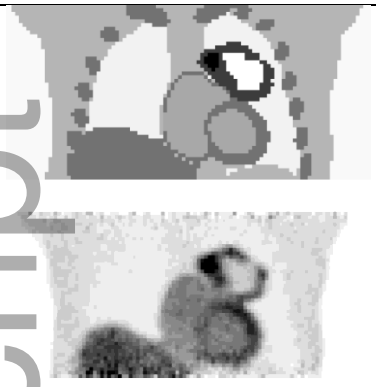
Comments	Threshold Estimator
Drever, et al.'s single-parameter FTS fit[259]: It is most notable for its use of the histogram's mode for more stable estimation of the background.	$T = a\left(I_{max} - I_{bkg}\right) + I_{bkg}$
Nestle, et al.'s single-parameter FTS fit[23]: This fit uses the mean of voxels greater than 70% of the lesion's maximum. The use of the mean instead of the maximum uptake reduces the variability.	$T = a\, I_{mean,70\%max} + I_{bkg}$
Daisne, et al.'s show a two-parameter FTS fit model [149]: The scaling parameter, $\mathrm{I_{max}}$, can be recast as a mean-value or volume-based measure for an ATS algorithm[260].	$T = a + b\, \frac{I_{bkg}}{I_{max}}$
Schaefer, et al.'s two-parameter FTS fit[30]: This fit is extended from Nestle's scheme above[23].	$T = \frac{a\, I_{mean,70\%max} + b\, I_{bkg}}{I_{max}}$
Erdi, et al.'s two-parameter FTS fit[48]: uses a fixed threshold of 42%. It was noted that it worked well for large lung tumors, however the authors go on to say that its use should be limited to homogeneous uptake distributions.	$T = a\, e^{-b\, V(T)}$
Black, et al.'s two-parameter ATS fit[53]: The use of the mean SUV to make the algorithm more stable to noise requires a threshold for its calculation.	$T = a + b\, I_{mean}\left(V(T)\right)$
Biehl, et al.'s two-parameter ATS fit[46]: The volume is the GTV defined by CT. This algorithm is shown to work for a wide range of tumor volumes in NSCLC.	$T = I_{max}\left(a + b\, ln(CT_{GTV})\right)$
Jentzen, et al.'s three-parameter ATS fit[51]: The parameters were fitted from phantom data. The inverse volume requires a threshold.	$T = \frac{a}{V(T)} + b\, \frac{I_{bkg}}{I_{max}} + c$
Nehmeh, et al.'s four-parameter ATS fit[52]: The fit used Monte Carlo simulation results to avoid cold wall effects.	$T = I_{max}\left(a + b\, V^{c}(T) e^{d/V(T)}\right)$
Burger, et al.'s Background Subtracted Lesion (BSL) [261]: Not meant as segmentation but rather a volume estimation scheme, an equivalent volume threshold can be found (Li et al.[262]) Note that this method tends to overestimate the volume by including spill-out.	Procedure: T, such that the volume from a threshold matches the volume used in the BSL measurement.

2750

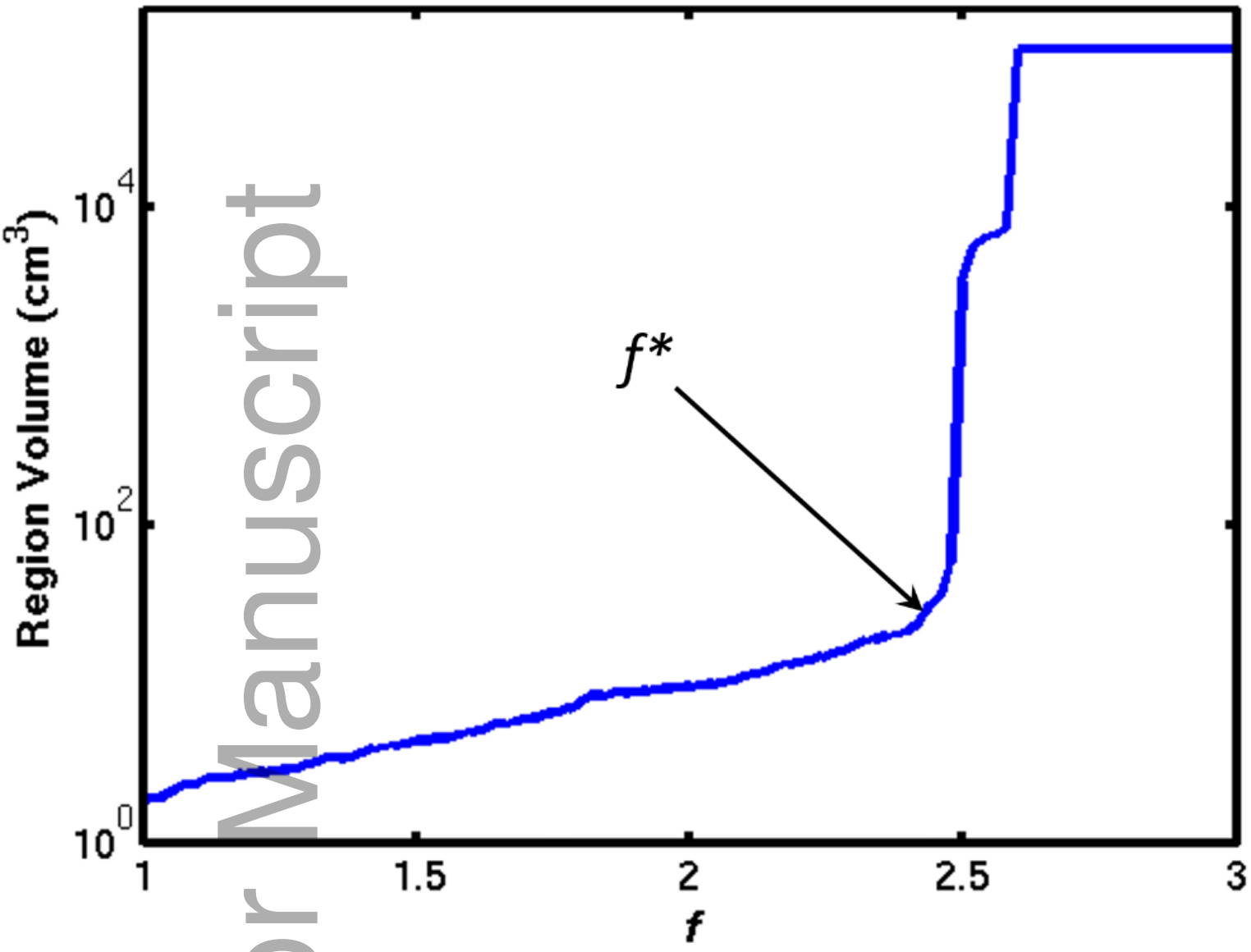
Table A2. A summary of the existing phantoms that are considered as potential candidates to provide data for a future PET-AS evaluation protocol.

Name or type and Reference	Example figure and brief description	Advantages	Disadvantages
Experimental phantoms			
IEC, NEMA NU 2-2001[173]	 Fig. T.A2.1. Image of a transaxial slice through the center of the spheres	Known ground truth, Variable sphere size, Widely available	Overly simple unrealistic lesion shapes; Uniform background; cold walls
"Porous phantom," Di Filippo, et al. 2004[172]	 Fig. T.A2.2. Perforated discs used for the construction of the hot spheres: multi resolution (left), multi contrast (right). Reprinted with permission from Med. Phys. J..	Adds variable contrast, eliminates cold walls	Simple shapes, Uniform uptake Tedious manufacturing
"Swiss cheese," Hunt, *et al.* 2007[171]	 Fig. T.A2.3. A cut view of the phantom, produced by rapid prototyping. Reprinted with permission from Med. Phys. J..	Similar to the "Porous phantom," above Easier manufacturing	Simple shapes, Uniform uptake
Tumor phantom, Shepherd, et al. 2012[115]	 Fig. T.A2.4. PET images and CT image based iso-surface of the tumor model phantom.	Irregular shapes including branching Based on clinical images Ground truth in PET image space from hybrid CT thresholding	Cold walls of glass compartments Does not recreate heterogeneity (requires internal medium of spatially varying absorbance)

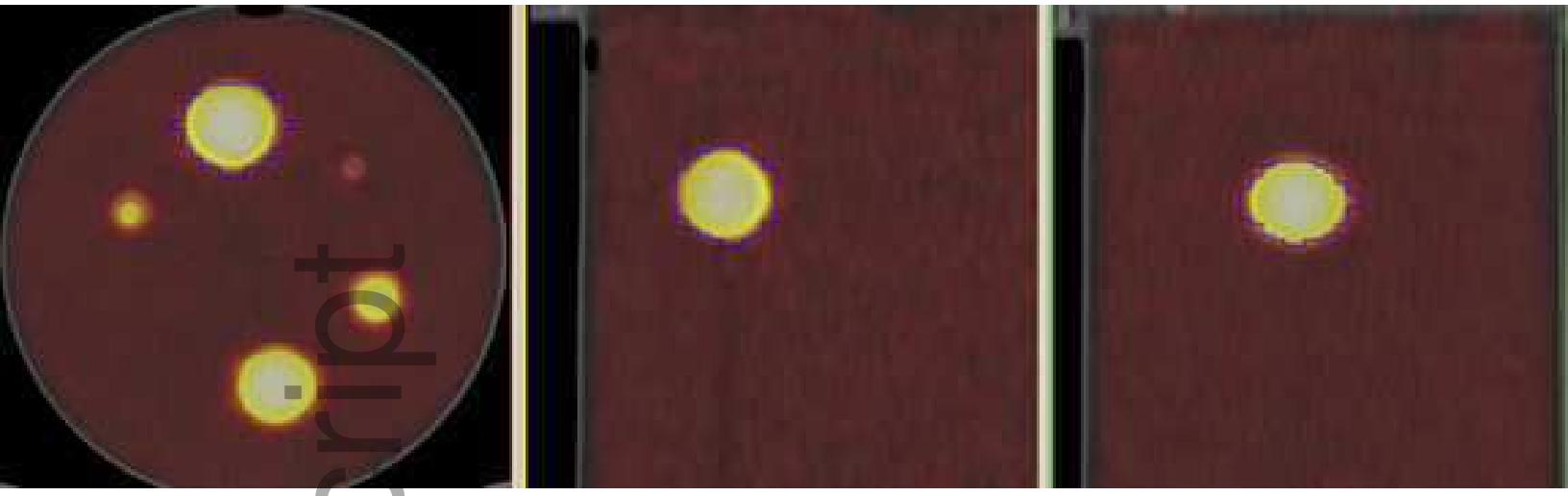
Molecular sieves, Zito, et al. 2012[174]	 Fig. T.A2.5. CT and PET images of 3 zeolites with superimposed ground truths.	Known ground truth; irregular shapes	Uniform lesion uptake; uniform background
Using Radioactive Ink in 2D and 3D printers, Larsson, et al. 2000[175], El-Ali, et al. 2003[176], Miller & Hutchins, 2008[269,177], Berthon et al, 2015[270]	 Fig. T.A2.6. PET image of the human brain phantom produced by incorporating a radioactive dye in rapid prototyping.[269] (Reprinted with permission.)	Can match the irregular shape and non-uniform uptake of real lesions	A specially adapted and calibrated printer is needed
NonU phantom, Kirov, et al. 2011[178]	 Fig. T.A2.7. Relative activity levels of the reference activity distribution obtained by displacing activity by thin foils with irregular cutouts.	Can match the irregular shape and non-uniform uptake of real lesions and background.	Expensive Need to improve accuracy of ground truth
Simulated phantoms			
Simulated PET images from forward projected reference images, 2015[181]	 Fig. T.A2.8. A schematic showing the generation of PET-like images from reference activity, tumor and attenuation distribution images. These reference images are forward projected, scaled,	Realistic tumor and normal tissues and organs simulations, irregular shapes and heterogen-eous activity distributions Known ground truth	Because geometric projection, rather than photon transport is used, many of the physical aspects of the real image acquisition process are ignored

	blurred and noise is added to simulate realistic PET data that is subsequently reconstructed.		
Monte Carlo simulations Le Maitre*, et al.* 2009 [271] Papadimitro-ulas, et al. 2013 [272]	 Fig. T.A2.9.	Realistic tumor and normal tissues and organs simulations, irregular shapes and heteroge-neous activity distributions; Known ground truth	Computationally expensive; requires extensive up front experience.

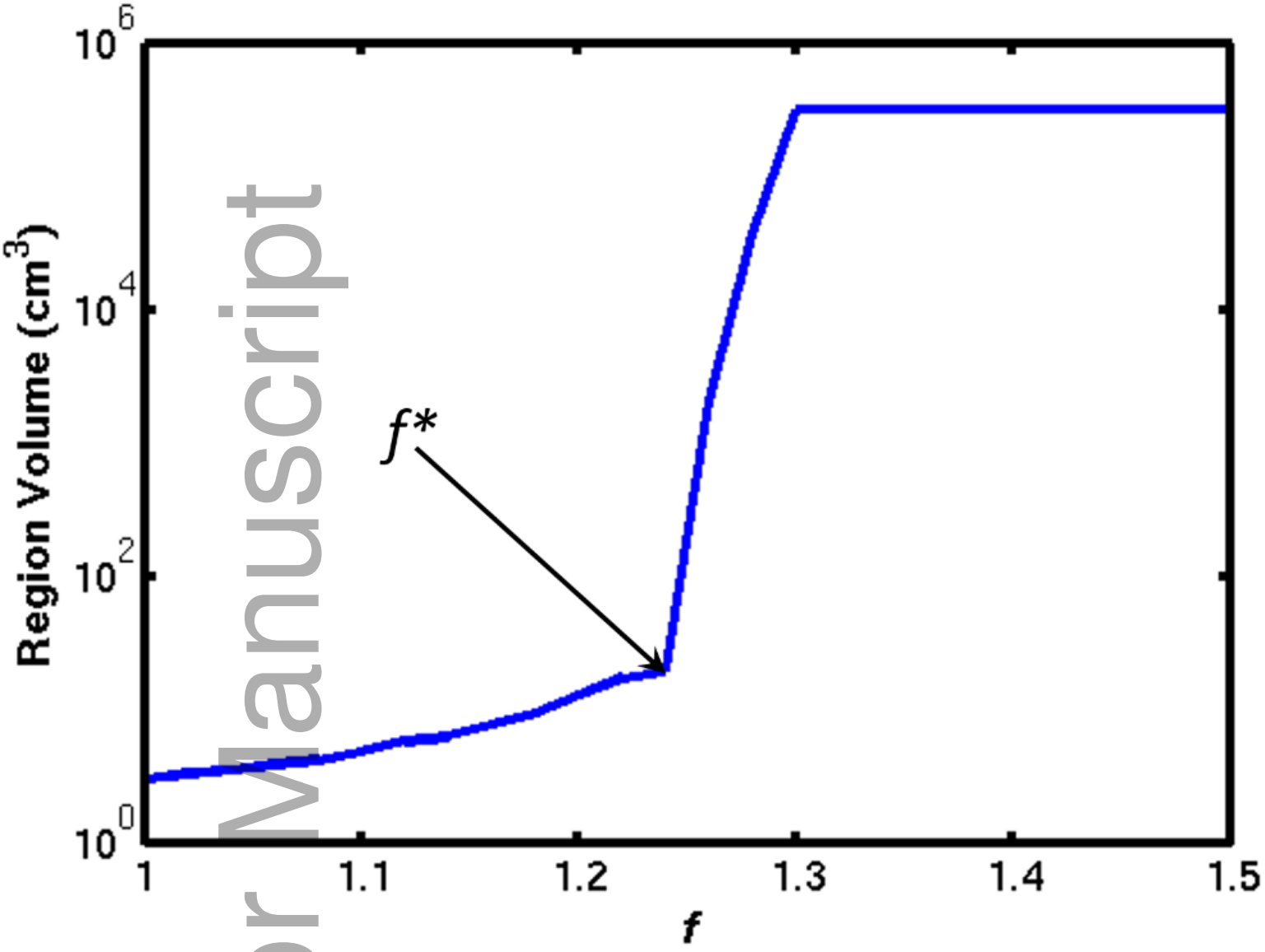
2755



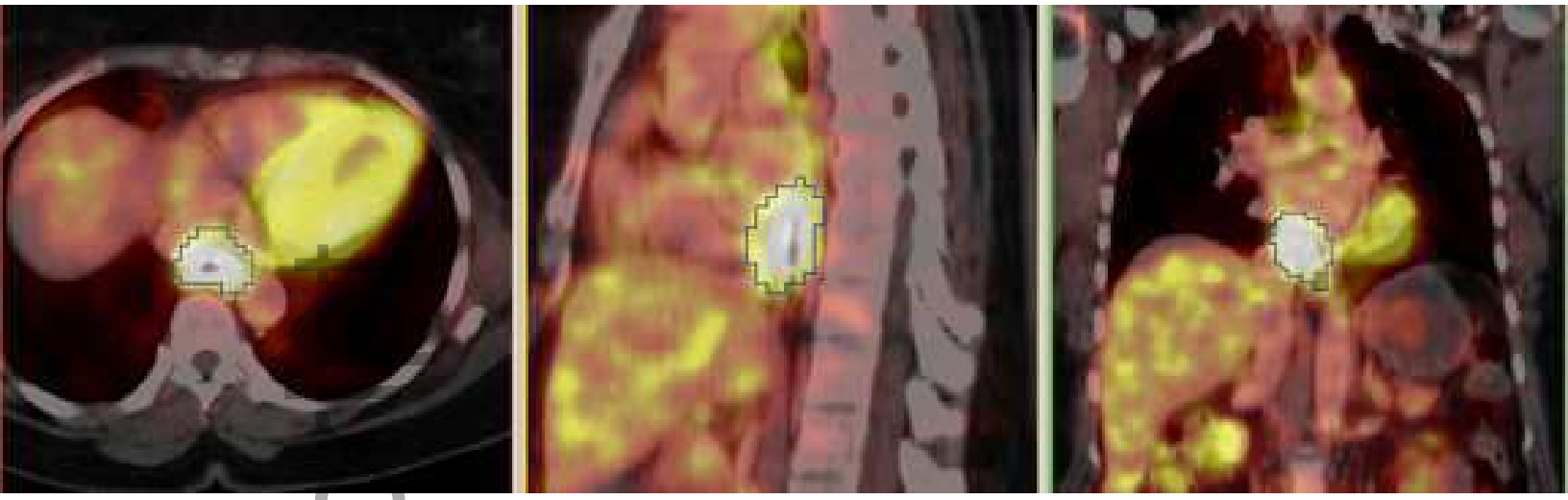
mp_12124_f1a.eps



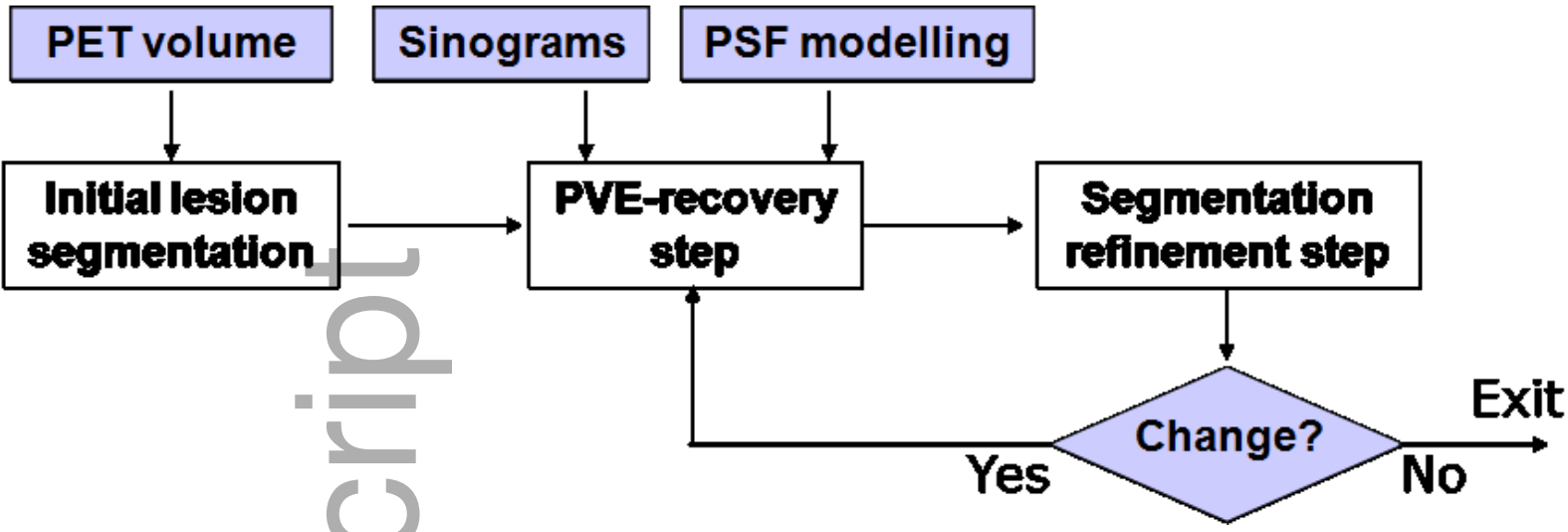
mp_12124_f1b.eps



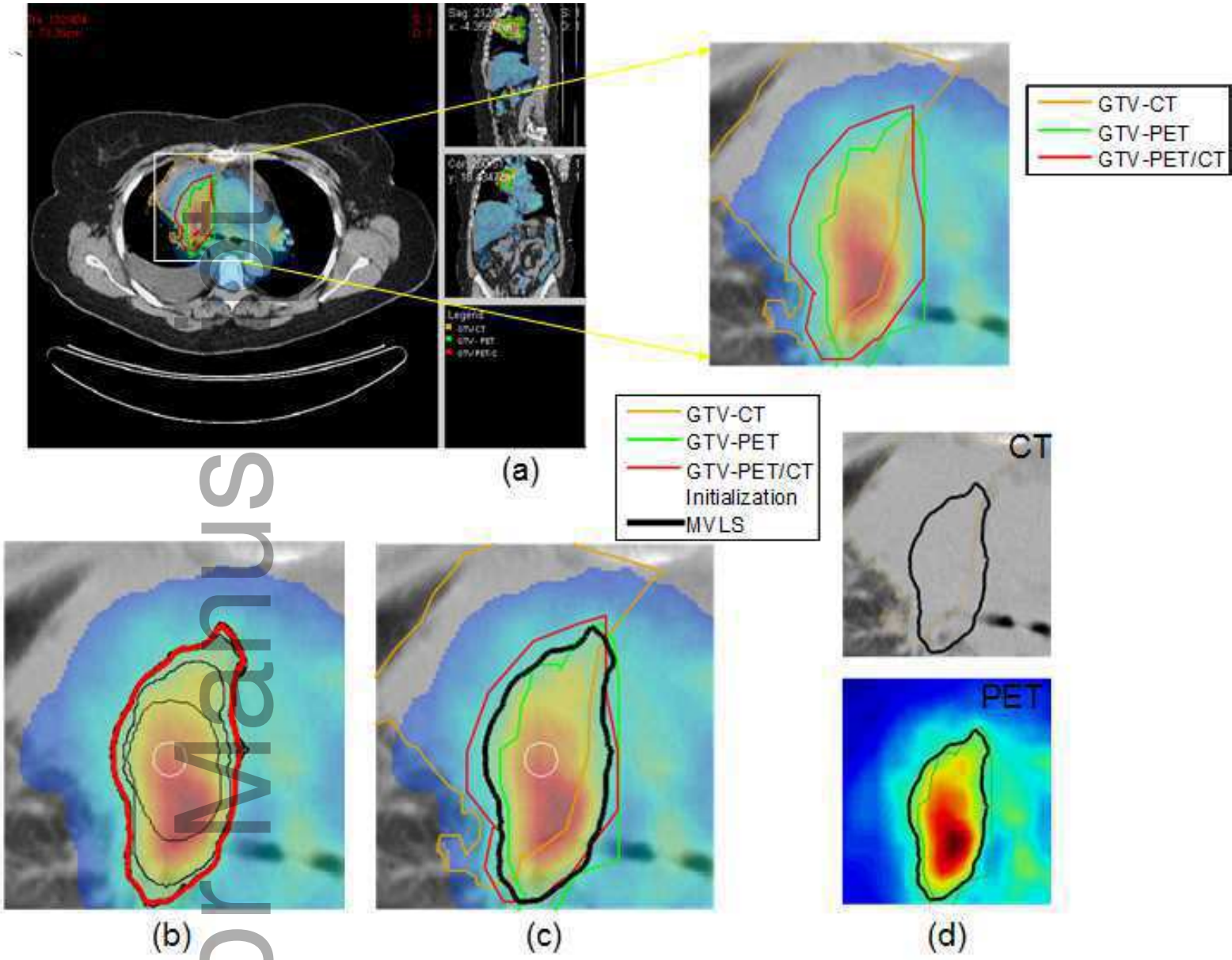
mp_12124_f1c.eps



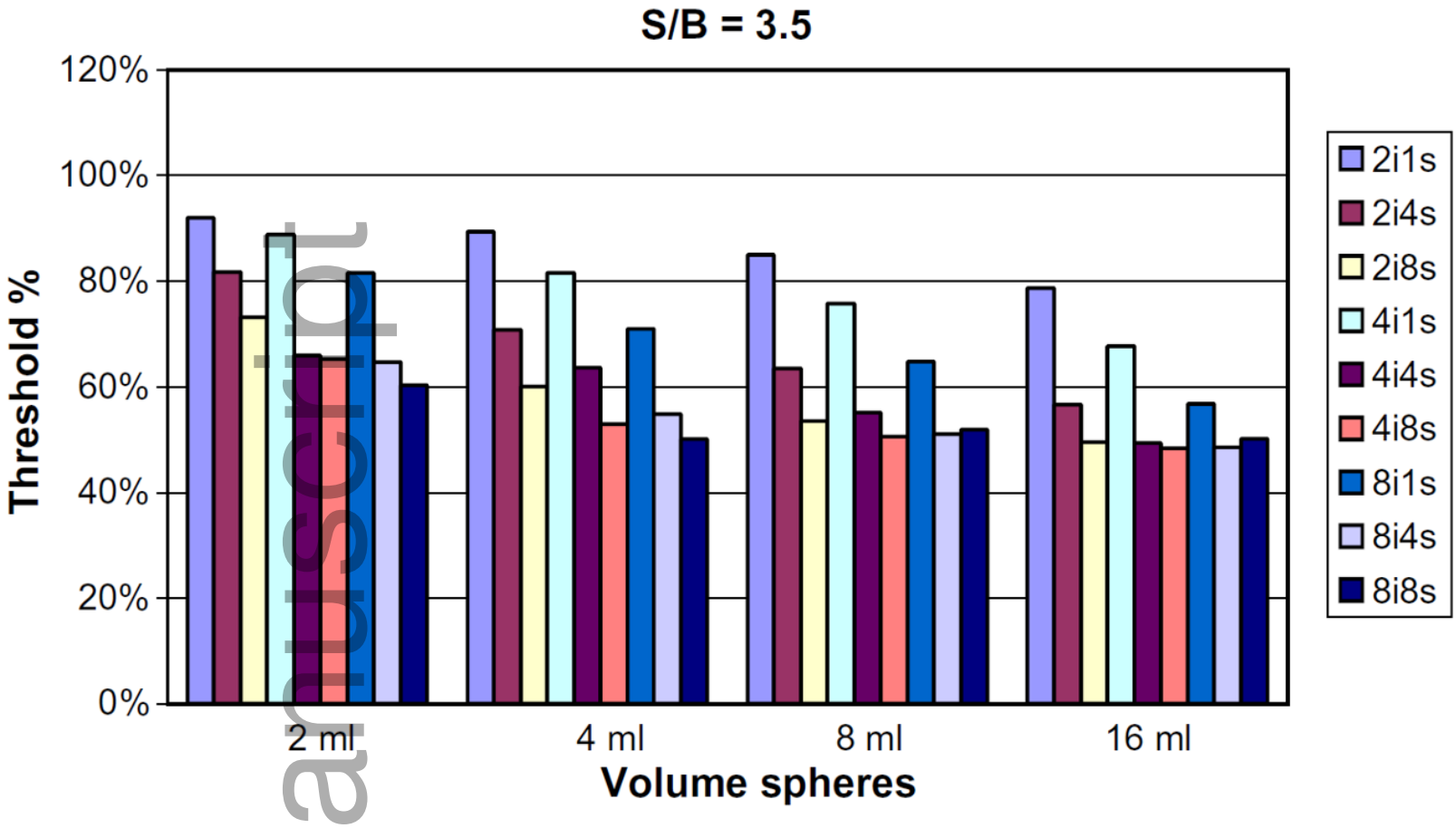
mp_12124_f1d.eps



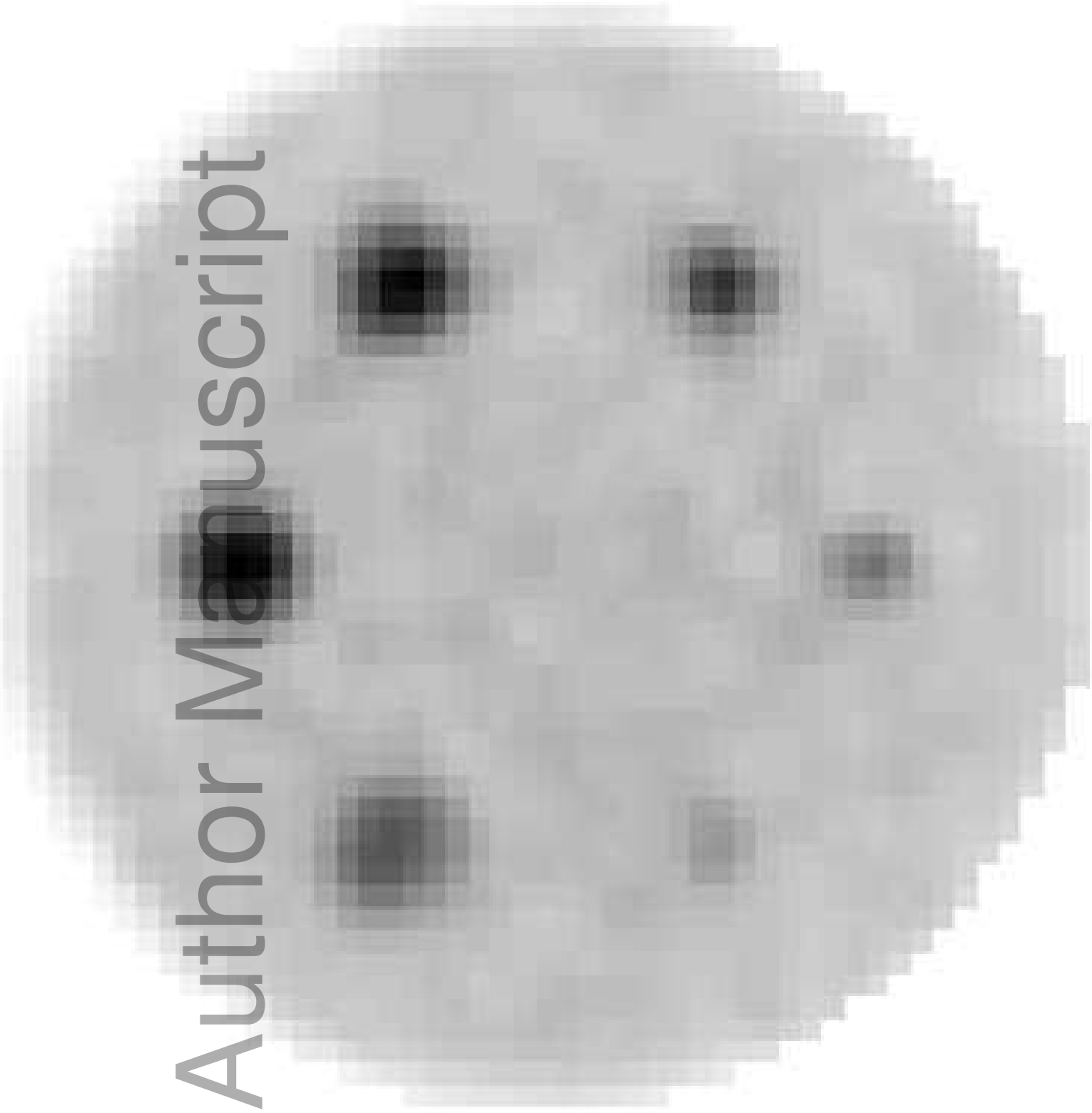
mp_12124_f2.eps



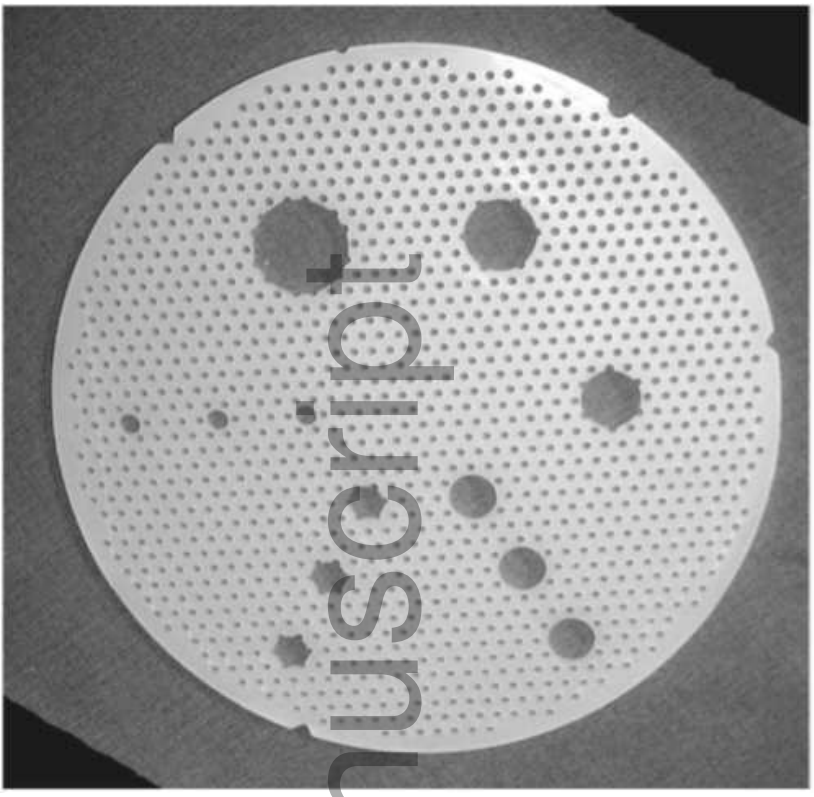
mp_12124_f3a.eps

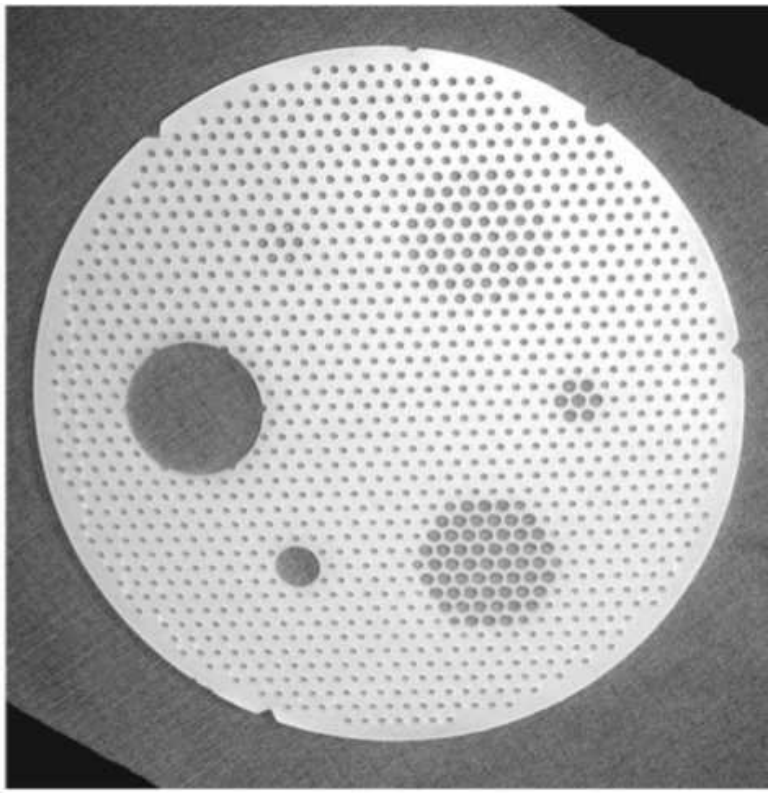


mp_12124_f3b.eps

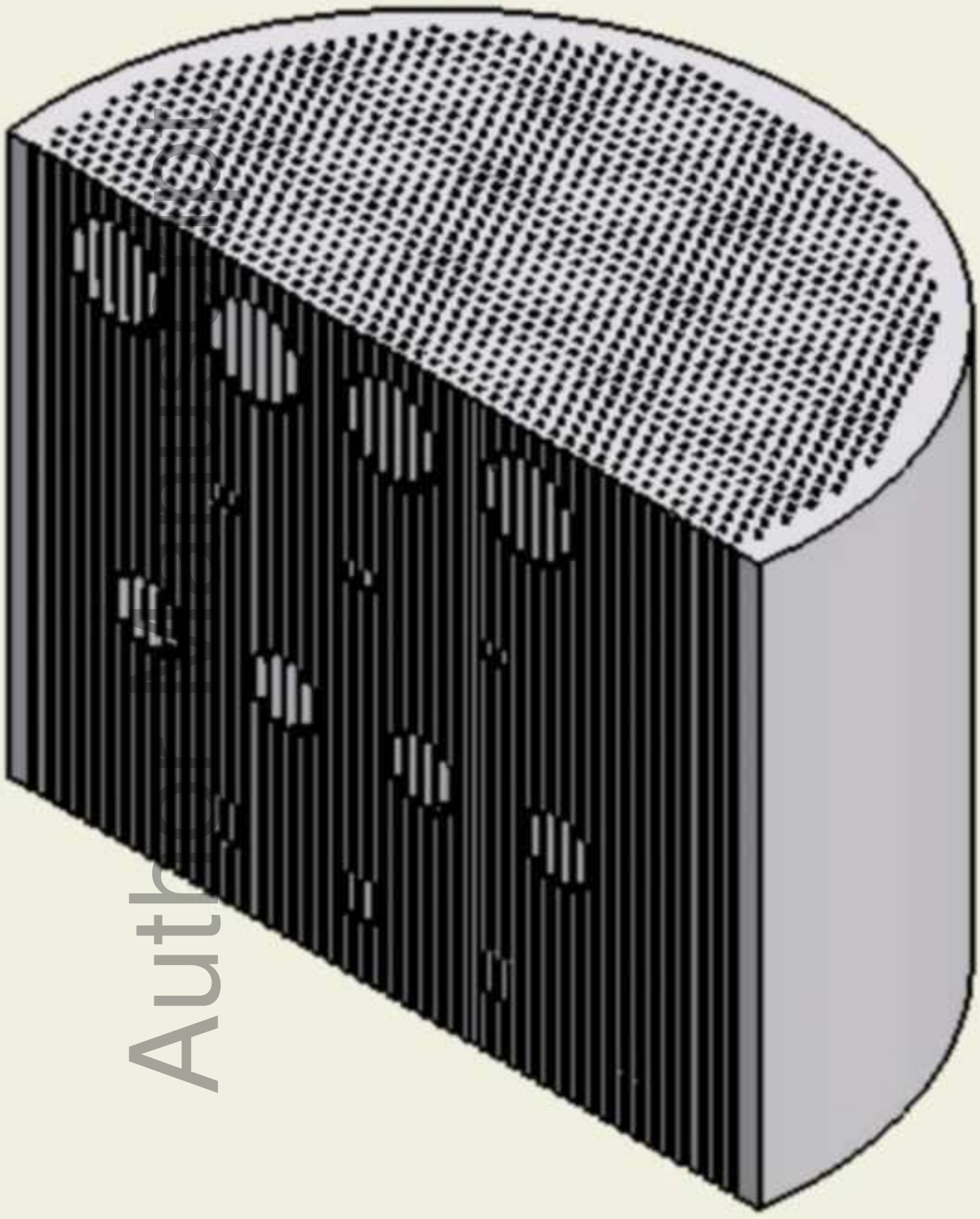


mp_12124_f3c.eps





mp_12124_f3d.eps



mp_12124_f4.eps