Article

Comparative Study With New Accuracy Metrics for Target Volume Contouring in PET Image Guided Radiation Therapy

SHEPHERD, Tony, *et al*.

Abstract

The impact of positron emission tomography (PET) on radiation therapy is held back by poor methods of defining functional volumes of interest. Many new software tools are being proposed for contouring target volumes but the different approaches are not adequately compared and their accuracy is poorly evaluated due to the ill-definition of ground truth. This paper compares the largest cohort to date of established, emerging and proposed PET contouring methods, in terms of accuracy and variability. We emphasize spatial accuracy and present a new metric that addresses the lack of unique ground truth. Thirty methods are used at 13 different institutions to contour functional volumes of interest in clinical PET/CT and a custom-built PET phantom representing typical problems in image guided radiotherapy. Contouring methods are grouped according to algorithmic type, level of interactivity and how they exploit structural information in hybrid images. Experiments reveal benefits of high levels of user interaction, as well as simultaneous visualization of CT images and PET gradients to guide interactive procedures. Method-wise [...]

Reference

SHEPHERD, Tony, *et al*. Comparative Study With New Accuracy Metrics for Target Volume Contouring in PET Image Guided Radiation Therapy. *IEEE Transactions on Medical Imaging* , 2012, vol. 31, no. 11, p. 2006-2024

DOI : 10.1109/TMI.2012.2202322

Comparative Study With New Accuracy Metrics for Target Volume Contouring in PET Image Guided Radiation Therapy

Tony Shepherd*, *Member, IEEE*, Mika Teräs, *Member, IEEE*, Reinhard R. Beichel, *Member, IEEE*, Ronald Boellaard, Michel Bruynooghe, Volker Dicken, Mark J. Gooding, *Member, IEEE*, Peter J. Julyan, John A. Lee, Sébastien Lefèvre, Michael Mix, Valery Naranjo, Xiaodong Wu, Habib Zaidi, *Senior Member, IEEE*, Ziming Zeng, and Heikki Minn

***Abstract*—The impact of positron emission tomography (PET) on radiation therapy is held back by poor methods of defining functional volumes of interest. Many new software tools are being proposed for contouring target volumes but the different approaches are not adequately compared and their accuracy is poorly evaluated due to the ill-definition of ground truth. This paper compares the largest cohort to date of established, emerging and proposed PET contouring methods, in terms of accuracy and variability. We emphasize spatial accuracy and present a new metric that addresses the lack of unique ground truth. Thirty methods are used at 13 different institutions to contour functional volumes of interest in clinical PET/CT and a custom-built PET phantom representing typical problems in image guided radiotherapy. Contouring methods are grouped according to algorithmic type, level of interactivity and how they exploit structural information in hybrid images. Experiments reveal benefits of high levels of user interaction, as well as simultaneous visualization of CT images and PET gradients to guide interactive procedures. Method-wise evaluation identifies the danger of over-automation and the value of prior knowledge built into an algorithm.**

***Index Terms*—Human computer interaction, image segmentation, oncology, performance evaluation, phantoms, positron emission tomography (PET).**

Manuscript received April 26, 2012; accepted May 24, 2012. Date of publication June 04, 2012; date of current version October 26, 2012. *Asterisk indicates corresponding author.*

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

*T. Shepherd is with the Turku PET Centre, Turku University Hospital, 20521 Turku, Finland, and also with the Department of Oncology and Radiotherapy, Turku University Hospital, 20521 Turku, Finland (e-mail: tony.shepherd@tyks.fi).

M. Teräs is with the Turku PET Centre, Turku University Hospital, 20521 Turku, Finland.

R. R. Beichel is with the Department of Electrical and Computer Engineering and Internal Medicine, University of Iowa, Iowa City, IA 52242 USA.

R. Boellaard is with the Department of Nuclear Medicine and PET Research, VU University Medical Centre, 1081 HV Amsterdam, The Netherlands.

M. Bruynooghe is with the SenoCAD Research GmbH, 76185 Karlsruhe, Germany.

V. Dicken is with the Fraunhofer MEVIS—Institute for Medical Image Computing, D-28359 Bremen, Germany.

M. J. Gooding is with the Mirada Medical, OX1 1BY Oxford, U.K.

P. J. Julyan is with the North Western Medical Physics, Christie Hospital NHS Foundation Trust, M20 4BX Manchester, U.K.

J. A. Lee is with the Belgian FNRS and center for Molecular Imaging, Radiotherapy, and Oncology (MIRO), Université Catholique de Louvain, B-1200 Brussels, Belgium.

S. Lefèvre is with the VALORIA Research Laboratory, University of South Brittany, F-56017 Vannes, France.

M. Mix is with the Department of Radiation Oncology, University Freiburg Medical Centre, D-79095 Freiburg, Germany.

V. Naranjo is with the Labhuman Inter-University Research Institute for Bioengineering and Human Centered Technology, 46022 Valencia, Spain.

X. Wu is with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242 USA.

H. Zaidi is with the Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva, Switzerland.

Z. Zeng is with the Department of Computer Science, University of Aberystwyth, SY23 3DB Aberystwyth, U.K..

H. Minn is with the Department of Oncology and Radiotherapy, Turku University Hospital, 20521 Turku, Finland..

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2012.2202322

I. Introduction

POSITRON emission tomography (PET) with the metabolic tracer ^{18}F-FDG is in routine use for cancer diagnosis and treatment planning. Target volume contouring for PET image-guided radiotherapy has received much attention in recent years, driven by the combination of PET with computed tomography (CT) for treatment planning [1], unprecedented accuracy of intensity modulated radiation therapy (IMRT) [2], and ongoing debates [3], [4] over the ability of the standardized uptake value (SUV) to define functional volumes of interest (VOIs) by simple thresholding. Many new methods are still threshold-based, but either automate the choice of SUV threshold specific to an image [5], [6] or apply thresholds to a combination (e.g., ratio) of SUV and an image-specific background value [7], [8]. More segmentation algorithms are entering PET oncology from the field of computer vision [9] including the use of image gradients [10], deformable contour models [11], [12], mutual information in hybrid images [13], [14], and histogram mixture models for heterogeneous regions [15], [16]. The explosion of new PET contouring algorithms calls for constraint in order to steer research in the right direction and avoid so-called *yapetism* (yet another PET image segmentation method) [17]. For this purpose, we identify different approaches and compare their performance.

Previous works to compare contouring methods in PET oncology [18]–[20] do not reflect the wide range of proposed and potential algorithms and fall short of measuring spatial accuracy. Nestle *et al.* [18] compare three threshold-based methods used on PET images of non-small cell lung cancer in terms of the absolute volume of the VOIs, ignoring spatial accuracy of the VOI surface that is important to treatment planning. Greco

et al. [19] compare one manual and three threshold-based segmentation schemes performed on PET images of head-and-neck cancer. This comparison also ignores spatial accuracy, being based on absolute volume of the VOI obtained by manual delineation of complementary CT and magnetic resonance imaging (MRI). Vees *et al.* [20] compare one manual, four threshold-based, one gradient-based and one region-growing method in segmenting PET gliomas and introduce spatial accuracy, measured by volumetric overlap with respect to manual segmentation of complimentary MRI. However, a single manual segmentation can not be considered the unique truth as manual delineation is prone to variability [21], [22].

Outside PET oncology, the society for Medical Image Computing and Computer Assisted Intervention (MICCAI) has run a "challenge" in recent years to compare emerging methods in a range of application areas. Each challenge takes the form of a double-blind experiment, whereby different methods are applied by their developers on common test-data and the results analyzed together objectively. In 2008, two examples of pathological segmentation involved multiple sclerosis lesions in MRI [23] and liver tumors in CT [24]. These tests involved 9 and 10 segmentation algorithms respectively, and evaluated their accuracy using a combination of the Dice similarity coefficient [25] and Hausdorff distance [26] with respect to a single manual delineation of each VOI. In 2009 and 2010, the challenges were to segment the prostate in MRI [27] and parotid in CT [28]. These compared 2 and 10 segmentation methods respectively, each using a combination of various overlap and distance measures to evaluate accuracy with respect to a single manual ground truth per VOI. The MICCAI challenges have had a major impact on segmentation research in their respective application areas, but this type of large-scale, double-blind study has not previously been applied to PET target volume delineation for therapeutic radiation oncology, and the examples above are limited by their dependence upon a single manual delineation to define ground truth of each VOI.

This paper reports on the design and results of a large-scale, multi-center, double-blind experiment to compare the accuracy of 30 established and emerging methods of VOI contouring in PET oncology. The study uses a new, probabilistic accuracy metric [29] that removes the assumption of unique ground truth, along with standard metrics of Dice similarity coefficient, Hausdorff distance and composite metrics. We use both a new tumor phantom [29] and patient images of head-and-neck cancer imaged by hybrid PET/CT. Experiments first validate the new tumor phantom and accuracy metric, then compare conceptual approaches to PET contouring by grouping methods according to how they exploit CT information in hybrid images, the level of user interaction and 10 distinct algorithm types. This grouping leads to conclusions about general approaches to segmentation, also relevant to other tools not tested here. Regarding the role of CT, conflicting reports in the literature further motivate the present experiments: while some authors found that PET tumor discrimination improves when incorporating CT visually [30] or numerically [31], others report on the detrimental effect of visualizing CT on accuracy [32] and inter/intra-observer variability [21], [22]. Further experiments directly evaluate each method in terms of accuracy and, where available, inter-/intra operator variability. Due to the large number of contouring methods, full details of their individual accuracies and all statistically significant differences are provided in the supplementary material and summarized in this paper.

TABLE I
THE 30 CONTOURING METHODS AND THEIR ATTRIBUTES

method	team	type	interactivity					CT use		
			max	high	mid	low	none	high	low	none
PL^a	01	PL					▲			■
WS^a	02	WS				▲				■
PL^b	03	PL			▲				■	
PL^c					▲				■	
PL^d					▲				■	
T2^a		T2					▲			■
MD^a	04	MD	▲							■
T4^a		T4				▲				■
T4^b						▲				■
T4^c						▲				■
MD$^b{}_{1,2}$	05	MD	▲							■
RG^a		RG		▲						■
HB	06	HB				▲		■		
WS^b	07	WS				▲				■
T1^a	08	T1				▲				■
T1^b						▲				■
T2^b		T2				▲				■
T2^c						▲				■
RG$^b{}_{1,2}$	09	RG		▲					■	
RG$^c{}_{1,2}$					▲					■
PL^e	10	PL					▲			■
PL^f						▲				■
GR	11	GR			▲					■
MD^c	12	MD	▲						■	
T1^c		T1					▲			■
T3^a		T3				▲				■
T3^b						▲				■
T2^d		T2					▲			■
T2^e					▲					■
PL^g	13	PL					▲			■

The rest of this paper is organized as follows. Section II describes all contouring algorithms and their groupings. Section III presents the new accuracy metric and describes phantom and patient images and VOIs. Experiments in Section IV evaluate the phantom and accuracy metric and compare segmentation methods as grouped and individually. Section V discusses specific findings about manual practices and the types of automation and prior knowledge built into contouring and Section VI gives conclusions and recommendations for future research in PET-based contouring methodology for image-guided radiation therapy.

II. CONTOURING METHODS

Thirteen contouring "teams" took part in the experiment. We identify 30 distinct "methods," where each is a unique combination of team and algorithm. Table I presents the methods along with labels (first column) used to identify them hereafter. Some teams used more than one contouring algorithm and some well-established algorithms such as thresholding were used by

more than one team, with different definitions of the quantity and its threshold. Methods are grouped according to algorithm type and distinguished by their level of dependence upon the user (Section II-B) and CT data (Section II-C) in the case of patient images. Contouring by methods MD^b, RG^b, and RG^c was repeated by two users in the respective teams, denoted by subscripts 1 and 2, and the corresponding segmentations are treated separately in our experiments.

Some of the methods are well known for PET segmentation while others are recently proposed. Of the recently proposed methods, some were developed specifically for PET segmentation (e.g., GR, $\mathrm{T2}^d$, and PL^g) while some were adapted and optimized for PET tumor contouring for the purpose of this study. The study actively sought new methods, developed or newly adapted for PET tumors, as their strengths and weaknesses will inform current research that aims to refine or replace state of the art tools, whether those tools are included here or not. Many of the algorithms considered operate on standardized uptake values (SUVs), whereby PET voxel intensity I is rescaled as $\mathrm{SUV} = I \times (\beta/a_{\mathrm{in}})$ to standardize with respect to initial activity a_{in} of the tracer in $Bq\ ml^{-1}$ and patient mass β in grams [33]. The SUV transformation only affects segmentation by fixed thresholding while methods that normalize with respect to a reference value in the image or apply thresholds at a percentage of the maximum value are invariant to the SUV transformation.

A. Method Types and Descriptions

Manual delineation methods (MD) use a computer mouse to delineate a VOI slice-by-slice, and differ by the modes of visualization such as overlaying structural or gradient images and intensity windowing. $\mathbf{MD}^a$ is performed by a board certified radiation oncologist and nuclear medicine physician, who has over a decade of research and clinical experience in PET-based radiotherapy planning. $\mathbf{MD}^b$ is performed by two independent, experienced physicians viewing only PET image data. For each dataset, the grey-value window and level were manually adjusted. $\mathbf{MD}^c$ performed on the PET images by a nuclear medicine physicist who used visual aids derived from the original PET: intensity thresholds, both for the PET and the PET image-gradient, were set interactively for the purpose of visual guidance.

Thresholding methods (T1–T4) are divided into four types according to whether the threshold is applied to signal (T1 and T2) or a combination of signal and background intensity (T3 and T4) and whether the threshold value is chosen *a priori*, based on recommendations in the literature or the team's own experience (T1 and T3) or chosen for each image, either automatically according to spatial criteria or visually by the user's judgement (T2 and T4). Without loss of generalization the threshold value may be absolute or percentage (e.g., of peak) intensity or SUV. $\mathbf{T1}^a$ and $\mathbf{T1}^b$ employ the widely used cutoff values of 2.5 SUV and 40% of the maximum in the VOI, as used for lung tumor segmentation in [34] and [35], respectively. Method $\mathrm{T1}^a$ is the only method of all in Table I that is directly affected by the conversion from raw PET intensity to SUVs. The maximum SUV used by method $\mathrm{T1}^b$ was taken from inside the VOI defined by $\mathrm{T1}^a$. To calculate SUV for the phantom image, where patient weight β is unavailable, all voxel values were rescaled with respect to a value of unity at one end of the phantom where intensity is near uniform, causing method $\mathrm{T1}^a$ to fail for phantom scan 2 as the maximum was below 2.5 for both VOIs. $\mathbf{T1}^c$ applies a threshold at 50% of the maximum SUV. Method $\mathbf{T2}^a$ is the thresholding scheme of [6], which automatically finds the optimum relative threshold level (RTL) based an estimate of the true absolute volume of the VOI in the image. The RTL is relative to background intensity, where background voxels are first labelled automatically by clustering. An initial VOI is estimated by a threshold of 40% RTL, and its maximum diameter is determined. The RTL is then adjusted iteratively until the absolute volume of the VOI matches that of a sphere of the same diameter, convolved with the point-spread function (PSF) of the imaging device, estimated automatically from the image. Methods $\mathbf{T2}^b$ and $\mathbf{T2}^c$ automatically define thresholds according to different criteria. They both use the results of method $\mathrm{T1}^a$ as an initial VOI, and define local background voxels by dilation. Method $\mathrm{T2}^b$ uses two successive dilations and labels the voxels in the second dilation as background. The auto-threshold is then defined as three standard deviations above the mean intensity in this background sample. Method $\mathrm{T2}^c$ uses a single dilation to define the background and finds the threshold that minimizes the within-class variance between VOI and background using the optimization technique in [36]. Finally, method $\mathrm{T2}^c$ applies a closing operation to eliminate any holes within the VOI, which may also have the effect of smoothing the boundary. Method $\mathbf{T2}^d$ finds the RTL using the method of [6] in common with method $\mathrm{T2}^a$ but with different parameters and initialization. Method $\mathrm{T2}^d$ assumes a PSF of 7 mm full-width at half-maximum (FWHM) rather than estimating this value from the image. The RTL was initialized with background defined by a manual bounding box rather than clustering and foreground defined by method $\mathrm{T3}^a$ with a 50% threshold rather than 40% RTL. Adaptive thresholding method $\mathbf{T2}^e$ starts with a manually defined bounding box then defines the VOI by the iso-contour at a percentage of the maximum value within the bounding box. Methods $\mathbf{T3}^a$ and $\mathbf{T3}^b$ are similar to $\mathrm{T1}^c$, but incorporate local background intensity calculated by a method equivalent to that Daisne *et al.* [7]. A threshold value is then 41% and 50% of the maximum plus background value, respectively. Method $\mathbf{T4}^a$ is an automatic SUV-thresholding method implemented in the "Rover" software [37]. After defining a search area that encloses the VOI, the user provides an initial threshold which is adjusted in two steps of an iterative process. The first step estimates background intensity I_b from the average intensity over those voxels that are below the threshold *and* within a minimum distance of the VOI (above the threshold). The second step redefines the VOI by a new threshold at 39% of the difference $I_{\max} - I_b$, where $I_{\max}$ is the maximum intensity in the VOI. Methods $\mathbf{T4}^b$ and $\mathbf{T4}^c$ use the source-to-background algorithm in [8]. The user first defines a background region specific to the given image, then uses parameters a and b to define the threshold $t = a\mu_{\mathrm{VOI}} + b\mu_{\mathrm{BG}}$, where $\mu_{\mathrm{VOI}}+$ and μ_{BG} are the mean SUV in the VOI and background respectively. The parameters are found in a calibration procedure by scanning spherical phantom VOIs of known volume. As this calibration was not performed

for the particular scanner used in the present experiments (GE Discovery), methods $\mathrm{T4}^b$ and $\mathrm{T4}^c$ use parameters previously obtained for Gemini and Biograph PET systems respectively.

Region growing methods (RG) use variants of the classical algorithm in [38], which begins at a "seed" voxel in the VOI and agglomerates connected voxels until no more satisfy criteria based on intensity. In $\mathbf{RG}^a$, the user defines a bounding sphere centred on the VOI, defining both the seed at the center of the sphere and a hard constraint at the sphere surface to avoid leakage into other structures. The acceptance criterion is an interactively adjustable threshold and the final VOI is manually modified in individual slices if needed. Methods $\mathbf{RG}^b$ and $\mathbf{RG}^c$ use the region growing tool in Mirada XD (Mirada Medical, Oxford, U.K.) with seed point location and acceptance threshold defined by the user. In RG^b only, the results are manually post-edited using the "adaptive brush" tool available in Mirada XD. This 3-D painting tool adapts in shape to the underlying image. Also in method RG^b only, CT images were fused with PET for visualization and the information used to modify the regions to exclude airways and unaffected bone.

Watershed methods (WS) use variants of the classical algorithm in [39]. The common analogy pictures a gradient-filtered image as a "relief map" and defines a VOI as one or more pools, created and merged by flooding a region with water. Method $\mathbf{WS}^a$, adapted from the algorithm in [40] for segmenting natural color images and remote-sensing images, makes use of the content as well as the location of user-defined markers. A single marker for each VOI (3×3 or 5×5 pixels depending on VOI size) is used along with a background region to train a fuzzy classification procedure where each voxel is described by a texture feature vector. Classification maps are combined with image gradient and the familiar "flooding" procedure is adapted for the case of multiple surfaces. Neither the method nor the user were specialized in medical imaging. Method $\mathbf{WS}^b$, similar way to that in [41], uses two procedures to overcome problems associated with local minima in image gradient. First, viscosity is added to the watershed, which closes gaps in the edge-map. Second, a set of internal and external markers are identified, indicating the VOI and background. After initial markers are identified in one slice by the user, markers are placed automatically in successive slices, terminating when the next slice is deemed no longer to contain the VOI according to a large drop in the "energy," governed by area and intensity, of the segmented cross section. If necessary, the user interactively overrides the automatic marker placement.

Pipeline methods (PL) are more complex, multi-step algorithms that combine elements of thresholding, region growing, watershed, morphological operations and techniques in [42], [43], [15]. Method $\mathbf{PL}^a$ is a deformable contour model adapted from white matter lesion segmentation in brain MRI. The main steps use a region-scalable fitting model [44] and a global standard convex scheme [45] in energy minimization based on the "Split Bregman" technique in [42]. Methods $\mathbf{PL}^b$–$\mathbf{PL}^d$ are variants of the "Smart Opening" algorithm, adapted for PET from the tool in [43] for segmenting lung nodules in CT data. In contrast to CT lung lesions, the threshold used in region growing can not be set *a priori* and is instead obtained from the image interactively. Method PL^b was used by an operator with limited PET experience. The user of method PL^c had more PET experience and, to aid selection of boundary points close to steep PET gradients, also viewed an overlay of local maxima in the edge-map of the PET image. Finally, method PL^d took the results of method PL^c and performed extra processing by dilation, identification of local gradient maxima in the dilated region, and thresholding the gradient at the median of these local maxima. Methods $\mathbf{PL}^e$ and $\mathbf{PL}^f$ use the so-called "poly-segmentation" algorithm without and with post editing respectively. PL^e is based on a multi-resolution approach, which segments small lesions using recursive thresholding and combines three segmentation algorithms for larger lesions. First, the watershed transform provides an initial segmentation. Second, an iterative procedure improves the segmentation by adaptive thresholding that uses the image statistics. Third, a region growing method based on regional statistics is used. The interactive variant (PL^f) uses a fast interactive tool for watershed-based subregion merging. This intervention is only necessary in at most two slices per VOI. Method $\mathbf{PL}^g$ is a new fuzzy segmentation technique for noisy and low resolution oncological PET images. PET images are first smoothed using a nonlinear anisotropic diffusion filter and added as a second input to the fuzzy C-means (FCM) algorithm to incorporate spatial information. Thereafter, the algorithm integrates the *à trous* wavelet transform in the standard FCM algorithm to handle heterogeneous tracer uptake in lesions [15].

The **gradient based method (GR)** is the novel edge-finding method in [10], designed to overcome the low signal-to-noise ratio and poor spatial resolution of PET images. As resolution blur distorts image features such as iso-contours and gradient intensity peaks, the method combines edge restoration methods with subsequent edge detection. Edge restoration goes through two successive steps, namely edge-preserving denoising and deblurring with a deconvolution algorithm that takes into account the resolution of a given PET device. Edge-preserving denoising is achieved by bilateral filtering and a variance-stabilizing transform [46]. Segmentation is finally performed by the watershed transform applied after computation of the gradient magnitude. Over-segmentation is addressed with a hierarchical clustering of the watersheds, according to their average tracer uptake. This produces a dendrogram (or tree-diagram) in which the user selects the branch corresponding to the tumor or target. User intervention is usually straightforward, unless the uptake difference between the target and the background is very low.

The **Hybrid method (HB)** is the multi-spectral algorithm in [14], adapted for PET/CT. This graph-based algorithm exploits the superior contrast of PET and the superior spatial resolution of CT. The algorithm is formulated as a Markov random field (MRF) optimization problem [47]. This incorporates an energy term in the objective function that penalizes the spatial difference between PET and CT segmentation.

B. Level of Interactivity

Levels of interactivity are defined on an ordinal scale of "max," "high," "mid," "low," and "none," where "max" and "none" refer to fully manual and fully automatic methods, respectively. Methods with a "high" level involve user initialization, which locates the VOI and/or representative voxels,

TABLE II
DETAILS OF PHANTOM AND PATIENT PET/CT IMAGES

Image type	PET (18F FDG)					CT			
	frame length (min)	width/height (pixels)	depth (slices)	pixel size (mm)	slice depth (mm)	width/height (pixels)	depth (slices)	pixel size (mm)	slice depth (mm)
phantom	10.0	256	47	1.17×1.17	3.27	512	47	0.59×0.59	3.75
patient	3.0	256	33,37	2.73×2.73	3.27	512	42,47	0.98×0.98	1.37

as well as run-time parameter adjustment and post-editing of the contours. "Mid"-level interactions involve user-initialization and either run-time parameter adjustment or other run-time information such as wrongly included/excluded voxels. "Low"-level interaction refers to initialization or minimal procedures to restart an algorithm with new information such as an additional mouse-click in the VOI.

C. Level of CT Use

We define the levels at which contouring methods exploit CT information in hybrid patient images as "high," "low," or "none," where "high" refers to *numerical* use of CT together with PET in calculations. The "low" group makes *visual* use of CT images to guide manual delineation, postediting or other interactions in semi-automatic methods. The "none" group refers to cases where CT is not used, or is viewed incidentally but has no influence on contouring as the algorithm is fully automatic. None of the methods operated on CT images alone.

III. EXPERIMENTAL METHODS

A. Images

We use two images of a new tumor phantom [29], manufactured for this study and two clinical PET images of different head-and-neck cancer patients. The phantom images are available online [48], along with ground truth sets described in Section III-C. All imaging used the metabolic tracer ^{18}F-Fluorodeoxyglucose (FDG) and a hybrid PET/CT scanner (GE Discovery), but CT images from phantom scans were omitted from the test set. Table II gives more details of each image type. The tumor phantom contains glass compartments of irregular shapes shown in Fig. 1(top row), mimicking real radiotherapy target volumes. The tumor compartment (a) has branches to recreate the more complex topology of some tumors. This and the nodal chain compartment (b) are based on cancer of the oral cavity and lymph node metastasis respectively, manually segmented from PET images of two head and neck cancer patients and formed by glass blowing. The phantom compartments and surrounding container were filled with low concentrations of FDG and scanned by a hybrid device (1, middle and bottom rows). Four phantom VOIs result from scans 1 and 2, with increasing signal to background ratio achieved by increasing FDG concentration in the VOIs. Details of the four phantom VOIs are given in the first four rows of Table III. Fig. 2 shows the phantom VOIs from scan 1, confirming qualitatively the spatial and radiometric agreement between phantom and patient VOIs.

For patient images, head and neck cancer was chosen as it poses particular challenges to PET-based treatment planning due to the many nearby organs at risk (placing extra demand on GTV contouring accuracy), the heterogeneity of tumor tissue

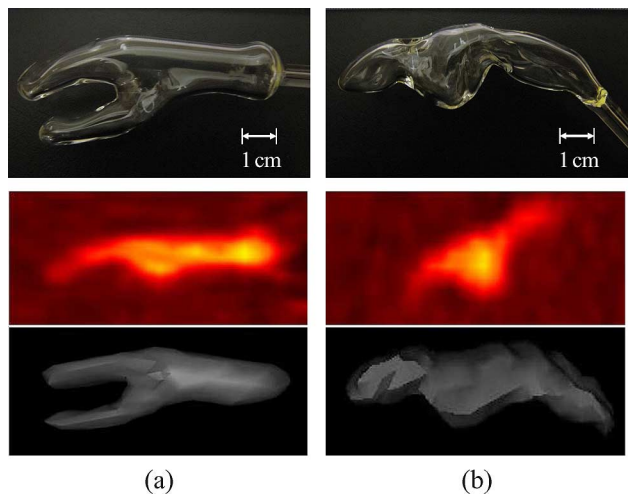

Fig. 1. (a) Tumor and (b) nodal chain VOIs of the phantom. *Top*: Digital photographs of glass compartments. *Middle*: PET images from scan 1 (sagittal view). *Bottom*: 3-D surface view from an arbitrary threshold of simultaneous CT, lying within the glass wall.

TABLE III
PROPERTIES OF VOI AND BACKGROUND (BG) DATA (VOLUMES IN CM^3 ARE ESTIMATED AS IN SECTION III-C)

VOI	image	initial activity (kBq ml^{-1})	volume (cm^3)	source of ground truth
tumour	phantom scan 1	8.7 (VOI)	6.71	thresholds of simultaneous CT image
node		4.9 (BG)	7.45	
tumour	phantom scan 2	10.7 (VOI)	6.71	
node		2.7 (BG)	7.45	
tumour	patient 1	2.4 ×10^5	35.00	multiple expert delineations
node			2.54	
tumour	patient 2	3.6 ×10^5	2.35	

and the common occurrence of lymph node metastasis. A large tumor of the oral cavity and a small tumor of the larynx were selected from two different patients, along with a metastatic lymph node in the first patient (Fig. 3). These target volumes were chosen as they were histologically proven and have a range of sizes, anatomical locations/surroundings, and target types (tumor and metastasis). Details of the three patient VOIs are given in the last three rows of Table III.

B. Contouring

With the exception of the hybrid method (HB) that does not apply to the PET-only phantom data, all methods contoured all seven VOIs. In the case of patient VOIs, participants had the option of using CT as well as PET, and were instructed to contour the gross tumor volume (GTV) and metastatic tissue of tumors and lymph node, respectively. All contouring methods were used at the sites of the respective teams using their own

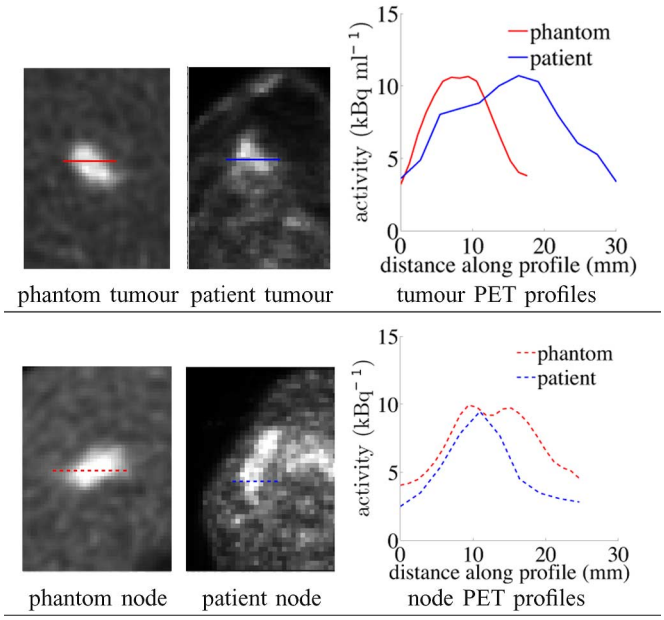

Fig. 2. Axial PET images of phantom and real tumor (top) and lymph node (bottom) VOIs with profile lines traversing each VOI. Plots on the right show the image intensity profiles sampled from each image pair.

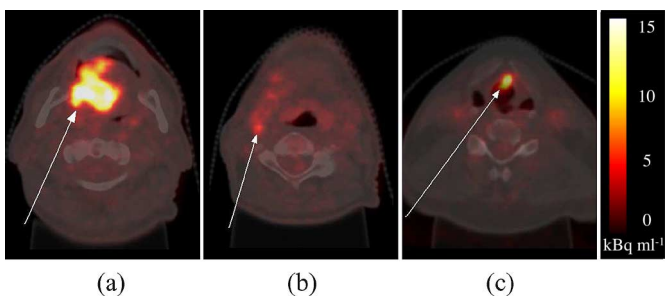

Fig. 3. Axial neck slices of ^{18}F-FDG PET images overlain on simultaneous CT. (a) and (b) Oral cavity tumor and lymph node metastasis in patient 1. (c) Laryngeal tumor in patient 2.

software and workstations. Screen-shots of each VOI were provided in axial, sagittal, and coronal views, with approximate centers indicated by cross-hairs and their voxel coordinates provided to remove any ambiguity regarding the ordering of axes and direction of increasing indexes. No other form of ground truth was provided. Teams were free to refine their algorithms and practice segmentation before accepting final contours. This practicing stage was done without any knowledge of ground truth and is considered normal practice. Any contouring results with sub-voxel precision were down-sampled to the resolution of the PET image grid and any results in millimeters were converted to voxel indexes. Finally, all contouring results were duplicated to represent VOIs first by the voxels on their surface, and second by masks of the solid VOI *including* the surface voxels. These two representations were used in surface-based and volume-based contour evaluation, respectively.

C. Contouring Evaluation

Accuracy measurement generally compares the contour being evaluated, which we denote $\mathcal{C}$, with some notion of ground truth, denoted $\mathcal{GT}$. We use a new probabilistic metric

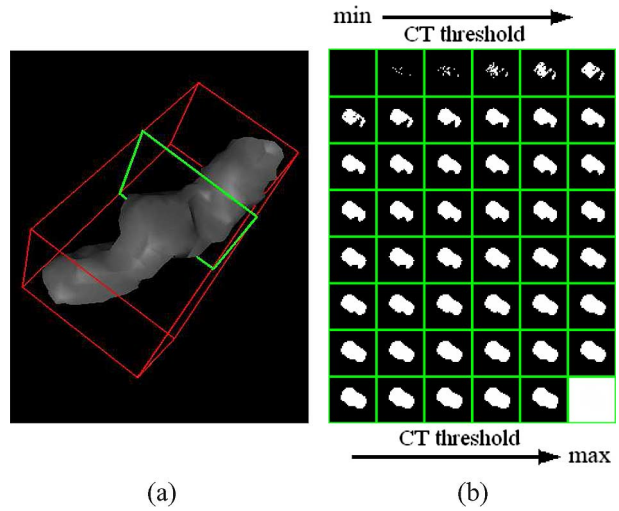

Fig. 4. (a) 3-D visualization of phantom VOI from CT thresholded at a density near the internal glass surface. (b) Arbitrary ground truth masks of the axial cross section in (a), from 50 thresholds of HU.

[29] denoted AUC', as well as a variant of the Hausdorff distance [26] denoted HD' and the standard metric of Dice similarity coefficient [25] (DSC). AUC' and HD' are standardized to the range $0 \ldots 1$ so that they can be easily combined or compared with DSC and other accuracy metrics occupying this range [49]–[51]. Treated separately, AUC', HD', and DSC allow performance evaluation with and without the assumption of unique ground truth, and in terms of both volumetric agreement (AUC' and DSC) and surface-displacement (HD') with respect to ground truth.

AUC' is a probabilistic metric based on receiver operating characteristic (ROC) analysis, in a scheme we call *inverse-ROC* (I-ROC). The I-ROC method removes the assumption of unique ground truth, instead using a set of p arbitrary ground truth definitions $\{\mathcal{GT}_i\}, i \in \{1 \ldots p\}$ for each VOI. While uniquely correct ground truth in the space of the PET image would allow deterministic and arguably superior accuracy evaluation, the I-ROC method is proposed for the case here, and perhaps all cases except numerical phantoms, where such truth is not attainable. The theoretical background of I-ROC is given in Appendix A and shows that the area under the curve (AUC) gives a probabilistic measure of accuracy provided that the arbitrary set can be ordered by increasing volume and share the topology and general form of the (unknown) true surface. The power of AUC' as an accuracy metric also relies on the ability to incorporate the best available knowledge of ground truth into the arbitrary set. This is done for phantom and patient VOIs as follows.

For phantom VOIs, the ground truth set is obtained by incrementing a threshold of Hounsfield units (HU) in the CT data from hybrid imaging (Fig. 4). Masks acquired for all CT slices in the following steps.

1) Reconstruct/down-sample the CT image to the same pixel grid as the PET image.
2) Define a bounding box in the CT image that completely encloses the glass VOI as well as $\mathcal{C}$.
3) Threshold the CT image at a value HU_i.

4) Treat all pixels below this value as being "liquid" and all above it as "glass."
5) Label all "liquid" pixels that are *inside* the VOI as positive, but ignore pixels outside the VOI.
6) Repeat for p thresholds $\mathrm{HU}_i, i \in \{1 \ldots p\}$ between natural limits $\mathrm{HU}_{\min}$ and $\mathrm{HU}_{\max}$.

This ground truth set is guaranteed to pass through the internal surface of the glass compartment and exploits the inherent uncertainty due to partial volume effects in CT. It follows from derivations in Appendix A.2–3 that AUC is equal to the probability that a voxel drawn at random from below the unknown CT threshold at the internal glass surface, lies inside the contour $\mathcal{C}$ being evaluated.

For patient VOIs, the ground truth set is the union of an increasing number of expert manual delineations. Experts contoured GTV and node metastasis on PET visualized with co-registered CT. In the absence of histological resection, we assume that the best source of ground truth information is manual PET segmentation by human experts at the imaging site, who have experience of imaging the particular tumor-type and access to extra information such as tumor stage, treatment follow-up and biopsy where available. However, we take the view that no single manual segmentation provides the unique ground truth, which therefore remains unknown. In total, three delineated each VOI on two occasions (denoted $N_{\text{experts}} = 3$ and $N_{\text{occasions}} = 2$) with at least a week in between. The resulting set of $p = N_{\text{experts}} \times N_{\text{occasions}}$ ground truth estimates were acquired to satisfy the requirements in Appendix A.3 as follows.

1) Define a bounding box in the CT image that completely encloses all $N_{\text{experts}} \times N_{\text{occasions}}$ manual segmentations $\{\mathcal{GT}_i\}$ and the contour $\mathcal{C}$ being evaluated.
2) Order the segmentations $\{\mathcal{GT}_i\}$ by absolute volume in cm^3.
3) Use the smallest segmentation as $\mathcal{GT}_2$.
4) form a new VOI from the union of the smallest and the next largest VOI in the set and use this as $\mathcal{GT}_3$.
5) Repeat until the largest VOI in the set has been used in the union of all $N_{\text{experts}} \times N_{\text{occasions}}$ VOIs.
6) Create homogeneous masks for $\mathcal{GT}_1$ and $\mathcal{GT}_p$, having all negative and all positive contents, respectively.

The patient ground truth set encodes uncertainty from inter-/intra-expert variability in manual delineation and AUC is the probability that a voxel drawn at random from the unknown manual contour at the true VOI surface, lies inside the contour $\mathcal{C}$ being evaluated. Finally, we rescale AUC to the range $\{0 \ldots 1\}$ by

$$\mathrm{AUC}' = \frac{\mathrm{AUC} - 0.5}{0.5}$$
$$0 \le \mathrm{AUC}' \le 1 = \text{maximum accuracy.} \quad (1)$$

Reference surfaces that profess to give the unique ground truth are required to measure the Hausdorff distance and Dice similarity. We obtain the "best guess" of the unique ground truth, denoted $\mathcal{GT}^*$ from the sets of ground truth definitions introduced above. For each phantom VOI we select the CT threshold having the closest internal volume in cm^3 to an independent estimate. This estimate is the mean of three repeated measurements of the volume of liquid contained by each glass compartment. For patient VOIs, $\mathcal{GT}^*$ is the union mask that has the closest absolute volume to the mean of all $N_{\text{experts}} \times N_{\text{occasions}}$ raw expert manual delineations.

TABLE IV
COMPOSITE ACCURACY METRICS THAT CONDENSE RANKING AND SIGNIFICANCE INFORMATION

n(n.s.d): the number between 0 and 4, of accuracy metrics AUC', DSC, HD and A*, for which a method scores an accuracy of no significant difference (n.s.d) from the best method according to that accuracy
n(>μ+σ): the number between 0 and 4, of accuracy metrics AUC', DSC, HD and A*, for which a method scores more than one standard deviation (σ) above the mean (μ) of that score achieved by all 32 methods (33 in the case of patient VOIs only)
median rank: the median, calculated over the 4 accuracy metrics, of the ranking of that method in the list of all 32 methods (33 for patient VOIs only) ordered by increasing accuracy

HD' first uses the reference surface $\mathcal{GT}^*$ to calculate the Hausdorff distance HD, being the maximum for any point on the surface $\mathcal{C}$ of the minimum distances from that point to any point on the surface of $\mathcal{GT}^*$. We then normalize HD with respect to a length scale r and subtract the result from 1

$$\mathrm{HD}' = \frac{1 - \min(\mathrm{HD}, r)}{r}$$
$$0 \le \mathrm{HD}' \le 1 = \text{maximum accuracy} \quad (2)$$

where $r = \sqrt[3]{(3/(4\pi))\mathrm{vol}(\mathcal{GT}^*)}$ is the radius of a sphere having the same volume as $\mathcal{GT}^*$ denoted $\mathrm{vol}(\mathcal{GT}^*)$. Equation (2) transforms HD to the desired range with 1 indicating maximum accuracy.

DSC also uses the reference surface $\mathcal{GT}^*$ and is calculated by

$$\mathrm{DSC} = \frac{2N_{\mathcal{C} \cap \mathcal{GT}^*}}{N_{\mathcal{C}} + N_{\mathcal{GT}^*}}$$
$$0 \le \mathrm{DSC} \le 1 = \text{maximum accuracy} \quad (3)$$

where N_v denotes the number of voxels in volume v defined by contours or their intersect.

Composite metrics are also used. First, we calculate a synthetic accuracy metric from the weighted sum

$$\mathrm{A}^* = 0.5\ \mathrm{AUC}' + 0.25\ \mathrm{DSC} + 0.25\ \mathrm{HD}' \quad (4)$$

which, in the absence of definitive proof of their relative power, assigns equal weighting to the benefits of the probabilistic (AUC') and deterministic approaches (DSC and HD'). By complementing AUC' with the terms using the best guess of unique ground truth, A* penalizes deviation from the "true" absolute volume, which is measured with greater confidence than spatial truth. Second, we create composite metrics based on the relative accuracy within the set of all methods. Three composite metrics are defined in Table IV and justified as follows. Metric n(n.s.d) favors a segmentation tool that is as good as the most accurate in a statistical sense and, in the presence of false significances due to the multiple comparison effect, gives more conservative rather than falsely high scores. Metric $\mathrm{n}(> \mu + \sigma)$ favors the methods in the positive tails of the population, which is irrespective of multiple comparison effects. The rank-based metric is also immune to the multiple comparison effect and we use the median rather than mean rank

to avoid misleading results for a method that ranks highly in only one of the metrics AUC', DSC, HD, and A*, considered an outlier.

Intra-operator variability was measured by the raw Hausdorff distance in millimeters between the first and second segmentation result from repeated contouring (no ground truth necessary). However, this was only done for some contouring methods. For fully automatic methods, variability is zero by design and was not explicitly measured. Of the remaining semi-automatic and manual methods, 11 were used twice by the same operator: MD_1^b, MD_2^b, RG^a, HB, WS^b, RG_1^b, RG_2^b, RG_1^c, RG_2^c, GR, and MD^c and for these we measure the intra-operator variability which allows extra, direct comparisons in Section IV-E.

IV. Experiments

This section motivates the use of the new phantom and accuracy metric (IV-A), then investigates contouring accuracy by comparing the pooled accuracy of methods grouped according to their use of CT data (Section IV-B), level of user interactivity (Section IV-C), and algorithm type (Section IV-D). Section IV-E evaluates methods individually, using condensed accuracy metrics in Table IV. With the inclusion of repeated contouring by methods MD^b, RG^b, and RG^c by a second operator, there are a total of $\mathrm{n} = 33$ segmentations of each VOI, with the exception of phantom VOIs where $\mathrm{n} = 32$ by the exclusion of method HB. Also, method $\mathrm{T1}^a$ failed to recover phantom VOIs in scan 1 as no voxels were above the predefined threshold. In this case a value of zero accuracy is recorded for two out of four phantom VOIs.

A. Phantom and AUC'

This experiment investigates the ability of the phantom to pose a realistic challenge to PET contouring, by testing the null-hypothesis that both phantom and patient VOIs lead to the same distribution of contouring accuracy across all methods used on both image types. First, we take the mean accuracy over the four phantom VOIs as a single score for each contouring method. Next, we measure the accuracy of the same methods used in patient images and take the mean over the three patient VOIs as a single score for each method. Finally, a paired-samples t-test is used for the difference of means between accuracy scores in each image type, with significant difference defined at a confidence level of $p \leq 0.05$. Fig. 5 shows the results separately for accuracy defined by AUC', DSC, and HD'. There is no significant difference between accuracy in phantom and patient images measured by AUC' or DSC. A significant difference is seen for HD', which reflects the sensitivity of HD' to small differences between VOI surfaces. In this case the phantom VOIs are even more difficult to contour accurately than the patient images, which could be explained by the absence of anatomical context in these images, used by operators of manual and semi-automatic contouring methods. A similar experiment found no significant difference between phantom and patient VOIs in terms of intra-operator variability. On the whole we accept the null-hypothesis meaning that the phantom and patient images pose the same challenge to contouring methods in terms of accuracy and variability.

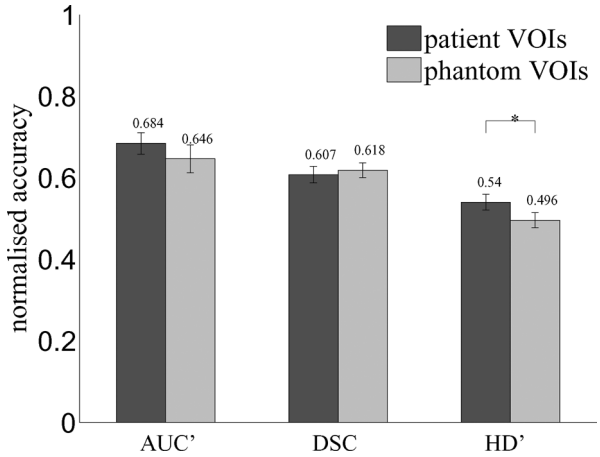

Fig. 5. Contouring accuracy in phantom and patient images, where "┌*┐" indicates significant difference.

TABLE V
Variance of AUC' and Standard Accuracy Metrics Calculated for All Seven VOIs (Second Column), and for the Four and Three VOIs in Phantom and Patient Images, Respectively

metric	all VOIs	phantom	patient
AUC'	0.028	0.035	0.021
DSC	0.011	0.010	0.012
HD'	0.011	0.010	0.011

Fig. 5 also supports the use of the new metric AUC'. Although values are generally higher than DSC and HD, which may be explained by the involvement of multiple ground truth definitions increasing the likelihood that a contour agrees with any one in the set, the variance of accuracy scores is greater for AUC' than the other metrics (Table V), which indicates higher sensitivity to small differences in accuracy between any two methods.

B. Role of CT in PET/CT Contouring

For contouring in patient images only, we test the benefit of exploiting CT information in contouring (phantom VOIs are omitted from this experiments as the CT was used for ground truth definitions and not made available during contouring). This information is in the form of anatomical structure in the case of visual CT-guidance ("low" CT use) and higher-level, image texture information in the case of method HB with "high" CT use. The null-hypothesis is that contouring accuracy is not affected by the level of use of CT information.

We compare each pair of groups i and j that differ by CT use, using a t-test for unequal sample sizes n_i and n_j, where the corresponding samples have mean accuracy μ_i and μ_j and standard deviation σ_i and σ_j. For the ith group containing $\mathrm{n}_{\mathrm{methods}}$ contouring methods, each segmenting $\mathrm{n}_{\mathrm{VOIs}}$ targets, the sample size $\mathrm{n}_i = \mathrm{n}_{\mathrm{methods}} \times \mathrm{n}_{\mathrm{VOIs}}$ and μ_i and σ_j are calculated over all $\mathrm{n}_{\mathrm{methods}} \times \mathrm{n}_{\mathrm{VOIs}}$ accuracy scores. We calculate the significance level from the t-value using the number of degrees of freedom given by the Welch–Satterthwaite formula for unequal sample sizes and sample standard deviations. Significant differences between groups are defined by confidence interval of $p \leq 0.05$. For patient images only, $\mathrm{n}_{\mathrm{VOIs}} = 3$ and for the grouping according to CT use in Table I, $\mathrm{n}_{\mathrm{methods}} = 1, 6$ and 26 for the

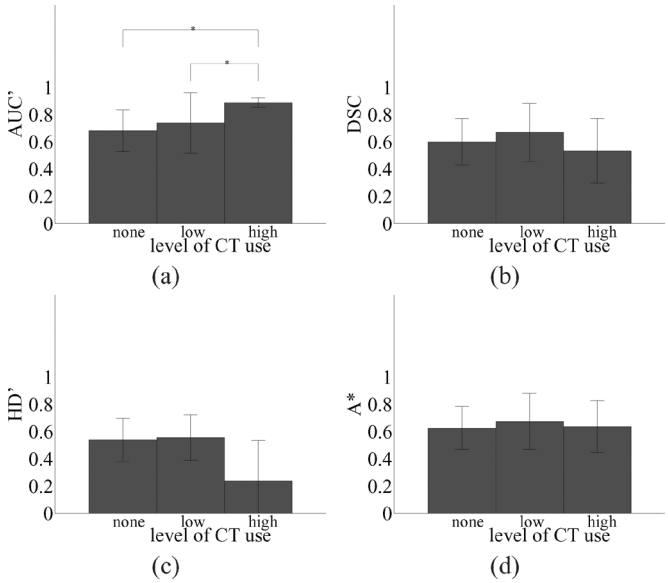

Fig. 6. Effect of CT use on contouring accuracy in patient images, measured by (a) AUC', (b) DSC, (c) HD', and (d) A*, where "$\sqsubset^{*}\sqsupset$" denotes significant difference between two levels of CT use.

groups with levels of CT use "high," "low," and "none," respectively (methods RG^b in the "low" group and MD^b and RG^c in the "none" group were used twice by different operators in the same team). We repeat for four accuracy metrics AUC', DSC, HD' and their weighted sum A*. Fig. 6 shows the results for all groups ordered by level of CT use, in terms of each accuracy metric in turn.

With the exception of AUC' the use of CT as a visual guidance ("low"), outperformed the "high" and "none" groups consistently but without significant difference. The fact that the "high" group (method HB only) significantly outperformed the lower groups in terms of AUC' alone indicates that the method had good spatial agreement with one of the union-of-experts masks for any given VOI, but this union mask did not have absolute volume most closely matching the independent estimates used in calculations of DSC and HD'. We conclude that the use of CT images as visual reference ("low" use) generally improves accuracy, as supported by the consistent improvement in three out of four metrics. This is in agreement with experiments in [30] and [31], which found the benefits of adding CT visually and computationally, in manual and automatic tumor delineation and classification, respectively.

C. Role of User Interaction

This experiment investigates the affect of user-interactivity on contouring performance. The null hypothesis is that contouring accuracy is not affected by the level of interactivity in a contouring method. We compare each pair of groups i and j that differ by level of interactivity, using a t-test for unequal sample sizes as above. For the grouping according to level of interactivity in Table I, groups with interactivity level "max," "high," "mid," "low," and "none" have $\mathrm{n}_{\mathrm{methods}} = 4, 3, 7, 13$ (12 for phantom images by removal of method HB) and 6, respectively (methods MD^b, RGMD^b and RGMD^c in the "max," "high," and "mid" groups, respectively, were used twice by different operators in the same team). We repeat for patient images ($\mathrm{n}_{\mathrm{VOIs}} = 3$), phantom images ($\mathrm{n}_{\mathrm{VOIs}} = 4$), and the combined set ($\mathrm{n}_{\mathrm{VOIs}} = 7$) and, as above, for each of the four accuracy metrics. Fig. 7 shows all results for all groups ordered by level of interactivity.

The trends for each of phantom, patient and all VOIs are consistent over all metrics. The most accurate methods were those in the "high" and "max" groups for phantom and patient images, respectively. For patient images, the "max" group is significantly more accurate than any other and this trend carries over to the pooled accuracies in both image types despite having less patient VOIs ($\mathrm{n} = 3$) than phantom VOIs ($\mathrm{n} = 4$). For phantom VOIs, with the exception of HD', there are no significant differences between "high" and "max" groups and these both significantly outperform the "low" and "none" groups in all metrics. For HD' alone, fully manual delineation is significantly less accurate than semi-automatic methods with "high" levels of interaction. This may reflect the lack of anatomical reference in the phantom images, which is present for patient VOIs and guides manual delineation. As high levels of interaction still appear most accurate, the reduced accuracy of fully manual methods is not considered likely to be caused by a bias of manual delineations toward manual ground truth, given the levels of inter-user variability. Overall, we conclude that manual delineation is more accurate than semi- or fully-automatic methods, and that the accuracy of semi-automatic methods improves with the level of interaction built in.

D. Accuracy of Algorithm Types

This experiment compares the accuracy of different algorithm types, defined in Section II-A. The null hypothesis is that contouring accuracy is the same for manual or any numerical method regardless of the general approach they take. We compare each pair of groups i and j that differ by algorithm type, using a t-test for unequal sample sizes as above. For the grouping according to algorithm type in Table I, $\mathrm{n}_{\mathrm{methods}} = 4, 3, 5, 2, 3, 5, 2, 1, 1$ (0 for phantom images by removal of method HB) and 7 for algorithm-types MD, T1, T2, T3, T4, RG, WS, GR, HB, and PL, respectively (methods MD^b in the MD, and RG^b and RG^c in the RG group were used twice by different operators in the same team). As above, we repeat for patient images ($\mathrm{n}_{\mathrm{VOIs}} = 3$), phantom images ($\mathrm{n}_{\mathrm{VOIs}} = 4$) and the combined set ($\mathrm{n}_{\mathrm{VOIs}} = 7$), and for each of the four accuracy metrics. Fig. 8 shows the results separately for all image sets and accuracy metrics.

Plot (b) reproduces the same anomalous success of the hybrid method (HB) in terms of AUC' alone, as explained above. Manual delineation exhibits higher accuracy than other algorithm types, ranking in the top three for any accuracy metric in phantom images and the top two for any metric in patient images. The pooled results over all images reveal manual delineation as the most accurate in terms of all four metrics. With the exception of T4 in terms of HD' (patient and combined image sets), the improvement of manual delineation over any of the thresholding variants T1–T4 is significant, despite these being the most widely used (semi-)automatic methods. A promising semi-automatic approach is the gradient-based (GR) group

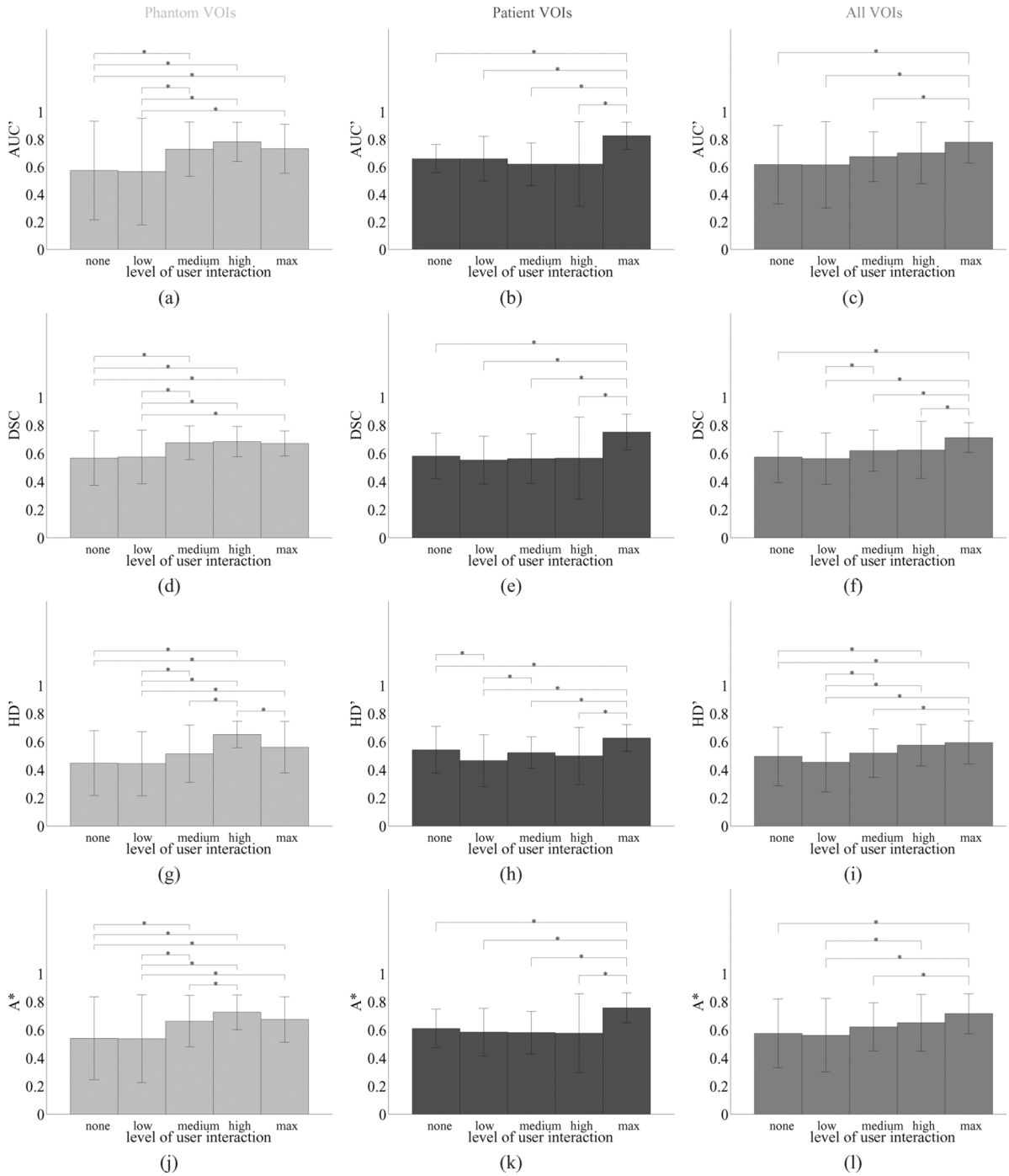

Fig. 7. Effect of user interaction on contouring accuracy measured by *top row:* AUC' for (a) phantom, (b) patient, and (c) both VOI types, *second row:* DSC for (d) phantom, (e) patient, and (f) both image types, *third row:* HD' for (g) phantom, (h) patient, and (i) both image types, and *bottom row:* A* for (j) phantom, (k) patient, and (l) both VOI types. Significant differences between any two levels of user interaction are indicated by "$\sqsubset^{*}\sqcup$."

(one method), which has the second highest accuracy by all metrics for the combined image set and significant difference from manual delineation. Conversely, the watershed group of methods that also rely on image gradients exhibit consistently low accuracy. This emphasized the problem of poorly-defined edges and noise-induced false edges typical of PET gradient filtering, which in turn suggests that edge-preserving noise reduction by the bi-lateral filter plays a large part in the success of method GR.

E. Accuracy of Individual Methods

The final experiments directly compare the accuracy of all methods. Where two algorithms have arguably minor difference, as in the case of PL^{c} and PL^{d} which differ by an extra processing step applied by PL^{d}, these are treated as separate methods because the change in contouring results is notable and can be attributed to the addition of the processing step, which is informative. Repeated segmentations by two different

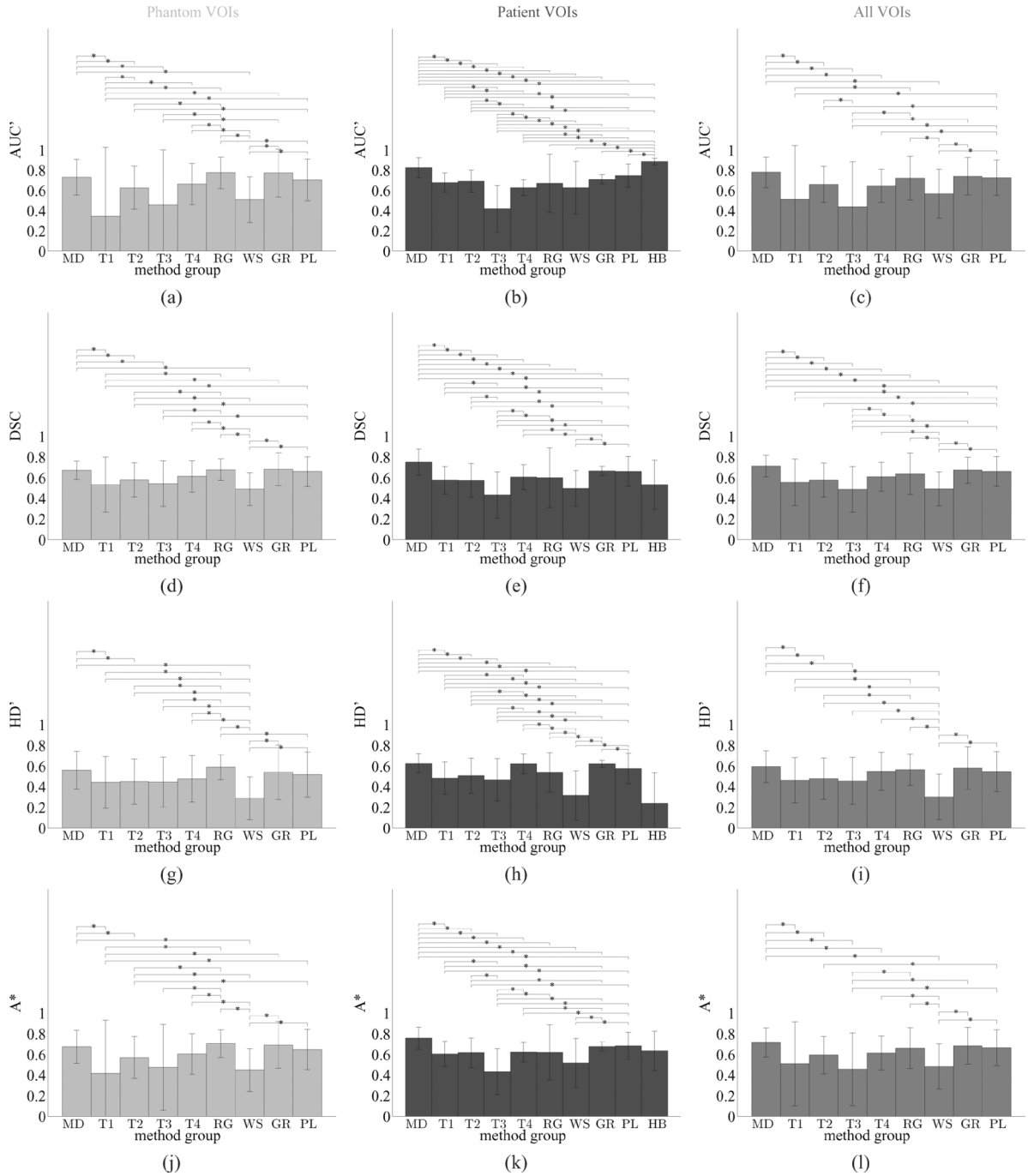

Fig. 8. Contouring accuracy of all algorithm types measured by *top row:* AUC' for (a) phantom, (b) patient, and (c) both VOI types, *second row:* DSC for (d) phantom, (e) patient, and (f) both image types, *third row:* HD' for (g) phantom, (h) patient, and (i) both image types and *bottom row:* A* for (j) phantom, (k) patient, and (l) both VOI types. Significant differences between any two algorithm types are indicated by "$\ulcorner^{*}\urcorner$."

users in the cases of methods $\mathrm{MD}^{b}_{1,2}$, $\mathrm{RG}^{b}_{1,2}$ and $\mathrm{RG}^{c}_{1,2}$ are counted as two individual results so there are a total of $\mathrm{n} = 32$ "methods," or $\mathrm{n} = 33$ for patient VOIs in PET/CT only by inclusion of hybrid method HB. The null hypothesis is that all n cases are equally accurate. We compare each pair of methods i and j that differ by method, using a t-test for equal sample sizes $\mathrm{n}_i = \mathrm{n}_j = \mathrm{n}_{\mathrm{VOIs}}$, where mean accuracy μ_i and μ_j and standard deviation σ_i and σ_j are calculated over all VOIs and there are $2n_{\mathrm{VOIs}} - 2$ degrees-of-freedom. As above, we repeat for all image sets and accuracy metrics. Fig. 9 shows the results separately for phantom, patient and combined image sets in terms of A* only. Full results for all metrics and significant differences between methods are given in the supplementary material.

The generally low values of A* in Fig. 9 and other metrics in the supplementary material highlight the problem facing accurate PET contouring. These results also reiterate the general finding that manual practices can be more accurate than semi- or fully-automatic contouring. For patient images, and the combined set, the most accurate contours are manually delineated by method MD^{c}. Also for these image sets the second and third

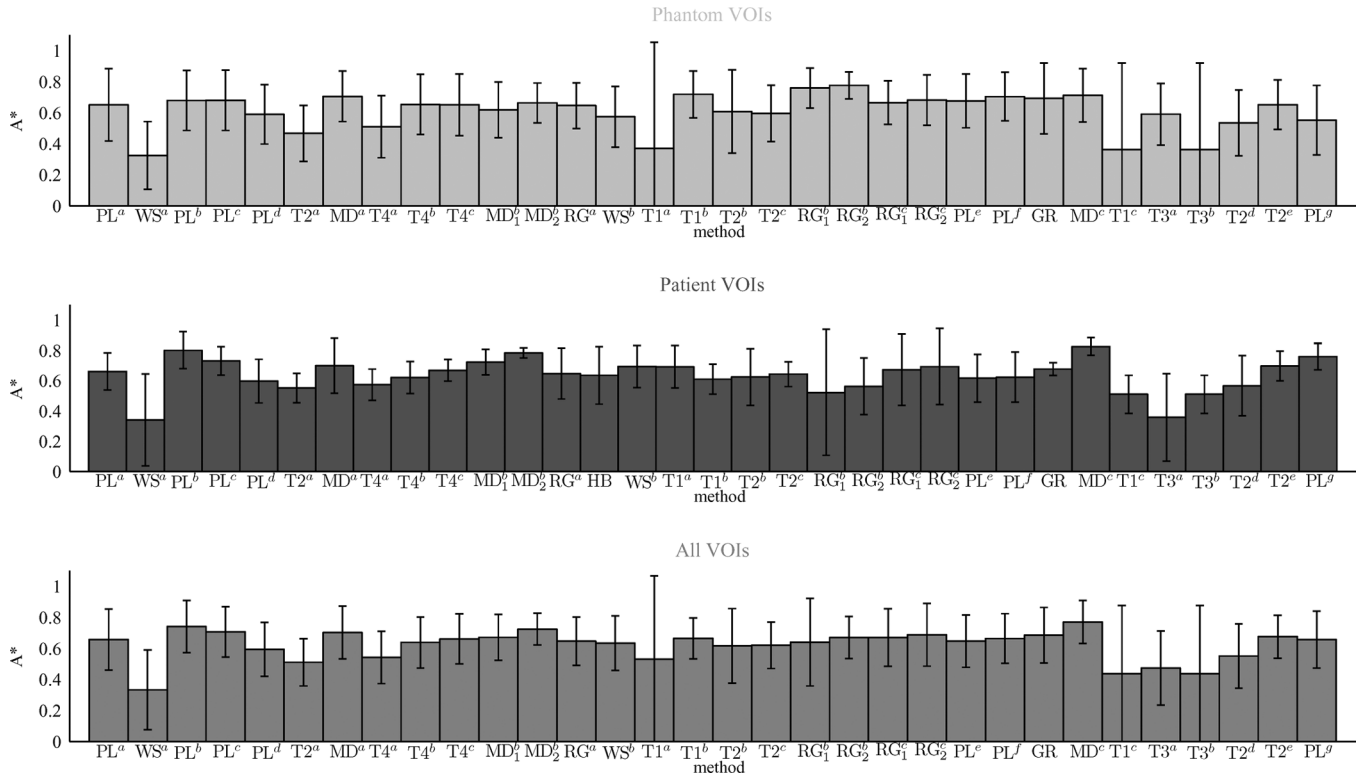

Fig. 9. Mean accuracy measured by A*, of each method used to contour VOIs in phantom (top), patient (middle), and the combined image set (bottom).

most accurate are another manual method (MD_2^b) and the "smart opening" algorithm (PL^b) with mid-level interactivity.

For phantom VOIs only, methods RG^b and $\mathrm{T1}^b$, with high- and low-level interactivity, out-perform manual method MD^c with no significant difference. Method RG^b is based on SRG with post-editing by the adaptive brush and showed low accuracy for patient VOIs with RG_2^b being significantly less accurate than the manual method MD^c (see supplementary material). Method $\mathrm{T1}^b$ is based on thresholding and showed low accuracy for patient VOIs, being significantly less accurate than the manual methods MD^c and MD_2^b (see supplementary material). Their high accuracy in phantom images alone could be explained by methods $\mathrm{T1}^b$ and RG^b being particularly suited to the relative homogeneity of the phantom VOIs.

Methods WS^a, $\mathrm{T1}^c$, and $\mathrm{T3}^b$ have the three lowest accuracies by mean A* across all three image sets. The poor performance of method WS^a could be explained by its origins (color photography and remote-sensing) and user having no roots or specialism in medical imaging. Threshold methods $\mathrm{T1}^c$ and $\mathrm{T3}^b$ give iso-contours at 50% of the local peak intensity without and with adjustment for background intensity, respectively. Their poor performance in all image types highlights the limitations of thresholding.

Table VI presents the composite metrics explained in Section III-C along with intra-operator variability where available (last two columns), measured by the Hausdorff distance in mm between two segmentations of the same VOI, averaged over the three patient or four phantom VOIs. This definition of intra-operator variability gives an anomalously high value if the two segmentations resulting from repeated contouring of the same VOI do not have the same topology, as caused by an internal hole in the first contouring by method RG_1^b. Notably, we find no correlation between intra-operator variability and the level of interactivity of the corresponding methods. The same is true for inter-operator variability (not shown) calculated by the Hausdorff distance between segmentations by different users of the same method (applicable to methods MD^b, RG^b, and RG^c). This finding contradicts the general belief that user input should be minimized to reduce variability. Table VI reaffirms the finding that manual delineation is the most accurate method type, with examples MD^c and $\mathrm{MD}_{1,2}^b$ scoring highly in all metrics. The most consistently accurate nonmanual methods are the semi- and fully-automatic methods PL^b and PL^c. More detailed method-wise comparisons are made in the next section.

V. Discussion

We have evaluated and compared 30 implementations of PET segmentation methods ranging from fully manual to fully automatic and representing the range from well established to never-before tested on PET data. Region growing and watershed algorithms are well established in other areas of medical image processing, while their use for PET target volume delineation is relatively new. Even more novel approaches are found in the "pipeline" group and the two distinct algorithms of gradient-based and hybrid segmentation. The gradient-based method [10] has already had an impact in the radiation oncology community and the HB method [14] is one of few in the literature to make numerical use of the structural information in fused PET/CT. The multispectral approach is in common with classification

TABLE VI
SUMMARIZED ACCURACY AND VARIABILITY OF PHANTOM (PH.) AND PATIENT (PT.) CONTOURING BY ALL METHODS ORDERED AS IN TABLE I AND USING RANKED AND OTHER COMPOSITE ACCURACY METRICS IN SECTION III-C. DATA ARE NOT AVAILABLE (N/A) FOR METHOD HB IN PHANTOM RESULTS AND MOST METHODS IN VARIABILITY RESULTS

method	n(n.s.d)		n($>\mu+\sigma$)		median rank		intra-operator HD (mm)	
	ph.	pt.	ph.	pt.	ph.	pt.	ph.	pt.
PL^a	4	3	0	0	17	19	n/a	n/a
WS^a	0	0	0	0	1.5	1.5	n/a	n/a
PL^b	4	4	0	3	24	31.5	n/a	n/a
PL^c	4	3	1	1	23.5	27	n/a	n/a
PL^d	3	2	0	0	10.5	12.5	n/a	n/a
T2^a	0	1	0	0	4	7	n/a	n/a
MD^a	4	4	2	0	28.5	23	n/a	n/a
T4^a	0	0	0	0	6	9	n/a	n/a
T4^b	4	1	0	0	18.5	15.5	n/a	n/a
T4^c	4	2	0	1	17.5	20.5	n/a	n/a
MD$^b{}_1$	3	3	0	1	13.5	25.5	3.9 ±0.9	4.4 ±1.2
MD$^b{}_2$	3	3	0	3	20.5	31.5	4.1 ±1.7	5.6 ±1.8
RG^a	3	3	1	0	14.5	17	3.7 ±0.6	2.4 ±0.1
HB	n/a	3	n/a	1	n/a	12	n/a	5.6 ±0.6
WS^b	2	2	0	1	8.5	26	3.3 ±3.0	7.4 ±6.7
T1^a	2	3	0	1	3	23	n/a	n/a
T1^b	4	1	1	0	28.5	11	n/a	n/a
T2^b	4	3	0	0	13.5	14	n/a	n/a
T2^c	3	1	0	0	11.5	16.5	n/a	n/a
RG$^b{}_1$	4	3	4	0	31	7	24.0 ±38.9	18.2 ±20.8
RG$^b{}_2$	4	2	4	0	31.5	8	4.5 ±2.4	3.3 ±2.0
RG$^c{}_1$	3	4	0	0	20	20.5	1.5 ±1.7	1.0 ±1.5
RG$^c{}_2$	4	4	0	0	25	22.5	2.6 ±2.0	2.7 ±0.4
PL^e	4	2	0	0	20	12	n/a	n/a
PL^f	4	3	0	0	27.5	14	n/a	n/a
GR	4	0	0	0	25	23	1.2 ±0.0	2.3 ±0.7
MD^c	4	4	1	4	28.5	32.5	2.9 ±0.7	3.8 ±1.2
T1^c	4	0	0	0	3	3.5	n/a	n/a
T3^a	3	1	0	0	10.5	2	n/a	n/a
T3^b	4	0	0	0	4.5	3.5	n/a	n/a
T2^d	0	2	0	0	7	7.5	n/a	n/a
T2^e	4	3	0	1	18.5	26	n/a	n/a
PL^g	3	4	0	3	8.5	29.5	n/a	n/a

experiments in [13] that showed favourable results over PET alone.

A. Manual Delineation

Freehand segmentation produced among the most accurate results, which may be counter-intuitive. One explanation comes from the incorporation of prior knowledge regarding the likely form and extent of pathology. In the case of the patient images alone, bias toward MD may be suspected as the ground-truth set is also built up from manual delineations. However, this does not explain the success of manual methods as they performed better still for phantom VOIs where the ground truth comes from CT thresholds. The use of multiple ground truth estimates by I-ROC may falsely favour manual delineation due to its inherent variability. However, this too does not explain the success of manual methods as they also perform well in terms of DSC and HD' that use a unique, "best-guess" of ground truth (at least one MD is among the five highest DSC and HD for each of the patient phantom VOI sets). These observations challenge the intuition, that manual delineation is less accurate. Although many (semi-)automatic methods outperform freehand delineation in the literature, the inherent bias toward positive results among published work makes this an unfair basis for intuition.

Of the four manual delineations (MD^a, MD^b_1, MD^b_2 and MD^c), method MD^c outperformed the rest in all of n(n.s.d), $\mathrm{n}(>\mu+\sigma)$, median rank and intra-operator variability where known, with significant improvement over $\mathrm{MD}^b_{1,2}$ in terms of AUC' for patient VOIs (although the multiple comparison effect can mean that one or more of these differences are falsely detected as significant). The obvious difference between these four is the user. It is interesting, and indicative of no bias in terms of user group, that the delineator of MD^c was a nuclear medicine physicist while the other users, in common with the experts providing ground truth estimates, were experienced physicians. However, while users of MD^a and $\mathrm{MD}^b_{1,2}$ only viewed the PET images during delineation, the physicist using MD^c also viewed an overlay of the PET gradient magnitude and, in the case of patient images, simultaneous CT. These modes of visual guidance could in part compensate for the relative lack of clinical experience, although no concrete conclusion can be made as clinical sites may disagree on the correct segmentation.

B. Automation versus User Guidance

Two method comparisons provide evidence that too much automation in a semi-automatic algorithm is detrimental to contouring accuracy. First, we compare the accuracy of methods PL^c and PL^d. Method PL^d starts with the same segmentation achieved by PL^c, then performs extra steps in the automatic pipeline intended to improve on the results. However, these extra steps reduce the final accuracy. Second, we compare the accuracy of methods $\mathrm{RG}^b_{1,2}$ and $\mathrm{RG}^c_{1,2}$. These differ in that $\mathrm{RG}^b_{1,2}$ also employs post-editing by the adaptive brush tool. While the adaptive brush may improve accuracy for phantom VOIs, accuracy is reduced for patient VOIs indicated by n(n.s.d) and median rank. This suggests that, where post-editing by unconstrained manual delineation generally improves accuracy in other methods, the automated component of the adaptive brush may influence the editing procedure, and this influence may be detrimental in cases where underlying image information is less reliable.

Conversely, two comparisons give a clear example of the benefits of user-intervention. First, methods PL^e and PL^f are almost the same with the difference that PL^f employs interactive post-editing by user-defined watershed markers and subregional merging. Method PL^f is consistently more accurate than PL^e over all 12 combinations of accuracy metric and image

type. A second example comes from comparing five thresholding schemes used at the same institution (team 13). Methods $\mathrm{T1}^c$, $\mathrm{T3}^a$ and $\mathrm{T3}^b$ use intensity thresholds of 50% maximum and 41% and 50% of maximum-plus-background, while $\mathrm{T2}^d$ and $\mathrm{T2}^e$ use thresholds chosen to match an estimate of the VOI's absolute volume and the user's visual judgement of VOI extent, respectively. Of these five, $\mathrm{T2}^e$ is most highly influenced by the user and ranks consistently higher than the other four in all 12 combinations of accuracy metric and image set, significantly outperforming $\mathrm{T1}^c$ once, $\mathrm{T3}^b$ twice, and $\mathrm{T3}^a$ three times (notwithstanding the possibility of false significance by the multiple comparison effect).

Fully automated contouring has the potential to reduce the user-time involved, whereas contouring speed is not included in the present evaluation strategy. This study focuses on accuracy, given that even fully automatic results can in principle be edited by medical professionals, who ultimately decide how much time is justified for a given treatment plan as well as just where the final contours should lie. The CPU-time of the more computationally expensive algorithms could be quantified as the subject of further work, but its relevance is debatable given that CPUs have different speeds and large data sets can be processed offline, allowing the medical professional to work on other parts of a treatment in parallel.

C. Building Prior Knowledge Into Contouring

As already seen from Fig. 9 method WS^a consistently gave the lowest accuracy. This method was adapted from an algorithm designed for segmenting remote sensing imagery and its user declared no expertise in medical image analysis. Conversely, two methods were adapted for the application of PET oncology, from other areas of medical image segmentation. Method PL^a has origins in white matter lesion segmentation in brain MRI and method PL^b is adapted from segmentation of lung nodules in CT images. These two examples far outperform method WS^a, with method PL^b having the joint second highest median ranking for patient images and no significant difference from the most accurate methods in terms of any metric for any image set.

Some methods were designed for PET oncology, incorporating numerical methods to overcome known challenges. Examples are method GR that overcomes poorly defined gradients around small volumes due in part to partial volume effects, and method PL^g allows for regional heterogeneity that is known to confound PET tumor segmentation. These methods rank reasonably highly, in patient images, ranking similarly to all manual delineations and the semi-automatic "smart opening" algorithm (PL^b), despite neither GR nor PL^g having any user intervention or making any use of simultaneous CT. Method PL^g performs relatively poorly in phantom images, where the problem of tissue heterogeneity is not reproduced.

The benefits of prior knowledge are also revealed by comparing three thresholding schemes $\mathrm{T4}^a$, $\mathrm{T4}^b$, and $\mathrm{T4}^c$ used by the same institution (team 04). Of these, method $\mathrm{T4}^a$ was considerably less accurate in terms of both n(n.s.d) and median rank. Methods $\mathrm{T4}^b$ and $\mathrm{T4}^c$ were calibrated using phantom data to build in prior knowledge of the imaging device. Even though the two devices used to calibrate $\mathrm{T4}^b$ and $\mathrm{T4}^c$ are from different vendors (Siemens and Biograph devices) than the one that acquired the test images (GE Discovery), they are consistently more accurate than method $\mathrm{T4}^a$ implemented at the same site, which does not learn from scanner characteristics but instead has an arbitrary parameter (39%). Methods $\mathrm{T4}^b$ and $\mathrm{T4}^c$ also outperform the majority of the other low-interactivity thresholding schemes, suggesting that the calibration is beneficial and generalizes across imaging devices. This apparent generalization is further evidenced by no significant differences between methods $\mathrm{T4}^b$ and $\mathrm{T4}^c$ in any individual metric for patient or phantom VOIs.

Finally, the *low* accuracy of methods $\mathrm{T4}^a$ and $\mathrm{T4}^a$ may be due to *erroneous* prior knowledge. These two implementations of the same algorithm [6] inherently approximate the volume of interest as a sphere. Both perform poorly, with median ranking from 4–7 over all four metrics in contouring both phantom and patient VOIs. These low accuracies are likely to arise from the spherical assumption rather than the initialization of the method, as the low accuracies are similar despite different methods of initialization described in Section II.

D. Accuracy Evaluation

Accuracy measurement is fundamentally flawed in many medical image segmentation tasks due to the ill-definition of the true surface of the VOI. It is most common to estimate the ground truth by manual delineation performed by a single expert (e.g., [52], [19], [53]). However, even among experts, inter- and intra-operator variability are inevitable and well documented in PET oncology [21], [22]. The new metric AUC' exploits this variability in a probabilistic framework, and we have also defined a single "best guess" ground truth, for use with traditional metrics of DSC and HD, from the union of a subset of expert contours. For patient VOIs, the I-ROC scheme incorporates knowledge and experience of multiple experts as well as structural and clinical information into accuracy measurement and rewards the ability of an algorithm to derive the same information from image data. The I-ROC method considers all ground truth estimates to be equally valid *a priori*, and any one estimate can become the operating point on the I-ROC curve built for a given contour under evaluation. This is in common with the simultaneous truth and performance level estimation (STAPLE) algorithm by Warfield *et al.* [54]. Theirs is also a probabilistic method, which uses maximum likelihood estimation to infer both the accuracy of the segmentation method under investigation and an estimate of the unique ground truth built from the initial set.

Other authors have evaluated segmentation accuracy using phantoms. The most common phantoms used in PET imaging contain simple compartments such as spherical VOIs, attempting to mimick tumors and metastases in head and neck cancer [10], [12], lung nodules [55] and gliomas [20], and cylindrical VOIs, attempting to mimick tumors [7]. The ground truth surface of such VOIs is precisely known due to their geometric form, but many segmentation algorithms are confounded by irregular surfaces and more complex topology such as branching seen in clinical cases and in the new phantom presented here. Another limitation of phantom images including

those used here is the difficulty of mimicking heterogeneous or multi-focal tumors as seen in some clinical data.

Digital images of histological resection can in some cases provide unique ground truth, removing the need to combine multiple estimates. A recent example demonstrates this for PET imaging of prostate cancer [56]. While this approach could provide the standard for accuracy evaluation where available, histology-based accuracy measurement is currently limited as described in [57], with errors introduced by deformation of the organ and co-registration of digital images (co-registration in [56] required first registering manually to an intermediate CT image). Furthermore, tumor excision is only appropriate for some applications. For head-and-neck cancer, the location of the disease often calls for noninvasive, *in vitro* treatment by radiotherapy and in such cases the proposed use of multiple ground truth estimates may provide a new standard.

Neither deterministic metrics with flawed, unique ground truth (DSC and HD) nor probabilistic methods like I-ROC or STAPLE, measure absolute accuracy. However, the relative accuracy of methods or method groups is of interest to our aim of guiding algorithm development. For this purpose, a large and varied cohort of segmentation methods is desirable, and the composite metrics based on method ranking, distributions of accuracy scores $\mathrm{n}(> \mu + \sigma)$ and the frequency of having no significant reduction in accuracy with respect to the most accurate n(n.s.d) become more reliable as the number of contouring tools increases. However, without a simultaneous increase in the number of VOIs, significance tests of the difference in accuracy of any one pair of methods becomes less reliable due to multiple comparison effects.

VI. Conclusion

The multi-center, double-blind comparison of segmentation methods presented here is the largest of its kind completed for VOI contouring in PET oncology. This application has an urgent need for improved software given the demands of modern treatment planning. The number and variety of contouring methods used in this paper alone confirms the need for constraint, if the research is to converge on a small number of contouring solutions for clinical use.

We found that structural images in hybrid PET/CT, now commonly available for treatment planning, should be used for visual reference during semi-automatic contouring while the benefits of high-level CT use by multispectral calculations are revealed only by the new accuracy metric. We also concluded that higher levels of user interaction improves contouring accuracy without increasing intra- or inter-operator variability. Indeed, manual delineation overall outperformed all semi- or fully-automatic methods. However, two methods ($\mathrm{T2}^b$ and PL^f) with a low-level of interactivity and two automatic methods (PL^a and PL^g) are characterized by accuracy scores that are frequently not significantly different from those of the best manual method. Contouring research should pursue a semi-automatic method that achieves the same level of accuracy as expert manual delineation, but must strike a balance between 1) guiding manual practices to reduce levels of variability and 2) not overinfluencing the expert or overriding his or her knowledge. To strike this balance, techniques that show promise are 1) visual guidance by both CT and PET-gradient images, 2) model-based handling of heterogeneity and blurred edges that characterize oncological VOIs in PET, and 3) departure from the reliance on the SUV transformation and iso-contours of this parameter or another scalar multiple of PET intensity, given its dependence on the imaging time window and countless other confounding factors.

These results go a long way towards constraining subsequent development of PET contouring methods, by identifying and comparing the distinct components and individual methods used or proposed in research and the clinic. In addition, we provide detailed results and statistical analyses in supplementary material for use by others in retrospective comparisons according to criteria or method groups not attempted here, as well as access to the phantom images and ground truth sets [48] that can be used to evaluate other contouring methods in the future. While our tests focused on head-and-neck oncology, only the fixed threshold method $\mathrm{T1}^a$ made any assumptions about the tracer or tumor site so results for the remaining methods tested here provide a benchmark for future comparisons. Recently proposed methods in [11], [12], and [58] would be of particular interest to test. However, if the number of tested methods increases without increasing the number of VOIs, the chance of falsely finding significant differences between a pair of methods increases due to the multiple comparison effect so the composite metrics are favoured over pair-wise comparisons for such a benchmark.

Future work using the data from the present study should categorize the 30 methods in terms of user-group and compare segmentation methods in more head and neck VOIs. Future work with a larger set of test data (images and VOIs) is expected to provide more statistically significant findings and should repeat for VOIs outside the realm of FDG in head-and-neck cancer and for images of different signal/background quality. For this purpose the experimental design including phantom, accuracy metrics and the grouping of contemporary segmentation methods, will generalize for other tumor types and PET tracers.

Appendix

In order to derive the new accuracy metric and explain its probabilistic nature, we recall the necessary components of conventional receiver operating characteristic (ROC) analysis, then demonstrate the principles of *inverse-ROC* (I-ROC) for a simple data classification problem and explain the extension to topological ground truth for contour evaluation.

A. Conventional ROC: Multiple Decision Makers

Receiver operating characteristic (ROC) analysis is well established in medical imaging as a means of evaluating region- and voxel-wise data classification [59]. Data comes in the form of $\mathrm{N} = \mathrm{N}_+ + \mathrm{N}_-$ measurements, comprising N_+ "positive" data with truth labels $+1$ and N_- "negative" data with labels -1. A binary classifier divides all N data into positive and negative sets, and has at least one internal parameter that affects this division. ROC analysis is performed by varying an internal parameter in p increments. In threshold classification, the threshold is the internal parameter and data above the threshold are counted as either true positive (TP) or false positive (FP)

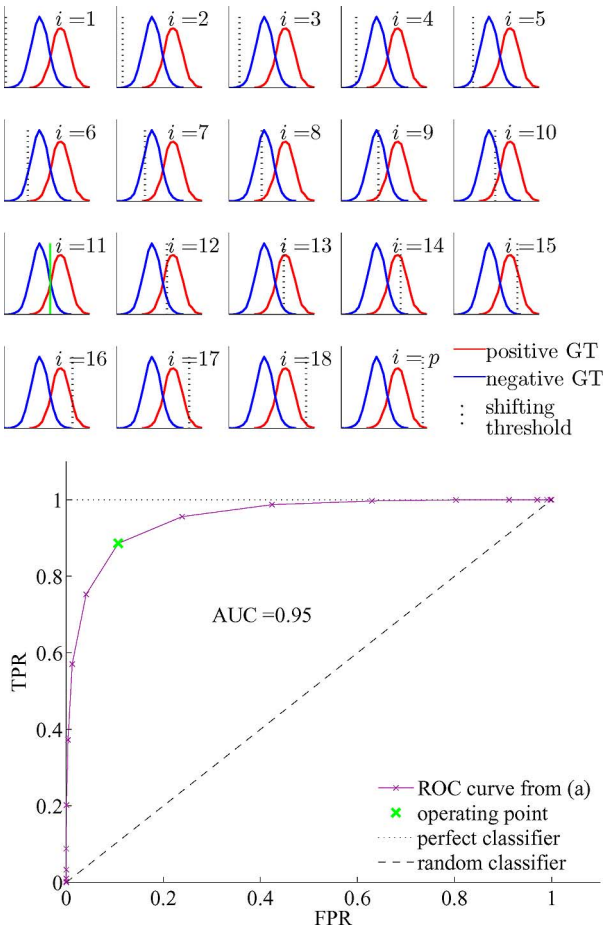

Fig. 10. Conventional ROC analysis of a threshold classifier performed by fixing the ground truth labelling and varying the threshold in $p = 19$ increments (top) to form a ROC curve (bottom). The operating point is marked green.

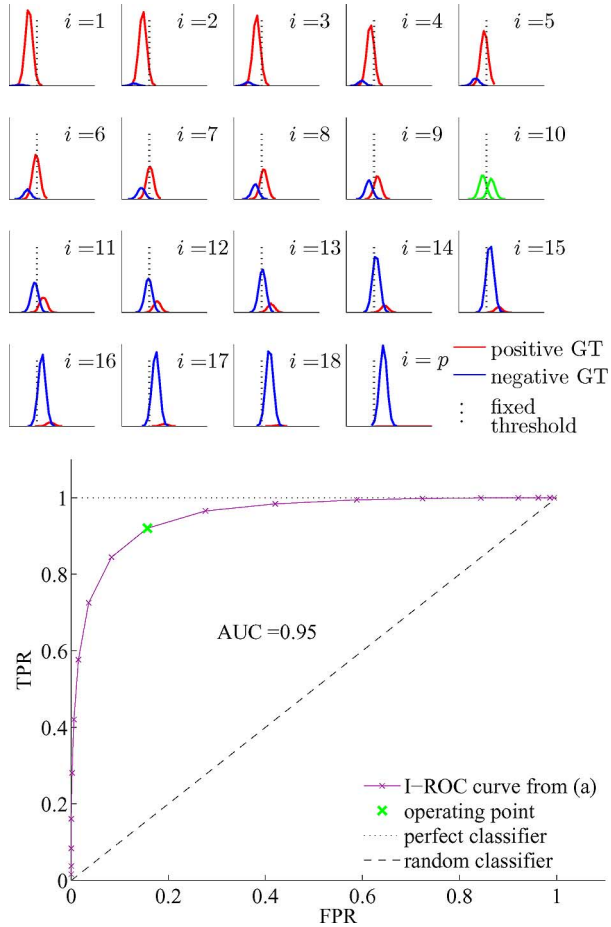

Fig. 11. I-ROC analysis of a threshold classifier performed by varying the ground truth distributions in $p = 19$ increments while the threshold is fixed (top) and plotting the corresponding $\{\mathrm{TPR}_i, \mathrm{FPR}_i\}$ pairs to form a ROC curve (bottom). The operating point is marked green.

according to agreement or otherwise, with the ground truth labels. Similarly, true negative (TN) or false negative (FN) classifications are counted below the threshold. The counts $\mathrm{N_{TP}}$, $\mathrm{N_{FP}}$, $\mathrm{N_{TN}}$, and $\mathrm{N_{FN}}$, of true/false positives and negatives yield the true positive ratio TPR_i and false positive ratio FPR_i for the ith threshold and the pair $\{\mathrm{TPR}_i, \mathrm{FPR}_i\}$ becomes a single point on a ROC curve. The whole curve is generated by varying the internal parameter between natural limits. For the threshold classifier in Fig. 10, the limits are the minimum and maximum value in all N data. The fixed ground truth in Fig. 10 are drawn from Gaussian distributions with $\mu_+ = 3.0$, $\mu_- = -3.0$ and $\sigma_+ = \sigma_- = 2.5$.

The ROC curve occupies the range $\{0 \dots 1\}$ in both TPR and FPR and has two limiting cases. The first limit is the diagonal line (— in Fig. 10) which has an area under the curve (AUC) of 0.5 and indicates failure to classify data better than random assignment of labels ± 1. The second limiting case ($\cdots$ in Fig. 10) has $\mathrm{AUC} = 1$ and indicates perfect classification. As a result, AUC is commonly used as a measure of classifier accuracy. ROC analysis simultaneously yields the operating point of the classifier, defined as the internal parameter setting (e.g., threshold) that minimizes the combined cost of false positives and false negatives.

If positive and negative ground truth are normally distributed, the ROC curve has exponential form and AUC can be calculated by fitting an analytic function and integrating between the limits 0 to 1. In this case, AUC is a monotonic function z^{-1} of the distance between the means μ_+, μ_- of the true distributions, scaled by their standard deviations σ_+, σ_-, where

$$z(\mathrm{AUC}) = \frac{\mu_+ - \mu_-}{\sqrt{\sigma_+^2 + \sigma_-^2}} \tag{5}$$

and AUC is equal to the Gaussian probability that a measurement drawn at random from the positive set will be correctly classified. If the assumption of normally distributed data is relaxed the probabilistic interpretation still holds, where the probability is that sought by a Wilcoxon signed ranks test and AUC is evaluated using the trapezium rule [60].

In summary, AUC is a probabilistic measure regardless of the underlying distributions and ROC analysis can be used as a metric combining sensitivity and specificity.

B. I-ROC: Multiple Ground Truth Representations

The new ROC technique is referred to as *inverse* as, rather than unique ground truth labelling and various arbitrary decision

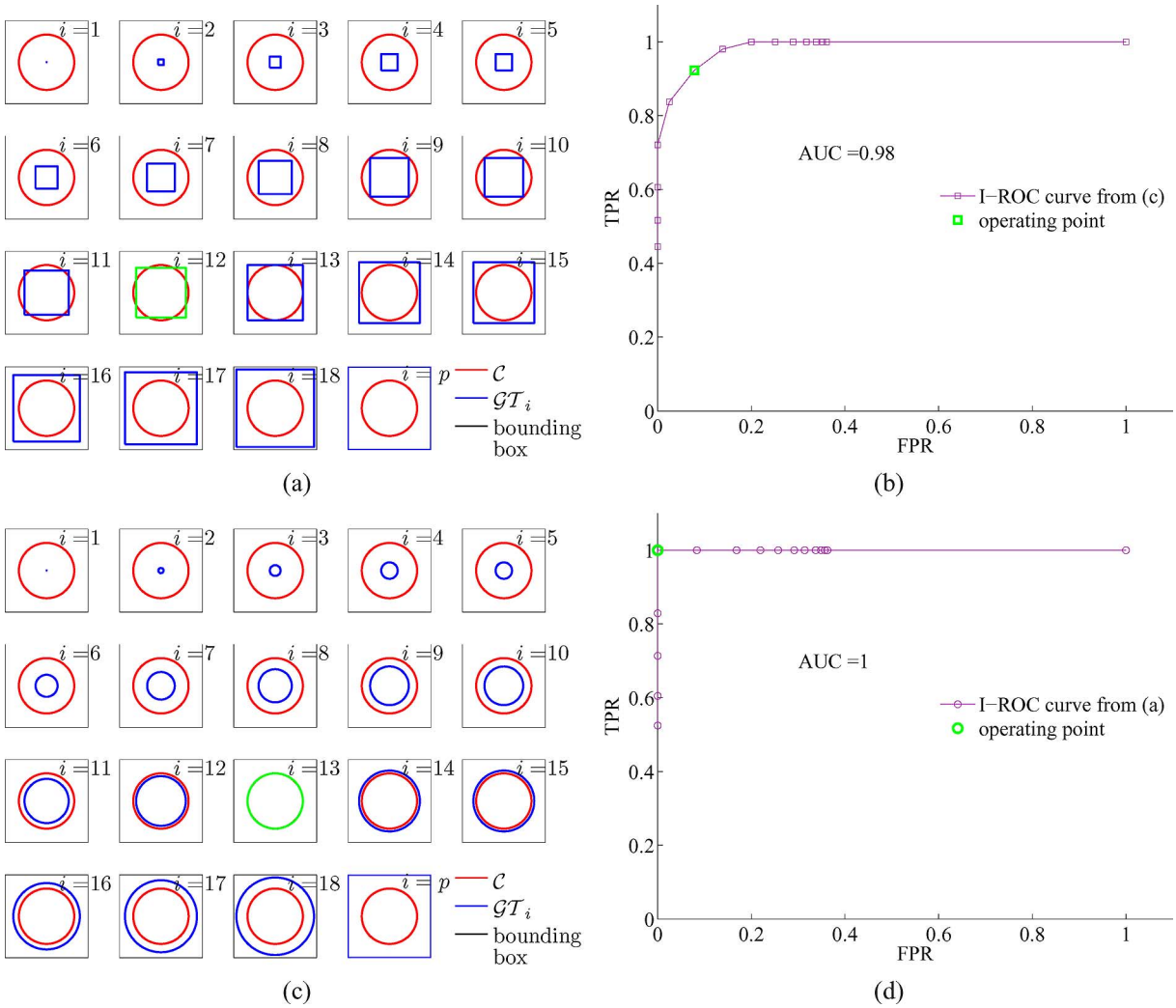

Fig. 12. Inverse-ROC analysis of a fixed contour (red circle) performed by varying ground truth contours as squares (a) or circles (c) of increasing size. ROC curves in (b) and (d) are built from the corresponding true and false counts that lie inside or outside the ith ground truth contour. Operating points are shown in green.

makers, it assumes a single classification and varies the definition of ground truth. Fig. 11 demonstrates this for the example of threshold classification. In common with Fig. 10, data being classified in Fig. 11 are a mixture of Gaussians with means μ_- and μ_+ separated by six units and standard deviations $\sigma_- = \sigma_+ = 2.5$, and the total number of data is fixed at $\mathrm{N}_+ + \mathrm{N}_- = 2 \times 10^4$. To simulate a change in ground truth labelling for the same underlying data, the means of the positive and negative distributions are shifted by δ_i so that $\mu_- = -3.0 + \delta_i$ and $\mu_+ = 3.0 + \delta_i$, where δ_i increases from an arbitrary (negative) minimum to an arbitrary (positive) maximum in $p = 19$ increments, and the proportion ρ of data in the positive set decreases as $\rho = 1 - i/p$. To classify data that has the ith ground truth labelling, we fix the threshold at $\mathrm{T} = 0$ for all $i \in \{1 \ldots p\}$. In line with the requirements of conventional ROC, the multiple ground truth definitions are as follows.

A.2(i) ordered by monotonically (in-) de-creasing N_+.

A.2(ii) obtained by independent means, not the threshold classifier being evaluated.

A.2(iii) incorporate the best knowledge of the unique (unknown) ground truth.

A.2(iv) "pass through" the unique (unknown) ground truth as closely as possible.

Requirement A.2(iii) is realized by fixing the difference of means $\mu_+ - \mu_-$ and having ρ increase with μ_+. Requirement A.2(iv) means that there exist labelings $\{\mathcal{GT}_i\}$ and $\{\mathcal{GT}_{i+1}\}$ with μ_+ and N_+ (similarly μ_- and N_-) either side of the operating point.

The shape of the ROC curve in Fig. 11, the operating point and, within the accuracy of the trapezium integration, the AUC are the same for the I-ROC as for the equivalent analysis in Fig. 10 by virtue of the choice of parameters, which merely serves to illustrate the ability to perform equivalent ROC analyses by shifting decision maker (ROC) or ground truth labelling (I-ROC).

C. I-ROC With Topographic Ground Truth

In the context of VOI contouring, the notion of "positives" refers to voxels inside a contour, which is a spatial distinction and may or may not correspond to voxel values above a threshold. Truth labels in turn are separated by a surface in image space, and stored as a binary mask of $\{\pm 1\}$. We refer

to $\{\mathcal{GT}_i\}$ as a contour or mask interchangeably. The I-ROC method evaluates the accuracy of a fixed result of a contouring algorithm denoted $\mathcal{C}$, using a set of arbitrary ground truth masks $\{\mathcal{GT}_i\}, i \in \{1 \dots p\}$. The term "arbitrary" refers to the fact that no single mask in the set is closest *a priori* to the unknown, unique ground truth and does not mean that their shapes are arbitrary. Following from the requirements for the shifting threshold in A.2, the natural limits $\mathcal{GT}_1$ and $\mathcal{GT}_p$ contain none and all of the image voxels (inside a bounding box), respectively, and the set $\{\mathcal{GT}_i\}, i \in \{1 \dots p\}$.

A.3(i) is ordered monotonically by volume where $\mathcal{GT}_i$ completely encloses $\mathcal{GT}_{i-1}$.

A.3(ii) is obtained independently of the contouring algorithm under evaluation.

A.3(iii) incorporates the best available knowledge of ground truth.

A.3(iv) "passes through" the unknown, unique ground truth surface as closely as possible.

Requirement A.3(i) can always be met by defining each $\mathcal{GT}_i$ as the union of contours from an original set. Requirements A.3(ii) and (iii) can also always be met, whereby suggested sources of independent information are complementary imaging or clinical information unseen to the tool under evaluation. Requirement A.3(iv) means that topology and general shape are conserved within the set as in the analogy of inflating a novelty balloon, and can also always be met by the procedure used to obtain all $\mathcal{GT}_i$, such as the suggested use of union masks.

If the general shape common to all $\{\mathcal{GT}_i\}$ is representative of the unknown ground truth then AUC is higher when the contour under evaluation shares this shape. Fig. 12 demonstrates this for the case where the ground truth set has a different (a) and the same (c) shape as a circular contour $\mathcal{C}$ under evaluation. Using a square ground truth set (a) gives $\mathrm{AUC} < 1$, equivalent to the case of overlapping histograms in Fig. 11, although the similar form of the curve and value of $\mathrm{AUC} = 0.98$ are only due to the parameters and shapes used for illustration. A circular set, chosen for its agreement with $\mathcal{C}$ to illustrate the possibility of achieving $\mathrm{AUC} = 1$, indicates perfect contouring accuracy. More generally, AUC approaches 1 as the contour $\mathcal{C}$ approaches any contour in the set $\{\mathcal{GT}_i\}$ and this indicates perfect agreement with the general form of the unknown, unique ground truth all in the set $\{\mathcal{GT}_i\}$ share this form. It follows that AUC is equal to the probability that a voxel drawn at random from inside the optimal $\mathcal{GT}_i$, which is not known *a priori*, lies inside the contour $\mathcal{C}$ being evaluated.

Formally, the I-ROC method will generalize for any shape of ground truth set or contour under evaluation if

$$\mathrm{N}(\in \mathcal{GT}_j) = \sum_{i=1}^{j} \mathrm{N}(\mathcal{GT}_i \vee \mathcal{GT}_j) \quad \text{and} \qquad \text{(a)}$$

$$\mathrm{N}(\in \mathcal{GT}_j) + \mathrm{N}(\notin \mathcal{GT}_j) = \text{constant} \quad \forall j \quad \text{(b)} \qquad (6)$$

where $\mathrm{N}(\in \mathcal{GT}_j)$ and $\mathrm{N}(\notin \mathcal{GT}_j)$ denote the number of voxels inside and outside the jth ground truth definition. Equation (6a) holds if requirement A.3(i) is met and (6b) is satisfied by the fixed bounding box enclosing the set $\{\mathcal{GT}_i\}$.

ACKNOWLEDGMENT

For retrospective patient data and manual ground truth delineation, the authors wish to thank S. Suilamo, K. Lehtiö, M. Mokka, and H. Minn at the Department of Oncology and Radiotherapy, Turku University Hospital, Finland. This study was funded by the Finnish Cancer Organisations.

REFERENCES

[1] R. Murakami, H. Uozumi, T. Hirai, R. Nishimura, S. Shiraishi, K. Oto, D. Murakami, S. Tomiguchi, N. Oya, S. Katsuragawa, and Y. Yamashita, "Impact of FDG-PET/CT fusion imaging on nodal staging and radiationtherapy planning for head-and-neck squamous cell carcinoma," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 66, p. 185, 2007.

[2] C. Nutting, "Intensity-modulated radiotherapy (IMRT): The most important advance in radiotherapy since the linear accelerator?," *Br. J. Radiol.*, vol. 76, p. 673, 2003.

[3] J. W. Keyes, "SUV: Standardised uptake or silly useless value?," *J. Nucl. Med.*, vol. 36, pp. 1836–1839, 1995.

[4] E. P. Visser, O. C. Boerman, and W. J. G. Oyen, "SUV: From silly useless value to smart uptake value," *J. Nucl. Med.*, vol. 51, pp. 173–175, 2010.

[5] Y. Nakamoto, K. R. Zasadny, H. Minn, and R. L. Wahl, "Reproducibility of common semi-quantitative parameters for evaluating lung cancer glucose metabolism with positron emission tomography using 2-deoxy- 2-[18F]Fluoro-D-Glucose.," *Mol. Imag. Biol.*, vol. 4, pp. 171–178, 2002.

[6] J. A. van Dalen, "A novel iterative method for lesion delineation and volumetric quantification with FDG PET," *Nucl. Med. Commun.*, vol. 28, pp. 485–493, 2007.

[7] J. F. Daisne, M. S. , A. Bol, T. D. M. Lonneux, and V. Grégoire, "Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: Influence of reconstruction algorithms," *Radiother. Oncol.*, vol. 69, pp. 247–250, 2003.

[8] A. Schaefer, S. Kremp, D. Hellwig, C. Rube, C.-M. Kirsch, and 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. Ima.*, vol. 35, pp. 1989–1999, 2008.

[9] H. Zaidi and I. El Naqa, "PET-guided delineation of radiation therapy treatment volumes: A survey of image segmentation techniques," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 37, pp. 2165–2187, 2010.

[10] X. Geets, J. A. Lee, A. Bol, M. Lonneux, and V. Grégoire, "A gradientbased method for segmenting FDG-PET images: Methodology and validation," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 34, pp. 1427–1438, 2007.

[11] I. El-Naqa, D. Yang, A. Apte, D. Khullar, S. Mutic, J. Zheng, J. D. Bradley, P. Grigsby, and J. O. Deasy, "Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning," *Med. Phys.*, vol. 34, pp. 4738–4749, 2007.

[12] H. Li, W. L. Thorstad, K. J. Biehl, R. Laforest, Y. Su, K. I. Shoghi, E. D. Donnelly, D. A. Low, and W. Lu, "A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours," *Med. Phys.*, vol. 35, pp. 3711–3721, 2008.

[13] H. Yu, C. Caldwell, K. Mah, and D. Mozeg, "Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning," *IEEE Trans. Med. Imag.*, vol. 28, no. 3, pp. 374–383, Mar. 2009.

[14] D. Han, J. Bayouth, Q. Song, A. Taurani, M. Sonka, J. Buatti, and X. Wu, "Globally optimal tumor segmentation in PET-CT images: A graph-based co-segmentation method," in *Proceedings, Information Processing in Medical Imaging (IPMI)*, 2011, vol. 6801, Lecture Notes in Computer Science, pp. 245–256.

[15] S. Belhassen and H. Zaidi, "A novel fuzzy c-means algorithm for unsupervised heterogeneous tumor quantification in PET," *Med. Phys.*, vol. 37, pp. 1309–1324, 2010.

[16] M. Hatt, C. C. le Rest, P. Descourt, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, O. Pradier, and D. Visvikis, "Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 77, pp. 301–308, 2010.

[17] J. A. Lee, "Segmentation of positron emission tomography images: Some recommendations for target delineation in radiation oncology," *Radiother. Oncol.*, vol. 96, pp. 302–307, 2010.

[18] U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welsch, D. Hellwig, C. Rübe, and C. Kirsch, "Comparison of different methods for delineation of 18F-FDG PETpositive tissue for target volume definition in radiotherapy of patients with nonsmall cell lung cancer," *J. Nucl. Med.*, vol. 46, pp. 1342–1348, 2005.

[19] C. Greco, S. A. Nehmeh, H. Schöder, M. Gönen, B. Raphael, H. E. Stambuk, J. L. Humm, S. M. Larson, and 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.*, vol. 31, pp. 439–445, 2008.

[20] H. Vees, S. Senthamizhchelvan, R. Miralbell, D. C. Weber, O. Ratib, and 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. Imag.*, vol. 36, pp. 182–193, 2009.

[21] A. C. Riegel, A. M. Berson, S. Destian, T. Ng, L. B. Tena, R. J. Mitnick, and P. S. Wong, "Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 65, pp. 726–732, 2006.

[22] S. L. Breen, J. Publicover, S. De Silva, G. Pond, K. Brock, B. O. Sullivan, B. Cummings, L. Dawson, A. Keller, J. Kim, J. Ringash, E. Yu, A. Hendler, and J. Waldron, "Intraobserver and interobserver variability in GTV delineation on FDG-PET-CT images of head and neck cancers," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 68, pp. 763–770, 2007.

[23] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, V. Jewells, and S. W. , Eds., "Workshop proceedings, 3D segmentation in the clinic: A grand challenge II—MS lesion segmentation," in *2008, Medical Image Computing and Computer-Assisted Invervention (MICCAI)* [Online]. Available: http://grand-challenge2008.bigr.nl/proceedings/mslesions/articles.html

[24] X. Deng and G. D. , Eds., "Workshop proceedings, 3D segmentation in the clinic: A grand challenge II—Liver tumour segmentation," in *2008, Medical Image Computing and Computer-Assisted Invervention (MICCAI)* [Online]. Available: http://grand-challenge2008.bigr.nl/proceedings/liver/articles.html

[25] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[26] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Haussdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[27] N. Hata, G. Fichtinger, S. Oguro, H. Elhawary, and T. van Walsum, "Prostate segmentation challenge 2009," in *2009, Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop: 3D Segmentation Challenge for Clinical Applications* [Online]. Available: http://wiki.na-mic.org/Wiki/index.php/2009

[28] V. Pekar, J. Kim, S. Allaire, A. Qazi, and D. A. Jaffray, "Head and neck auto-segmentation challenge 2010," in *2010, Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop: Medical Image Analysis for the Clinic: A Grand Challenge* [Online]. Available: http://www.grand-challenge2010.ca/

[29] T. Shepherd, M. Teräs, and H. Sipilä, "New physical tumour phantom and data analysis technique exploiting hybrid imaging and partial volume effects for segmentation evaluation in radiation oncology," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 37, p. S221, 2011.

[30] H. Schöder, H. W. D. Yeung, M. Gonen, D. Kraus, and S. M. Larson, "Head and neck cancer: Clinical usefulness and accuracy of PET/CT image fusion," *Radiology*, vol. 231, pp. 65–72, 2004.

[31] H. Yu, C. Caldwell, K. Mah, I. Poon, J. Balogh, R. MacKenzie, N. Khaouam, and R. Tirona, "Automated radiation targeting in headandneck cancer using region-based texture analysis of PET and CT images," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 75, pp. 618–25, 2010.

[32] J. Daisne, T. Duprez, B. Weynand, M. Lonneux, M. Hamoir, H. Reychler, and V. Grégoire, "Tumor volume in pharyngolaryngeal squamous cell carcinoma: Comparison at CT, MRI, and FDG PET and validation with surgical specimen. radiology," *Radiology*, vol. 233, pp. 93–100, 2004.

[33] K. R. Zasadny and R. L. Wahl, "Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-D-glucose: Variations with body weight and a method of correction," *Radiology*, vol. 189, pp. 847–850, 1993.

[34] D. Hellwig, T. P. Graeter, D. Ukena, A. Groeschel, G. W. Sybrecht, H.-J. Schaefers, and C.-M. Kirsch, "18F FDG PET for mediastinal staging of lung cancer: Which SUV threshold makes sense?," *J. Nucl. Med.*, vol. 48, pp. 1761–1766, 2007.

[35] Y. Erdi, O. Mawlawi, S. M. Larson, M. Imbriaco, H. Yeung, R. Finn, and J. L. Humm, "Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding," *Cancer*, vol. 80, pp. 2505–2509, 1997.

[36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man Cybern.*, vol. 9, pp. 62–66, 1979.

[37] ABX Advanced Biochemical Compounds, ROVER: ROI Visualisation, Evaluation and Image Registration 2010 [Online]. Available: http://www.abx.de/rover/index.php/id-3d-regions-of-interest.html

[38] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.

[39] S. Beucher and F. Meyer, *The Morphological Approach to Segmentation: The Watershed Transformation*. New York: Marcel Dekker, 1993, ch. 12, pp. 433–481.

[40] S. Lefèvre, "Knowledge from markers in watershed segmentation," in *Proceedings, IAPR International Conference on Computer Analysis of Images and Patterns*, 2007, vol. 4673, Lecture Notes in Computer Sciences, pp. 579–586.

[41] López-Mir, V. Naranjo, J. Angulo, E. Villanueva, M. A. Niz, and S. López-Celada, "Aorta segmentation using the watershed algorithm for an augmented reality system in laprascopic surgery," in *Proc. IEEE Int. Conf. Image Process.*, 2011, pp. 2649–2652.

[42] Y. Y. Yang, C. M. Li, C. Y. Kao, and S. Osher, "Split Bregman method for minimization of region-scalable fitting energy for image segmentation," in *Proceedings, International Symposium on Visual Computing*, 2010, Lecture Notes in Computer Sciences, pp. 117–128.

[43] J.-M. Kuhnigk, V. Dicken, L. Bornemann, A. Bakai, D. Wormanns, S. Krass, and H.-O. Peitgen, "Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans," *IEEE Trans. Med. Imag.*, vol. 25, no. 4, pp. 417–434, Apr. 2006.

[44] C. M. Li, C. Y. Kao, C. John, and Z. H. Ding, "Minimization of region-scalable fitting energy for image segmentation," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1940–1949, Oct. 2008.

[45] T. F. Chan, S. Esedoglu, and M. Nikolova, "Algorithms for finding global minimizers of denoising and segmentation models," *SIAM J. Appl. Math.*, vol. 66, pp. 1632–1648, 2006.

[46] J. A. Lee, X. Geets, V. Gregoire, and A. Bol, "Edge-preserving filtering of images with low photon counts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1014–1027, Jun. 2008.

[47] S. Li, "Markov random field models in computer vision," in *Computer Vision ECCV'94*, ser. Lecture Notes in Computer Science, J.-O. Eklundh, Ed. Berlin, Germany: Springer, 1994, vol. 801, pp. 361–370.

[48] T. Shepherd, "Contour evaluation," Tumour phantom for contour evaluation 2012 [Online]. Available: http://www.turkupetcentre.fi/files/tumourphantom/

[49] T. T. Tanimoto, IBM Internal Report 1957.

[50] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, pp. 327–352, 1977.

[51] R. Real and J. M. Vargas, "The probabilistic basis of Jaccard's index of similarity," *Systematic Biol.*, vol. 45, pp. 385–390, 1996.

[52] M. Chupin, A. R. Mukuna-Bantumbakulu, D. Hasboun, E. Bardinet, S. Baillet, S. Kinkingnéhun, L. Lemieux, B. Dubois, and L. Garnerob, "Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with alzheimers disease," *NeuroImage*, vol. 34, pp. 996–1019, 2007.

[53] D. W. Shattuck, G. Prasada, M. Mirzaa, K. L. Narra, and A. W. Togaa, "Online resource for validation of brain segmentation methods," *NeuroImage*, vol. 45, pp. 431–439, 2009.

[54] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[55] Q. Black, I. Grills, L. Kestin, C. Wong, J. Wong, A. Martinez, and D. Yan, "Defining a radiotherapy target with positron emission tomography," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 60, pp. 1272–1282, 2004.

[56] J. H. Chang, D. L. Joon, S. T. Lee, S. J. Gong, A. M. Scott, I. D. Davis, D. Clouston, D. Bolton, C. S. Hamilton, and V. Khoo, "Histopathological correlation of ^{11}C-choline PET scans for target volume definition in radical prostate radiotherapy," *Radiother. Oncol.*, vol. 99, pp. 187–192, 2011.

[57] R. J. Hicks and M. P. Mac Manus, "18F-FDG PET in candidates for radiation therapy: Is it important and how do we validate its impact?," *J. Nucl. Med.*, vol. 44, pp. 30–32, 2003.

[58] M. Hatt, C. C. le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," *IEEE Trans. Med. Imag.*, vol. 28, no. 6, pp. 881–893, Jun. 2009.

[59] J. A. Swets, "ROC analysis applied to the analysis of medical imaging techniques," *Invest. Radiol.*, vol. 14, pp. 109–121, 1979.

[60] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, pp. 1145–1159, 1997.