



HAL
open science

Comparative study with new accuracy metrics for target volume contouring in pet image guided radiation therapy

T. Shepherd, M. Teras, R.R. Beichel, R. Boellard, M. Bruynooghe, V. Dicken, M.J. Gooding, P.J. Julyan, J.A. Lee, Sébastien Lefèvre, et al.

► To cite this version:

T. Shepherd, M. Teras, R.R. Beichel, R. Boellard, M. Bruynooghe, et al.. Comparative study with new accuracy metrics for target volume contouring in pet image guided radiation therapy. IEEE Transactions on Medical Imaging, 2012, 31 (11), pp.2006-2024. 10.1109/TMI.2012.2202322 . hal-00763488

HAL Id: hal-00763488

<https://hal.science/hal-00763488>

Submitted on 13 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 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 emphasise spatial accuracy and present a new metric that addresses the lack of unique ground truth. 30 methods are used at 13 different institutions to contour functional VOIs 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 visualisation 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.

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 CT for treatment planning [1], unprecedented accuracy of intensity modulated

radiation therapy (IMRT) [2] and on-going debates [3], [4] over the ability of the standardised 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 (eg 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], [19], [20] do not reflect the wide range of proposed and potential algorithms and fall short of measuring spatial accuracy. [18] compare 3 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 3 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 MRI. Vees *et al.* [20] compare one manual, 4 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 analysed together objectively. In 2008, two examples of

T. Shepherd, and H. Minn are with the Turku PET Centre, Turku University Hospital, and Department of Oncology and Radiotherapy, University of Turku, Finland. M. Teräs is with the Turku PET Centre, Turku University Hospital, Finland. R. Beichel is with the Department of Electrical & Computer Engineering and Internal Medicine, University of Iowa, USA. R. Boellaard is with the Department of Nuclear Medicine & PET Research, VU University Medical Centre, Amsterdam, The Netherlands. M. Bruynooghe is with SenoCAD Research GmbH, Germany. V. Dicken is with Fraunhofer MEVIS - Institute for Medical Image Computing, Bremen, Germany. M. J. Gooding is with Mirada Medical, Oxford, UK. P. J. Julyan is with North Western Medical Physics, Christie Hospital NHS Foundation Trust, Manchester, UK. J. A. Lee is with the Belgian FNRS and the center for Molecular Imaging, Radiotherapy, and Oncology (MIRO), Université Catholique de Louvain, Brussels, Belgium. S. Lefèvre is with the VALORIA Research Laboratory, University of South Brittany, France. M. Mix is with the Department of Radiation Oncology, University Freiburg Medical Centre, Germany. V. Naranjo is with the Labhuman Inter-University Research Institute for Bioengineering and Human Centered Technology, Valencia, Spain. X. Wu is with the Department of Electrical & Computer Engineering, University of Iowa, USA. H. Zaidi is with the Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Switzerland. Z. Zeng is with the Department Computer Science, University of Aberystwyth, UK.

71 pathological segmentation involved multiple sclerosis lesions
 72 in MRI [23] and liver tumours in CT [24]. These tests
 73 involved 9 and 10 segmentation algorithms respectively, and
 74 evaluated their accuracy using a combination of the Dice
 75 similarity coefficient [25] and Hausdorff distance [26] with
 76 respect to a single manual delineation of each VOI. In 2009
 77 and 2010, the challenges were to segment the prostate in
 78 MRI [27] and parotid in CT [28]. These compared 2 and 10
 79 segmentation methods respectively, each using a combination
 80 of various overlap and distance measures to evaluate accuracy
 81 with respect to a single manual ground truth per VOI. The
 82 MICCAI challenges have had a major impact on segmentation
 83 research in their respective application areas, but this type
 84 of large-scale, double-blind study has not previously been
 85 applied to PET target volume delineation for therapeutic
 86 radiation oncology, and the examples above are limited by
 87 their dependence upon a single manual delineation to define
 88 ground truth of each VOI.

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

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

127 II. CONTOURING METHODS

128 Thirteen contouring 'teams' took part in the experiment. We
 129 identify 30 distinct 'methods', where each is a unique com-
 130 bination of team and algorithm. Table I presents the methods
 131 along with labels (first column) used to identify them hereafter.
 Some teams used more than one contouring algorithm and

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			PL								▲		■

132 some well-established algorithms such as thresholding were
 133 used by more than one team, with different definitions of the
 134 quantity and its threshold. Methods are grouped according to
 135 algorithm type and distinguished by their level of dependence
 136 upon the user (section II-B) and CT data (section II-C) in
 137 the case of patient images. Contouring by methods MD^b,
 138 RG^b and RG^c was repeated by two users in the respective
 139 teams, denoted by subscripts 1 and 2, and the corresponding
 140 segmentations are treated separately in our experiments.

141 Some of the methods are well known for PET segmentation
 142 while others are recently proposed. Of the recently proposed
 143 methods, some were developed specifically for PET segmen-
 144 tation (e.g. GR, T2^d and PL^g) while some were adapted and
 145 optimised for PET tumour contouring for the purpose of this
 146 study. The study actively sought new methods, developed
 147 or newly adapted for PET tumours, as their strengths and
 148 weaknesses will inform current research that aims to refine or
 149 replace state of the art tools, whether those tools are included
 150

151 here or not. Many of the algorithms considered operate on
 152 standardised uptake values (SUVs), whereby PET voxel in-
 153 tensity I is rescaled as $SUV = I \times (\beta/a_{in})$ to standardise with
 154 respect to initial activity a_{in} of the tracer in $Bq\ ml^{-1}$ and
 155 patient mass β in grams [33]. The SUV transformation only
 156 affects segmentation by fixed thresholding while methods that
 157 normalise with respect to a reference value in the image or
 158 apply thresholds at a percentage of the maximum value are
 159 invariant to the SUV transformation.

160 A. Method types and descriptions

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

175 **Thresholding methods (T1 - T4)** are divided into 4 types
 176 according to whether the threshold is applied to signal (T1 &
 177 T2) or a combination of signal and background intensity (T3
 178 & T4) and whether the threshold value is chosen *a priori*,
 179 based on recommendations in the literature or the team's
 180 own experience (T1 & T3) or chosen for each image, either
 181 automatically according to spatial criteria or visually by the
 182 user's judgement (T2 & T4). Without loss of generalisation
 183 the threshold value may be absolute or percentage (e.g. of
 184 peak) intensity or SUV. $T1^a$ & $T1^b$ employ the widely used
 185 cut-off values of 2.5 SUV and 40% of the maximum in the
 186 VOI, as used for lung tumour segmentation in [34] and [35]
 187 respectively. Method $T1^a$ is the only method of all in table
 188 I that is directly affected by the conversion from raw PET
 189 intensity to SUVs. The maximum SUV used by method $T1^b$
 190 was taken from inside the VOI defined by $T1^a$. To calculate
 191 SUV for the phantom image, where patient weight β is
 192 unavailable, all voxel values were re-scaled with respect to
 193 a value of unity at one end of the phantom where intensity is
 194 near uniform, causing method $T1^a$ to fail for phantom scan 2
 195 as the maximum was below 2.5 for both VOIs. $T1^c$ applies
 196 a threshold at 50% of the maximum SUV. Method $T2^a$ is
 197 the thresholding scheme of [6], which automatically finds the
 198 optimum relative threshold level (RTL) based an estimate of
 199 the true absolute volume of the VOI in the image. The RTL
 200 is relative to background intensity, where background voxels
 201 are first labelled automatically by clustering. An initial VOI
 202 is estimated by a threshold of 40% RTL, and its maximum
 203 diameter is determined. The RTL is then adjusted iteratively
 204 until the absolute volume of the VOI matches that of a sphere
 205 of the same diameter, convolved with the point-spread function
 206 (PSF) of the imaging device, estimated automatically from the

207 image. Methods $T2^b$ & $T2^c$ automatically define thresholds
 208 according to different criteria. They both use the results of
 209 method $T1^a$ as an initial VOI, and define local background
 210 voxels by dilation. Method $T2^b$ uses two successive dilations
 211 and labels the voxels in the second dilation as background.
 212 The auto-threshold is then defined as 3 standard deviations
 213 above the mean intensity in this background sample. Method
 214 $T2^c$ uses a single dilation to define the background and finds
 215 the threshold that minimises the within-class variance between
 216 VOI and background using the optimization technique in [36].
 217 Finally, method $T2^c$ applies a closing operation to eliminate
 218 any holes within the VOI, which may also have the effect
 219 of smoothing the boundary. Method $T2^d$ finds the RTL using
 220 the method of [6] in common with method $T2^a$ but with
 221 different parameters and initialisation. Method $T2^d$ assumes a
 222 PSF of 7 mm full width at half maximum (FWHM) rather than
 223 estimating this value from the image. The RTL was initialized
 224 with background defined by a manual bounding box rather
 225 than clustering and foreground defined by method $T3^a$ with a
 226 50% threshold rather than 40% RTL. Adaptive thresholding
 227 method $T2^e$ starts with a manually defined bounding box
 228 then defines the VOI by the iso-contour at a percentage of
 229 the maximum value within the bounding box. Methods $T3^a$
 230 & $T3^b$ are similar to $T1^c$, but incorporate local background
 231 intensity calculated by a method equivalent to that Daisne
 232 *et al.* [37]. A threshold value is then 41% and 50% of the
 233 maximum plus background value, respectively. Method $T4^a$
 234 is an automatic SUV-thresholding method implemented in
 235 the 'Rover' software [38]. After defining a search area that
 236 encloses the VOI, the user provides an initial threshold which
 237 is adjusted in two steps of an iterative process. The first step
 238 estimates background intensity I_b from the average intensity
 239 over those voxels that are below the threshold *and* within
 240 a minimum distance of the VOI (above the threshold). The
 241 second step re-defines the VOI by a new threshold at 39% of
 242 the difference $I_{max} - I_b$, where I_{max} is the maximum intensity
 243 in the VOI. Methods $T4^b$ & $T4^c$ use the source-to-background
 244 algorithm in [8]. The user first defines a background region
 245 specific to the given image, then uses parameters a and b to
 246 define the threshold $t = a\mu_{VOI} + b\mu_{BG}$, where μ_{VOI} and μ_{BG}
 247 are the mean SUV in the VOI and background respectively. The
 248 parameters are found in a calibration procedure by scanning
 249 spherical phantom VOIs of known volume. As this calibration
 250 was not performed for the particular scanner used in the
 251 present experiments (GE Discovery), methods $T4^b$ and $T4^c$
 252 use parameters previously obtained for Gemini and Biograph
 253 PET systems respectively.

254 **Region growing methods (RG)** use variants of the classical
 255 algorithm in [39], which begins at a 'seed' voxel in the
 256 VOI and agglomerates connected voxels until no more satisfy
 257 criteria based on intensity. In RG^a , the user defines a bounding
 258 sphere centred on the VOI, defining both the seed at the centre
 259 of the sphere and a hard constraint at the sphere surface to
 260 avoid leakage into other structures. The acceptance criterion
 261 is an interactively adjustable threshold and the final VOI is
 262 manually modified in individual slices if needed. Methods
 263 RG^b & RG^c use the region growing tool in Mirada XD
 264 (Mirada Medical, Oxford, UK) 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 3D painting tool adapts in shape to the underlying image. Also in method RG^b only, CT images were fused with PET for visualisation and the information used to modify the regions to exclude airways and unaffected bone.

Watershed methods (WS) use variants of the classical algorithm in [40]. 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 WS^a , adapted from the algorithm in [41] for segmenting natural colour 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 WS^b , similar way to that in [42], 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 [43], [44], [15]. Method 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 [45] and a global standard convex scheme [46] in energy minimization based on the 'Split Bregman' technique in [43]. Methods $PL^b - PL^d$ are variants of the 'Smart Opening' algorithm, adapted for PET from the tool in [44] 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 PL^e & 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 3 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 sub-region merging. This intervention is only necessary in at most two slices per VOI. Method 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)** method 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 isocontours 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 [47]. 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 tumour 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 [48]. 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 initialisation, which locates the VOI and/or representative voxels, as well as run-time parameter adjustment and post-editing of the contours. 'Mid'-level interactions involve user-initialisation and either run-time parameter adjustment or other run-time information such as wrongly included/excluded voxels. 'Low'-level interaction refers to initialisation 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

378 'none', where 'high' refers to *numerical* use of CT together
 379 with PET in calculations. The 'low' group makes *visual* use of
 380 CT images to guide manual delineation, post-editing or other
 381 interactions in semi-automatic methods. The 'none' group
 382 refers to cases where CT is not used, or is viewed incidentally
 383 but has no influence on contouring as the algorithm is fully
 384 automatic. None of the methods operated on CT images alone.

III. EXPERIMENTAL METHODS

A. Images

387 We use two images of a new tumour phantom [29], man-
 388 ufactured for this study and two clinical PET images of
 389 different head-and-neck cancer patients. The test images are
 390 available on-line [49], along with ground truth sets described
 391 in section III-C. All imaging used the metabolic tracer ^{18}F -
 392 Fluorodeoxyglucose (FDG) and a hybrid PET/CT scanner
 393 (GE Discovery), but CT images from phantom scans were
 394 omitted from the test set. Table II gives more details of each
 395 image type. The tumour phantom contains glass compartments
 396 of irregular shapes shown in figure 1 (top row), mimicking
 real radiotherapy target volumes. The tumour compartment

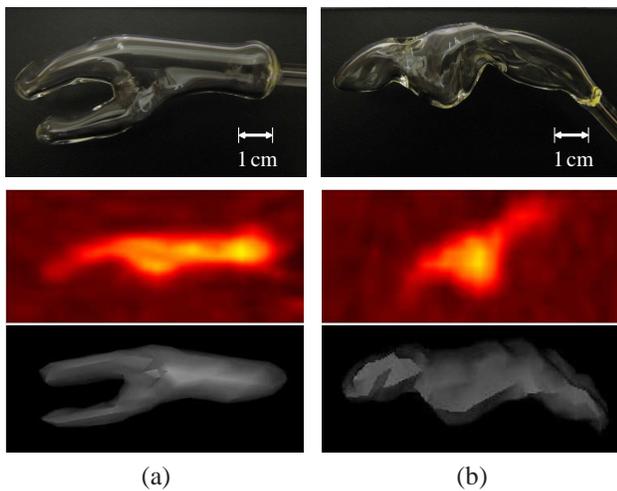


Fig. 1: (a) tumour and (b) nodal chain VOIs of the phantom. *Top*: Digital photographs of glass compartments. *Middle*: PET images from scan 1 (sagittal view). *Bottom*: 3D surface view from an arbitrary threshold of simultaneous CT, lying within the glass wall.

397 (a) has branches to recreate the more complex topology of
 398 some tumours. This and the nodal chain compartment (b) are
 399 based on cancer of the oral cavity and lymph node metastasis
 400 respectively, manually segmented from PET images of two
 401 head and neck cancer patients and formed by glass blowing.
 402 The phantom compartments and surrounding container were
 403 filled with low concentrations of FDG and scanned by a hybrid
 404 device (1, middle and bottom rows). Four phantom VOIs result
 405 from scans 1 and 2, with increasing signal to background ratio
 406 achieved by increasing FDG concentration in the VOIs. Details
 407 of the 4 phantom VOIs are given in the first 4 rows of table
 408 III. Figure 2 shows the phantom VOIs from scan 1, confirming
 409 qualitatively the spatial and radiometric agreement between
 410 phantom and patient VOIs.
 411

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 node	phantom scan 1	8.7 (VOI) 4.9 (BG)	6.71 7.45	thresholds of simultaneous CT image
tumour node	phantom scan 2	10.7 (VOI) 2.7 (BG)	6.71 7.45	
tumour node	patient 1	2.4×10^5	35.00 2.54	
tumour node	patient 2	3.6×10^9	2.35	

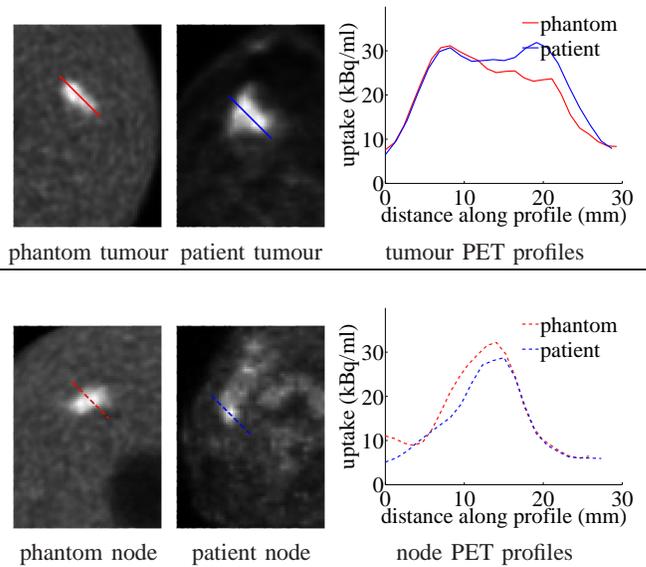


Fig. 2: Axial PET images of phantom and real tumour (top) and lymph node (bottom) VOIs with profile lines traversing each VOI. Plots on the right show the image intensity profiles sampled from each image pair.

412 For patient images, head and neck cancer was chosen as it
 413 poses particular challenges to PET-based treatment planning
 414 due to the many nearby organs at risk (placing extra demand on
 415 GTV contouring accuracy), the heterogeneity of tumour tissue
 416 and the common occurrence of lymph node metastasis. A large
 417 tumour of the oral cavity and a small tumour of the larynx were
 418 selected from two different patients, along with a metastatic
 419 lymph node in the first patient (figure 3). These target volumes
 420 were chosen as they were histologically proven and have a
 421 range of sizes, anatomical locations/surroundings and target
 422 types (tumour and metastasis). Details of the 3 patient VOIs
 423 are given in the last 3 rows of table III.

B. Contouring

424 With the exception of the hybrid method (HB) that does not
 425 apply to the PET-only phantom data, all methods contoured
 426 all 7 VOIs. In the case of patient VOIs, participants had
 427 the option of using CT as well as PET, and were instructed
 428 to contour the gross tumour volume (GTV) and metastatic
 429 tissue of tumours and lymph node respectively. All contouring
 430

TABLE II: Details of phantom and patient PET/CT images.

Image type	PET (^{18}F FDG)						CT					
	frame (min)	length	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		

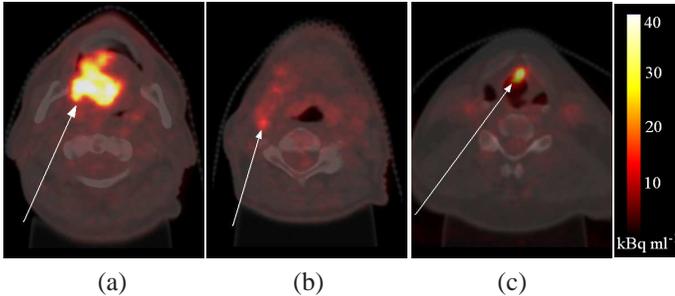


Fig. 3: Axial neck slices of ^{18}F -FDG PET images overlain on simultaneous CT. (a) & (b) Oral cavity tumour & lymph node metastasis in patient 1 (c) Laryngeal tumour in patient 2.

431 methods were used at the sites of the respective teams using
 432 their own software and workstations. Screen-shots of each
 433 VOI were provided in axial, sagittal and coronal views, with
 434 approximate centres indicated by cross-hairs and their voxel
 435 coordinates provided to remove any ambiguity regarding the
 436 ordering of axes and direction of increasing indices. No other
 437 form of ground truth was provided. Teams were free to refine
 438 their algorithms and practice segmentation before accepting
 439 final contours. This practicing stage was done without any
 440 knowledge of ground truth and is considered normal practice.
 441 Any contouring results with sub-voxel precision were down-
 442 sampled to the resolution of the PET image grid and any
 443 results in mm were converted to voxel indices. Finally, all
 444 contouring results were duplicated to represent VOIs first by
 445 the voxels on their surface, and second by masks of the solid
 446 VOI including the surface voxels. These two representations
 447 were used in surface-based and volume-based contour evalua-
 448 tion respectively.

449 C. Contouring evaluation

450 Accuracy measurement generally compares the contour being
 451 evaluated, which we denote \mathcal{C} , with some notion of ground
 452 truth, denoted \mathcal{GT} . We use a new probabilistic metric [29]
 453 denoted AUC' , as well as a variant of the Hausdorff distance
 454 [26] denoted HD' and the standard metric of Dice similarity
 455 coefficient [25] (DSC). AUC' and HD' are standardised to the
 456 range $0 \dots 1$ so that they can be easily combined or compared
 457 with DSC and other accuracy metrics occupying this range
 458 [50], [51], [52]. Treated separately, AUC' , HD' and DSC
 459 allow performance evaluation with and without the assumption
 460 of unique ground truth, and in terms of both volumetric
 461 agreement (AUC' and DSC) and surface-displacement (HD')
 462 with respect to ground truth.

463 AUC' is a probabilistic metric based on receiver operating
 464 characteristic (ROC) analysis, in a scheme we call *inverse-*

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

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

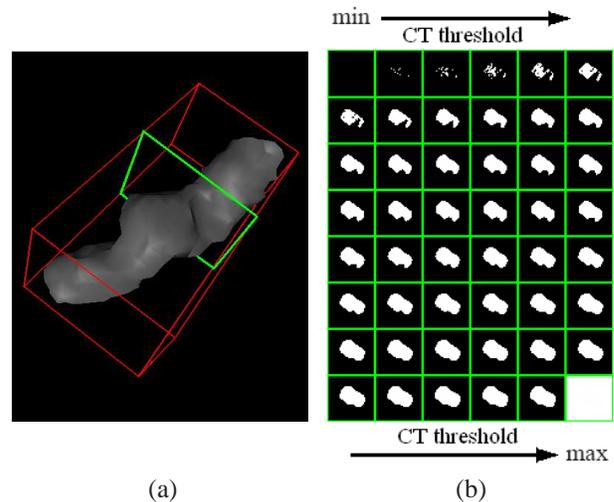


Fig. 4: (a) 3D visualisation of phantom VOI from CT thresholded at a density near the internal glass surface. (b) Arbitrary ground truth masks of the axial cross section in (a), from 50 thresholds of HU.

483 CT slices in the following steps:
 484

- 485 (i) reconstruct/down-sample the CT image to the same
 486 pixel grid as the PET image
- 487 (ii) define a bounding box in the CT image that completely
 488 encloses the glass VOI as well as \mathcal{C}
- 489 (iii) threshold the CT image at a value HU_i
- 490 (iv) treat all pixels below this value as being 'liquid' and
 491 all above it as 'glass'

(v) label all 'liquid' pixels that are *inside* the VOI as positive, but ignore pixels outside the VOI.

(vi) repeat for p thresholds $HU_i, i \in \{1 \dots p\}$ between natural limits HU_{\min} and 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 visualised 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 tumour-type and access to extra information such as tumour 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, 3 delineated each VOI on 2 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:

- (i) 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
- (ii) order the segmentations $\{\mathcal{GT}_i\}$ by absolute volume in cm^3
- (iii) use the smallest segmentation as \mathcal{GT}_2
- (iv) 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
- (v) repeat until the largest VOI in the set has been used in the union of all $N_{\text{experts}} \times N_{\text{occasions}}$ VOIs
- (vi) 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 $\{0 \dots 1\}$ by

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

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.

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 normalise HD with respect to a length scale r and subtract the result from 1

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

where $r = \sqrt[3]{\frac{3}{4\pi} \text{vol}(\mathcal{GT}^*)}$ is the radius of a sphere having the same volume as \mathcal{GT}^* denoted $\text{vol}(\mathcal{GT}^*)$. 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

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

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

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

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^* penalises 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) favours a segmentation tool that is

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

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 n(> μ + σ) favours 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.

568 **Intra-operator variability** was measured by the raw Haus-
 569 dorff distance in mm between the first and second seg-
 570 mentation result from repeated contouring (no ground truth
 571 necessary). However, this was only done for some contouring
 572 methods. For fully automatic methods, variability is zero by
 573 design and was not explicitly measured. Of the remaining
 574 semi-automatic and manual methods, 11 were used twice by
 575 the same operator: MD^b₁, MD^b₂, RG^a, HB, WS^b, RG^b₁,
 576 RG^b₂, RG^c₁, RG^c₂, GR and MD^c and for these we measure
 577 the intra-operator variability which allows extra, direct com-
 578 parisons in section IV-E.

579 IV. EXPERIMENTS

580 This section motivates the use of the new phantom and
 581 accuracy metric (IV-A), then investigates contouring accuracy
 582 by comparing the pooled accuracy of methods grouped ac-
 583 cording to their use of CT data (section IV-B), level of user
 584 interactivity (section IV-C) and algorithm type (section IV-D).
 585 Section IV-E evaluates methods individually, using condensed
 586 accuracy metrics in table IV. With the inclusion of repeated
 587 contouring by methods MD^b, RG^b and RG^c by a second
 588 operator, there are a total of $n = 33$ segmentations of each
 589 VOI, with the exception of phantom VOIs where $n = 32$ by the
 590 exclusion of method HB. Also, method T1^a failed to recover
 591 phantom VOIs in scan 1 as no voxels were above the pre-
 592 defined threshold. In this case a value of zero accuracy is
 593 recorded for two out of 4 phantom VOIs.

594 A. Phantom and AUC'

595 This experiment investigates the ability of the phantom
 596 to pose a realistic challenge to PET contouring, by testing
 597 the null-hypothesis that both phantom and patient VOIs lead
 598 to the same distribution of contouring accuracy across all
 599 methods used on both image types. First, we take the mean
 600 accuracy over the 4 phantom VOIs as a single score for each
 601 contouring method. Next, we measure the accuracy of the same
 602 methods used in patient images and take the mean over the
 603 3 patient VOIs as a single score for each method. Finally,
 604 a paired-samples t-test is used for the difference of means
 605 between accuracy scores in each image type, with significant
 606 difference defined at a confidence level of $p \leq 0.05$. Figure 5
 607 shows the results separately for accuracy defined by AUC',
 608 DSC and HD'. There is no significant difference between
 609 accuracy in phantom and patient images measured by AUC'
 610 or DSC. A significant difference is seen for HD', which
 611 reflects the sensitivity of HD' to small differences between
 612 VOI surfaces. In this case the phantom VOIs are even more
 613 difficult to contour accurately than the patient images, which
 614 could be explained by the absence of anatomical context in
 615 these images, used by operators of manual and semi-automatic
 616 contouring methods. A similar experiment found no significant
 617 difference between phantom and patient VOIs in terms of
 618 intra-operator variability. On the whole we accept the null-
 619 hypothesis meaning that the phantom and patient images pose
 620 the same challenge to contouring methods in terms of accuracy
 621 and variability.

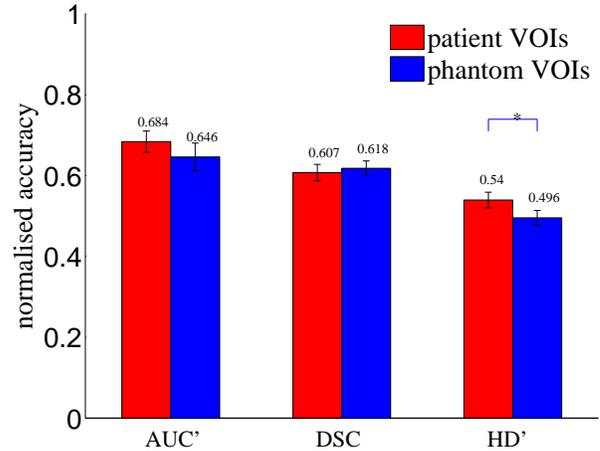


Fig. 5: Contouring accuracy in phantom and patient images, where '⌈*⌋' indicates significant difference.

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

TABLE V: Variance of AUC' and standard accuracy metrics calculated for all 7 VOIs (second column), and for the 4 and 3 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

629 B. Role of CT in PET/CT contouring

630 For contouring in patient images only, we test the benefit of
 631 exploiting CT information in contouring (phantom VOIs are
 632 omitted from this experiments as the CT was used for ground
 633 truth definitions and not made available during contouring).
 634 This information is in the form of anatomical structure in the
 635 case of visual CT-guidance ('low' CT use) and higher-level,
 636 image texture information in the case of method HB with
 637 'high' CT use. The null-hypothesis is that contouring accuracy
 638 is not affected by the level of use of CT information.
 639

640 We compare each pair of groups i and j that differ by CT
 641 use, using a t-test for unequal sample sizes n_i and n_j , where
 642 the corresponding samples have mean accuracy μ_i and μ_j and
 643 standard deviation σ_i and σ_j . For the i^{th} group containing
 644 n_{methods} contouring methods, each segmenting n_{VOIs} targets, the
 645 sample size $n_i = n_{\text{methods}} \times n_{\text{VOIs}}$ and μ_i and σ_j are calculated
 646 over all $n_{\text{methods}} \times n_{\text{VOIs}}$ accuracy scores. We calculate the
 647 significance level from the t-value using the number of degrees
 648 of freedom given by the Welch-Satterthwaite formula for un-
 649 equal sample sizes and sample standard deviations. Significant

650 differences between groups are defined by confidence interval
 651 of $p \leq 0.05$. For patient images only, $n_{\text{VOIs}} = 3$ and for the
 652 grouping according to CT use in table I, $n_{\text{methods}} = 1, 6$ and 26
 653 for the groups with levels of CT use 'high', 'low' and 'none'
 654 respectively (methods RG^b in the 'low' group and MD^b & RG^c
 655 in the 'none' group were used twice by different operators in
 656 the same team). We repeat for 4 accuracy metrics AUC' , DSC ,
 657 HD' and their weighted sum A^* . Figure 6 shows the results
 658 for all groups ordered by level of CT use, in terms of each
 accuracy metric in turn.

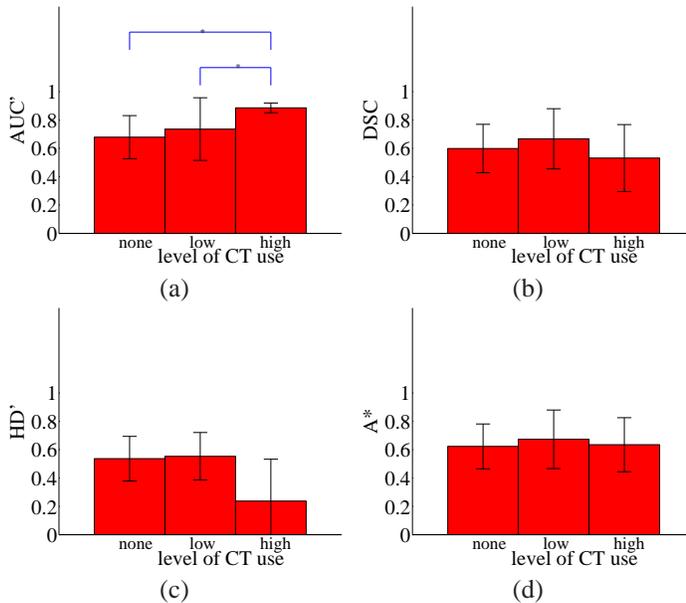


Fig. 6: Effect of CT use on contouring accuracy in patient images, measured by (a) AUC' , (b) DSC , (c) HD' and (d) A^* , where 'Γ*Γ' denotes significant difference between two levels of CT use.

659 With the exception of AUC' the use of CT as a visual
 660 guidance ('low'), out-performed the 'high' and 'none' groups
 661 consistently but without significant difference. The fact that the
 662 'high' group (method HB only) significantly out-performed
 663 the lower groups in terms of AUC' alone indicates that the
 664 method had good spatial agreement with one of the union-of-
 665 experts masks for any given VOI, but this union mask did not
 666 have absolute volume most closely matching the independent
 667 estimates used in calculations of DSC and HD' . We conclude
 668 that the use of CT images as visual reference ('low' use)
 669 generally improves accuracy, as supported by the consistent
 670 improvement in 3 out of 4 metrics. This is in agreement
 671 with experiments in [30] and [31], which found the benefits
 672 of adding CT visually and computationally, in manual and
 673 automatic tumour delineation and classification respectively.
 674

675 C. Role of user interaction

676 This experiment investigates the affect of user-interactivity
 677 on contouring performance. The null hypothesis is that con-
 678 touring accuracy is not affected by the level of interactivity
 679 in a contouring method. We compare each pair of groups i

and j that differ by level of interactivity, using a t-test for
 680 unequal sample sizes as above. For the grouping according
 681 to level of interactivity in table I, groups with interactivity
 682 level 'max', 'high', 'mid', 'low' and 'none' have $n_{\text{methods}} =$
 683 4, 3, 7, 13 (12 for phantom images by removal of method
 684 HB) and 6 respectively (methods MD^b , RGMD^b and RGMD^c
 685 in the 'max', 'high' and 'mid' groups respectively were used
 686 twice by different operators in the same team). We repeat for
 687 patient images ($n_{\text{VOIs}} = 3$), phantom images ($n_{\text{VOIs}} = 4$) and
 688 the combined set ($n_{\text{VOIs}} = 7$) and, as above, for each of the
 689 4 accuracy metrics. Figure 7 shows all results for all groups
 690 ordered by level of interactivity.
 691

The trends for each of phantom, patient and all VOIs
 692 are consistent over all metrics. The most accurate methods
 693 were those in the 'high' and 'max' groups for phantom and
 694 patient images respectively. For patient images, the 'max'
 695 group is significantly more accurate than any other and this
 696 trend carries over to the pooled accuracies in both image
 697 types despite having less patient VOIs ($n = 3$) than phantom
 698 VOIs ($n = 4$). For phantom VOIs, with the exception of HD' ,
 699 there are no significant differences between 'high' and 'max'
 700 groups and these both significantly out-perform the 'low' and
 701 'none' groups in all metrics. For HD' alone, fully manual
 702 delineation is significantly less accurate than semi-automatic
 703 methods with 'high' levels of interaction. This may reflect the
 704 lack of anatomical reference in the phantom images, which
 705 is present for patient VOIs and guides manual delineation.
 706 As high levels of interaction still appear most accurate, the
 707 reduced accuracy of fully manual methods is not considered
 708 likely to be caused by a bias of manual delineations toward
 709 manual ground truth, given the levels of inter-user variability.
 710 Overall, we conclude that manual delineation is more accurate
 711 than semi- or fully-automatic methods, and that the accuracy of
 712 semi-automatic methods improves with the level of interaction
 713 built in.
 714

715 D. Accuracy of algorithm types

This experiment compares the accuracy of different al-
 716 gorithm types, defined in section II-A. The null hypothesis
 717 is that contouring accuracy is the same for manual or any
 718 numerical method regardless of the general approach they
 719 take. We compare each pair of groups i and j that differ
 720 by algorithm type, using a t-test for unequal sample sizes as
 721 above. For the grouping according to algorithm type in table I,
 722 $n_{\text{methods}} = 4, 3, 5, 2, 3, 5, 2, 1, 1$ (0 for phantom images by removal
 723 of method HB) and 7 for algorithm-types MD, T1, T2, T3,
 724 T4, RG, WS, GR, HB and PL respectively (methods MD^b in
 725 the MD, and RG^b & RG^c in the RG group were used twice by
 726 different operators in the same team). As above, we repeat for
 727 patient images ($n_{\text{VOIs}} = 3$), phantom images ($n_{\text{VOIs}} = 4$) and
 728 the combined set ($n_{\text{VOIs}} = 7$), and for each of the 4 accuracy
 729 metrics. Figure 8 shows the results separately for all image
 730 sets and accuracy metrics.
 731

Plot (b) reproduces the same anomalous success of the
 732 hybrid method (HB) in terms of AUC' alone, as explained
 733 above. Manual delineation exhibits higher accuracy than other
 734 algorithm types, ranking in the top 3 for any accuracy metric
 735

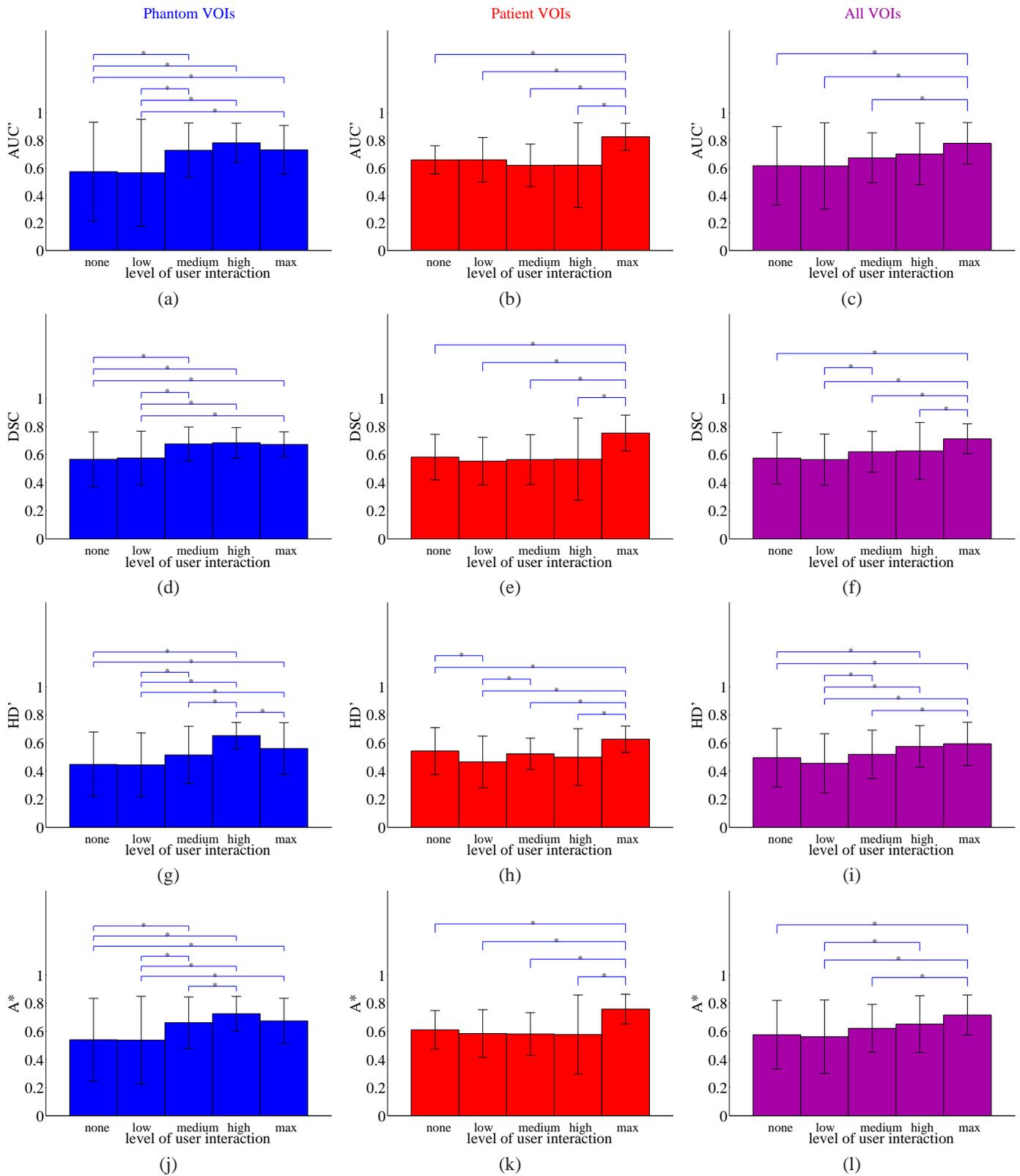


Fig. 7: Effect of user interaction on contouring accuracy measured by *top row*: AUC' for (a) phantom, (b) patient and (c) both VOI types, *second row*: DSC for (d) phantom (e) patient and (f) both image types, *third row*: HD' for (g) phantom, (h) patient and (i) both image types, and *bottom row*: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

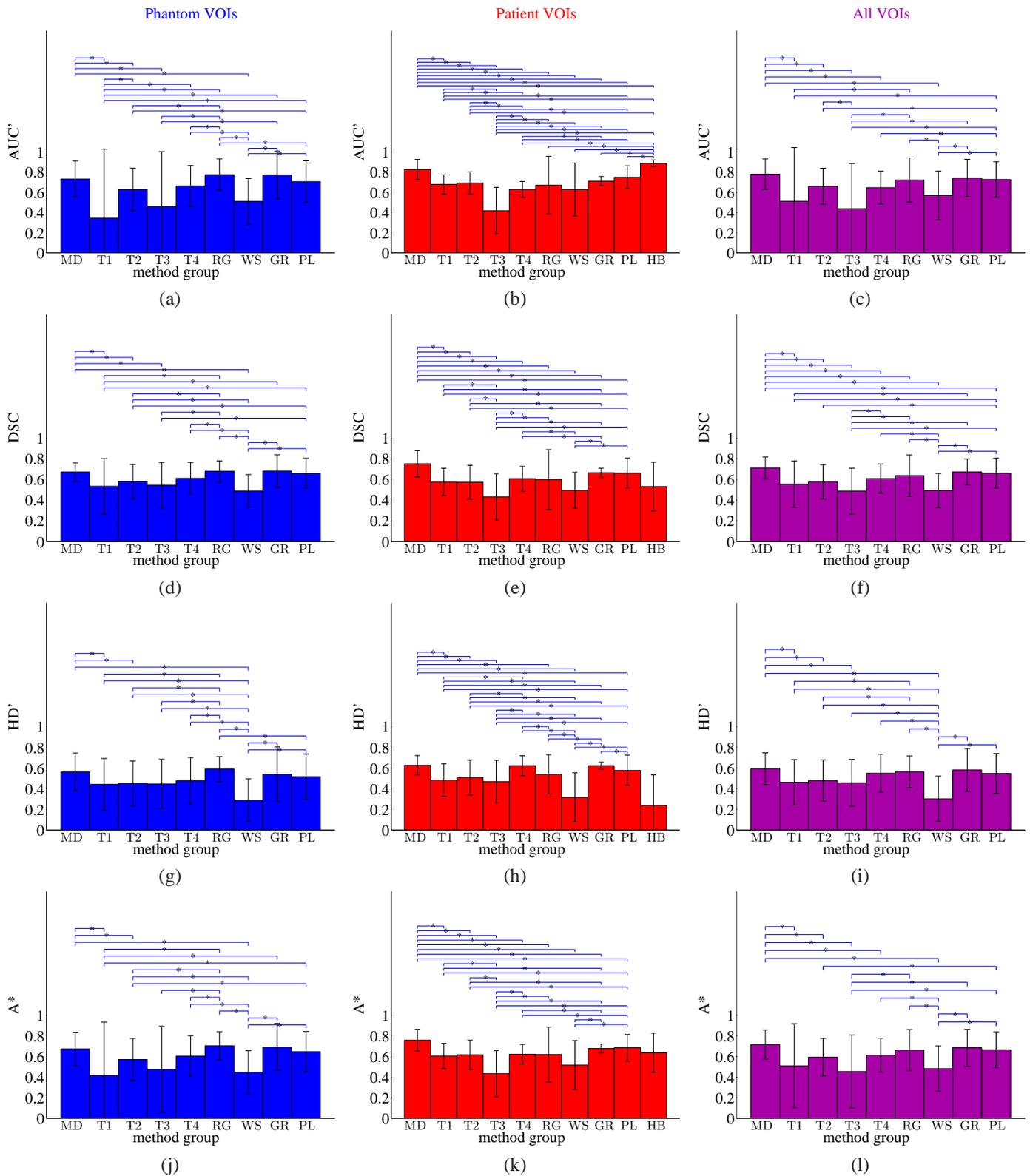


Fig. 8: Contouring accuracy of all algorithm types measured by *top row*: AUC' for (a) phantom, (b) patient and (c) both VOI types, *second row*: DSC for (d) phantom (e) patient and (f) both image types, *third row*: HD' for (g) phantom, (h) patient and (i) both image types and *bottom row*: A* for (j)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

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 4 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 (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 users in the cases of methods $MD^{b_{1,2}}$, $RG^{b_{1,2}}$ and $RG^{c_{1,2}}$ are counted as two individual results so there are a total of $n = 32$ 'methods', or $n = 33$ 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 $n_i = n_j = n_{VOIs}$, where mean accuracy μ_i and μ_j and standard deviation σ_i and σ_j are calculated over all VOIs and there are $2n_{VOIs} - 2$ degrees of freedom. As above, we repeat for all image sets and accuracy metrics. Figure 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 figure 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 most accurate are another manual method (MD^{b_2}) and the 'smart opening' algorithm (PL^b) with mid-level interactivity.

For phantom VOIs only, methods RG^b and $T1^b$, 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^{b_2} being significantly less accurate than the manual method MD^c (see supplementary material). Method $T1^b$ is based on thresholding and showed

low accuracy for patient VOIs, being significantly less accurate than the manual methods MD^c and MD^{b_2} (see supplementary material). Their high accuracy in phantom images alone could be explained by methods $T1^b$ and RG^b being particularly suited to the relative homogeneity of the phantom VOIs.

Methods WS^a , $T1^c$ and $T3^b$ have the 3 lowest accuracies by mean A^* across all 3 image sets. The poor performance of method WS^a could be explained by its origins (colour photography and remote-sensing) and user having no roots or specialism in medical imaging. Threshold methods $T1^c$ and $T3^b$ 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 3 patient or 4 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^{b_1} . 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 minimised to reduce variability. Table VI reaffirms the finding that manual delineation is the most accurate method type, with examples MD^c and $MD^{b_{1,2}}$ scoring highly in all metrics. The most consistently accurate non-manual 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 experiments in [13] that showed favourable results over PET alone.

A. Manual delineation

Free-hand segmentation produced among the most accurate results, which may be counter-intuitive. One explanation

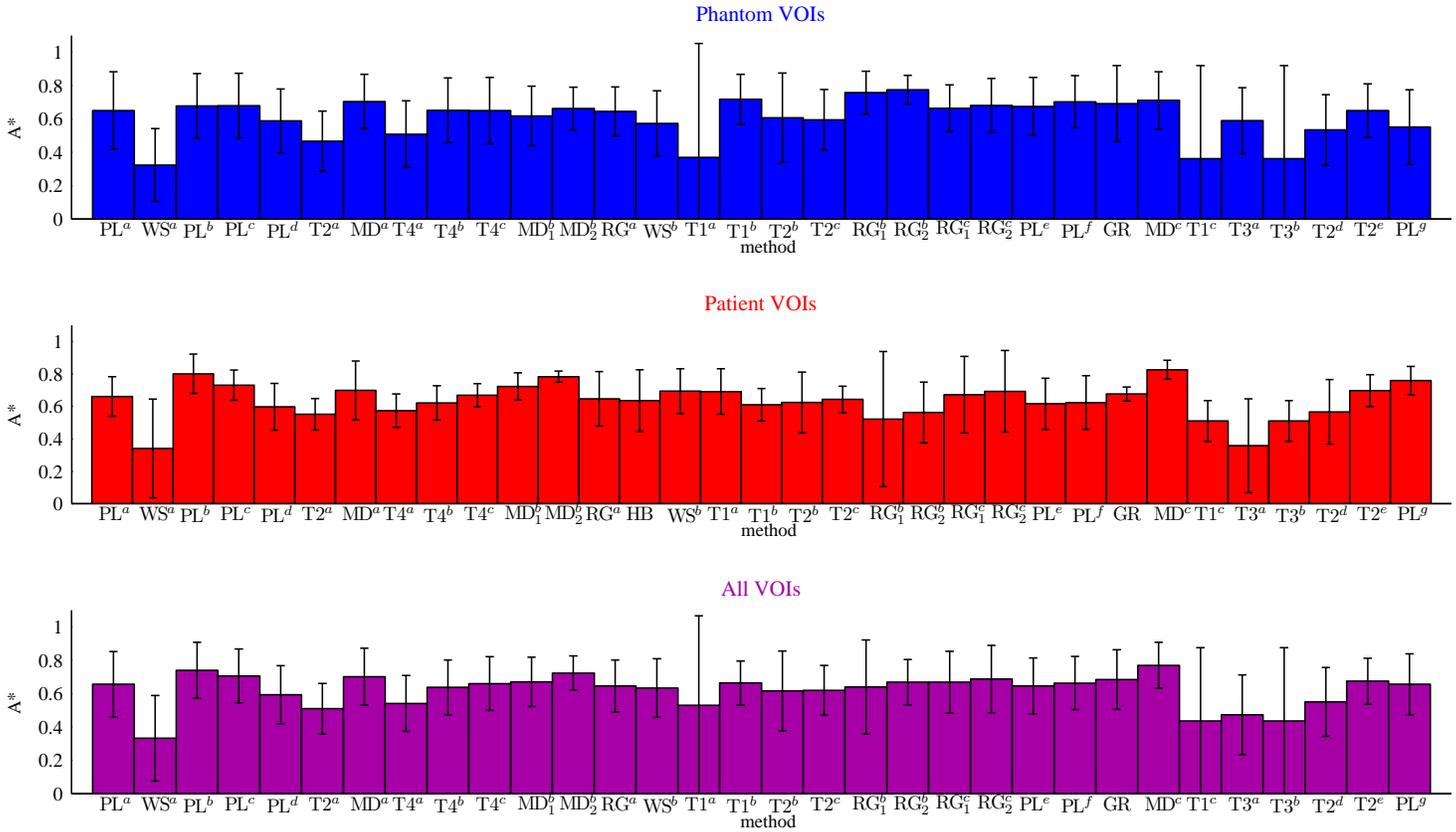


Fig. 9: Mean accuracy measured by A^* , of each method used to contour VOIs in phantom (top), patient (middle) and the combined image set (bottom).

847 comes from the incorporation of prior knowledge regarding
 848 the likely form and extent of pathology. In the case of the
 849 patient images alone, bias toward MD may be suspected as
 850 the ground-truth set is also built up from manual delineations.
 851 However, this does not explain the success of manual methods
 852 as they performed better still for phantom VOIs where the
 853 ground truth comes from CT thresholds. The use of multiple
 854 ground truth estimates by I-ROC may falsely favour manual
 855 delineation due to its inherent variability. However, this too
 856 does not explain the success of manual methods as they also
 857 perform well in terms of DSC and HD' that use a unique,
 858 'best-guess' of ground truth (at least one MD is among the
 859 5 highest DSC and HD for each of the patient phantom VOI
 860 sets). These observations challenge the intuition, that manual
 861 delineation is less accurate. Although many (semi-)automatic
 862 methods out-perform free-hand delineation in the literature,
 863 the inherent bias toward positive results among published work
 864 makes this an unfair basis for intuition.

865 Of the 4 manual delineations (MD^a , $MD^b_{1,2}$ and
 866 MD^c), method MD^c out-performed the rest in all of $n(n.s.d)$,
 867 $n(> \mu + \sigma)$, median rank and intra-operator variability where
 868 known, with significant improvement over $MD^b_{1,2}$ in terms of
 869 AUC^c for patient VOIs (although the multiple comparison ef-
 870 fect can mean that one or more of these differences are falsely
 871 detected as significant). The obvious difference between these
 872 4 is the user. It is interesting, and indicative of no bias in

873 terms of user group, that the delineator of MD^c was a nuclear
 874 medicine physicist while the other users, in common with the
 875 experts providing ground truth estimates, were experienced
 876 physicians. However, while users of MD^a and $MD^b_{1,2}$ only
 877 viewed the PET images during delineation, the physicist using
 878 MD^c also viewed an overlay of the PET gradient magnitude
 879 and, in the case of patient images, simultaneous CT. These
 880 modes of visual guidance could in part compensate for the
 881 relative lack of clinical experience, although no concrete
 882 conclusion can be made as clinical sites may disagree on the
 883 correct segmentation.

884 B. Automation vs. user guidance

885 Two method comparisons provide evidence that too much
 886 automation in a semi-automatic algorithm is detrimental to
 887 contouring accuracy. First, we compare the accuracy of meth-
 888 ods PL^c and PL^d . Method PL^d starts with the same seg-
 889 mentation achieved by PL^c , then performs extra steps in
 890 the automatic pipeline intended to improve on the results.
 891 However, these extra steps reduce the final accuracy. Second,
 892 we compare the accuracy of methods $RG^b_{1,2}$ and $RG^c_{1,2}$.
 893 These differ in that $RG^b_{1,2}$ also employs post-editing by the
 894 adaptive brush tool. While the adaptive brush may improve
 895 accuracy for phantom VOIs, accuracy is reduced for patient
 896 VOIs indicated by $n(n.s.d)$ and median rank. This suggests

TABLE VI: Summarised 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

interactive post-editing by user-defined watershed markers and sub-regional 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 5 thresholding schemes used at the same institution (team 13). Methods T1^c, T3^a and T3^b use intensity thresholds of 50% maximum and 41% & 50% of maximum-plus-background, while T2^d and 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, T2^e is most highly influenced by the user and ranks consistently higher than the other 4 in all 12 combinations of accuracy metric and image set, significantly out-performing T1^c once, T3^b twice and 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 off-line, allowing the medical professional to work on other parts of a treatment in parallel.

C. Building prior knowledge into contouring

As already seen from figure 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 out-perform 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 tumour 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 3 thresholding schemes T4^a, T4^b and T4^c used by the same institution (team 04). Of these, method T4^a

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

was considerably less accurate in terms of both n(n.s.d) and median rank. Methods $T4^b$ and $T4^c$ were calibrated using phantom data to build in prior knowledge of the imaging device. Even though the two devices used to calibrate $T4^b$ and $T4^c$ 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 $T4^a$ implemented at the same site, which does not learn from scanner characteristics but instead has an arbitrary parameter (39%). Methods $T4^b$ and $T4^c$ also out-perform the majority of the other low-interactivity thresholding schemes, suggesting that the calibration is beneficial and generalises across imaging devices. This apparent generalisation is further evidenced by no significant differences between methods $T4^b$ and $T4^c$ in any individual metric for patient or phantom VOIs.

Finally, the low accuracy of methods $T4^a$ and $T4^a$ 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 4 metrics in contouring both phantom and patient VOIs. These low accuracies are likely to arise from the spherical assumption rather than the initialisation of the method, as the low accuracies are similar despite different methods of initialisation 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. [53], [19], [54]). 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 sub-set 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.* [55]. 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 mimic tumours and metastases in head and neck cancer [10], [12], lung nodules [56] and gliomas [20] and cylindrical VOIs, attempting to mimic tumours [37]. 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 tumours 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 [57]. While this approach could provide the standard for accuracy evaluation where available, histology-based accuracy measurement is currently limited as described in [58], with errors introduced by deformation of the organ and co-registration of digital images (co-registration in [57] required first registering manually to an intermediate CT image). Furthermore, tumour excision is only appropriate for some applications. For head-and-neck cancer, the location of the disease often calls for non-invasive, *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 $n(>\mu+\sigma)$ 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. CONCLUSIONS

The multi-centre, 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 out-performed all semi- or fully-automatic methods. However, two methods ($T2^b$ 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 (i) guiding manual practices to reduce levels of variability and (ii) not over-influencing the expert or overriding his or her knowledge. To strike this balance, techniques that show promise are (i) visual guidance by both CT and PET-gradient images, (ii) model-based handling of heterogeneity and blurred edges that characterise oncological VOIs in PET and (iii) 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 test images and ground truth sets [49] 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 $T1^a$ made any assumptions about the tracer or tumour site so results for the remaining methods tested here provide a benchmark for future comparisons. Recently proposed methods in [11], [12] and [59] 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 categorise 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 generalise for other tumour types and PET tracers.

VII. ACKNOWLEDGEMENTS

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.

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.1 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 [60]. Data comes in the form of $N = N_+ + N_-$ 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) 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 N_{TP} , N_{FP} , N_{TN} and N_{FN} , of true/false positives and negatives yield the true positive ratio TPR_i and false positive ratio FPR_i for the i^{th} threshold and the pair $\{TPR_i, FPR_i\}$ 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 figure 10, the limits are the minimum and maximum value in all N data. The fixed ground truth in figure 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 figure 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 (\dots in figure 10) has $AUC = 1$ 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 minimises 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

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

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 [61].

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.

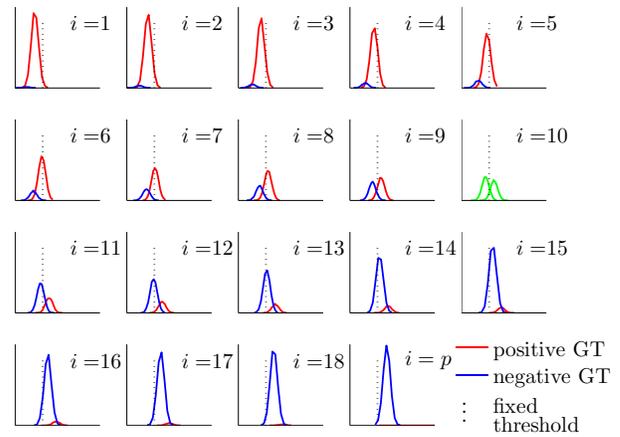
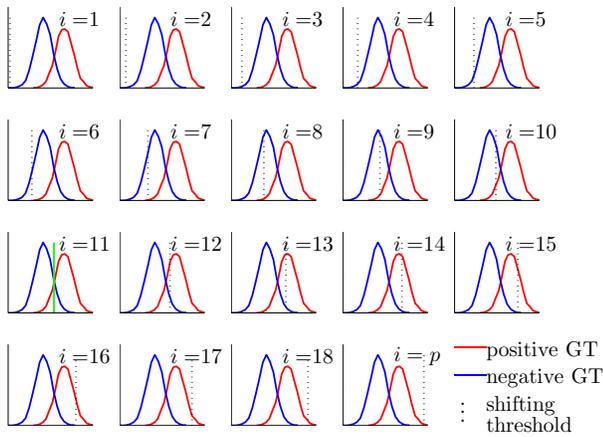


Fig. 10: Conventional ROC analysis of a threshold classifier performed by fixing the ground truth labelling and varying the threshold in $p = 19$ increments (top) to form a ROC curve (bottom). The operating point is marked green.

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 $\{TPR_i, FPR_i\}$ pairs to form a ROC curve (bottom). The operating point is marked green.

A.2 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 makers, it assumes a single classification and varies the definition of ground truth. Figure 11 demonstrates this for the example of threshold classification. In common with figure 10, data being classified in figure 11 are a mixture of Gaussians with means μ_- and μ_+ separated by 6 units and standard deviations $\sigma_- = \sigma_+ = 2.5$, and the total number of data is fixed at $N_+ + N_- = 2 \times 10^4$. 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 i^{th} ground truth labelling, we fix the threshold at $T=0$ for all $i \in \{1 \dots p\}$. In line with the requirements of conventional ROC, the multiple ground truth definitions are

- 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, and
- A.2(iv) 'pass through' the unique (unknown) ground truth as closely as possible.

Requirement A.2(iii) is realised by fixing the difference of means $\mu_+ - \mu_-$ and having ρ increase with μ_+ . Requirement A.2(iv) means that there exist labellings $\{\mathcal{GT}_i\}$ 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 figure 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 figure 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).

A.3 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

1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196

1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217

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, and

A.3(iv) 'passes through' the un-known, 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. Figure 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 $AUC < 1$, equivalent to the case of overlapping histograms in figure 11, although the similar form of the curve and value of $AUC = 0.98$ 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 $AUC = 1$, 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 generalise for any shape of ground truth set or contour under evaluation if

$$N(\in \mathcal{GT}_j) = \sum_{i=1}^j N(\mathcal{GT}_i \vee \mathcal{GT}_j) \quad \text{and} \quad (6)$$

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

where $N(\in \mathcal{GT}_j)$ and $N(\notin \mathcal{GT}_j)$ denote the number of voxels inside and outside the j^{th} ground truth definition. Equation

6(a) holds if requirement A.3(i) is met and 6(b) is satisfied by the fixed bounding box enclosing the set $\{\mathcal{GT}_i\}$.

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 radiation-therapy planning for head-and-neck squamous cell carcinoma." *International Journal of Radiation Oncology Biology Physics*, vol. 66, p. 185, 2007.
- [2] C. Nutting, "Intensity-modulated radiotherapy (IMRT): the most important advance in radiotherapy since the linear accelerator?" *British Journal of Radiology*, vol. 76, p. 673, 2003.
- [3] J. W. Keyes, "SUV: Standardised Uptake or Silly Useless Value?" *Journal of Nuclear Medicine*, 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," *Journal of Nuclear Medicine*, 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." *Molecular Imaging Biology*, vol. 4, pp. 171–178, 2002.
- [6] J. A. van Dalen, "A novel iterative method for lesion delineation and volumetric quantification with FDG PET," *Nuclear Medicine Communications*, vol. 28, pp. 485 – 493, 2007.
- [7] J. F. Daisne, M. S. and A. Bol, T. D. M. Lonnew, and V. Grégoire, "Tridimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms," *Radiotherapy & Oncology*, 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," *European Journal of Nuclear Medicine and Molecular Imaging*, 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," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 37, 2010.
- [10] X. Geets, J. A. Lee, A. Bol, M. Lonnew, and V. Grégoire, "A gradient-based method for segmenting FDG-PET images: Methodology and validation," *European Journal of Nuclear Medicine and Molecular Imaging*, 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," *Medical Physics*, vol. 34, 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," *Medical Physics*, 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 Transactions on Medical Imaging*, vol. 28, pp. 374 – 383, 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) Lecture Notes in Computer Science*, vol. 6801, 2011, pp. 245 – 256.
- [15] S. Belhassen and H. Zaidi, "A novel fuzzy c-means algorithm for unsupervised heterogeneous tumor quantification in pet," *Medical Physics*, vol. 37, pp. 1309 – 1324, 2010.
- [16] M. Hatt, C. C. le Rest, P. Descourt, A. Dekker, D. D. Ruyscher, M. Oellers, P. Lambin, O. Pradier, and D. Visvikis, "Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications," *International Journal of Radiation Oncology Biology Physics*, vol. 77, pp. 301 – 308, 2010.
- [17] J. A. Lee, "Segmentation of positron tomography images: Some recommendations for target delineation in radiation oncology," *Radiotherapy and Oncology. Special Issue: PET in Radiotherapy Planning*, vol. 96, pp. 302 – 307, 2010.

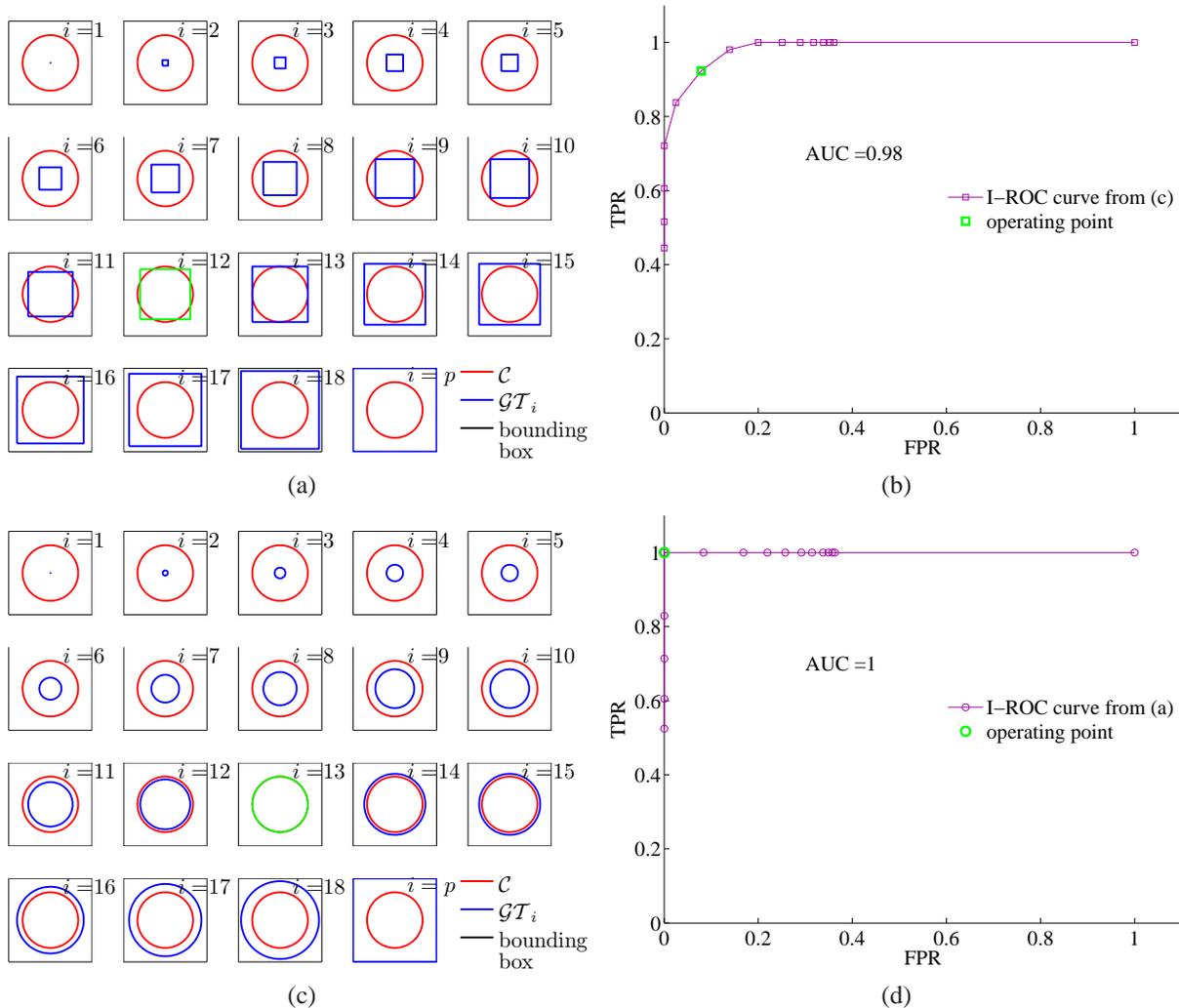


Fig. 12: Inverse-ROC analysis of a fixed contour (red circle) performed by varying ground truth contours as squares (a) or circles (c) of increasing size. ROC curves in (b) and (d) are built from the corresponding true and false counts that lie inside or outside the i^{th} ground truth contour. Operating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

1341 [18] U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welsch, D. Hell- 1367
 1342 wig, C. Rube, and C. Kirsch, "Comparison of different methods for 1368
 1343 delineation of 18F-FDG PET-positive tissue for target volume definition 1369
 1344 in radiotherapy of patients with nonsmall cell lung cancer," *Journal of 1370
 1345 Nuclear Medicine*, vol. 46, 2005. 1371
 1346 [19] C. Greco, S. A. Nehmeh, H. Schöder, M. Gönen, B. Raphael, H. E. 1372
 1347 Stambuk, J. L. Humm, S. M. Larson, and N. Y. Lee, "Evaluation of 1373
 1348 different methods of 18f-fdg-pet target volume delineation in the 1374
 1349 radiotherapy of head and neck cancer," *American Journal of Clinical 1375
 1350 Oncology*, 2008. 1376
 1351 [20] H. Veess, S. Senthambichelvan, R. Miralbell, D. C. Weber, O. Ratib, 1377
 1352 and H. Zaidi, "Assessment of various strategies for 18F-FET PET-guided 1378
 1353 delineation of target volumes in high-grade glioma patients," *European 1379
 1354 Journal of Nuclear Medicine and Molecular Imaging*, vol. 36, pp. 182 1380
 1355 – 193, 2009. 1381
 1356 [21] A. C. Riegel, A. M. Berson, S. Destian, T. NG, L. B. Tena, R. J. Mitnick, 1382
 1357 and P. S. Wong, "Variability of gross tumor volume delineation in head- 1383
 1358 and-neck cancer using CT and PET/CT fusion," *International Journal of 1384
 1359 Radiation Oncology Biology Physics*, vol. 65, pp. 726 – 732, 2006. 1385
 1360 [22] S. L. Breen, J. Publicover, S. De Silva, G. Pond, K. Brock, B. O 1386
 1361 Sullivan, B. Cummings, L. Dawson, A. Keller, J. Kim, J. Ringash, E. Yu, 1387
 1362 A. Hendler, and J. Waldron, "Intraobserver and interobserver variability 1388
 1363 in GTV delineation on FDG-PET-CT images of head and neck cancers," 1389
 1364 *International Journal of Radiation Oncology Biology Physics*, vol. 68, 1390
 1365 pp. 763 – 770, 2007. 1391
 1366 [23] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, 1392
 V. Jewells, and S. W. (Eds.), "Workshop proceedings, 3D segmentation 1367
 in the clinic: A grand challenge II - MS lesion segmentation," 1368
 2008, medical Image Computing and Computer-Assisted Intervention 1369
 (MICCAI). [Online]: [http://grand-challenge2008.bigr.nl/proceedings/](http://grand-challenge2008.bigr.nl/proceedings/mslesions/articles.html) 1370
 mslesions/articles.html 1371
 [24] X. Deng and G. D. (Eds.), "Workshop proceedings, 3D segmentation 1372
 in the clinic: A grand challenge II - Liver tumour segmentation," 1373
 2008, medical Image Computing and Computer-Assisted Intervention 1374
 (MICCAI). [Online]: [http://grand-challenge2008.bigr.nl/proceedings/](http://grand-challenge2008.bigr.nl/proceedings/liver/articles.html) 1375
 liver/articles.html 1376
 [25] L. R. Dice, "Measures of the amount of ecologic association between 1377
 species," *Ecology*, vol. 26, p. 97//302, 1945. 1378
 [26] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images 1379
 using the Hausdorff distance," *IEEE Transactions on Pattern Analysis 1380
 and Machine Intelligence*, vol. 15, pp. 850 – 863, 1993. 1381
 [27] N. Hata, G. Fichtinger, S. Oguro, H. Elhawary, and 1382
 T. van Walsum, "Prostate segmentation challenge 2009," 2009, 1383
 medical Image Computing and Computer Assisted Intervention 1384
 (MICCAI) Workshop: 3D Segmentation Challenge for Clinical 1385
 Applications. [Online]: [http://wiki.na-mic.org/Wiki/index.php/2009_](http://wiki.na-mic.org/Wiki/index.php/2009_prostate_segmentation_challenge_MICCAI) 1386
 prostate_segmentation_challenge_MICCAI 1387
 [28] V. Pekar, J. Kim, S. Allaire, A. Qazi, and D. A. Jaffray, "Head 1388
 and neck auto-segmentation challenge 2010," 2010, medical Image 1389
 Computing and Computer Assisted Intervention (MICCAI) Workshop: 1390
 Medical Image Analysis for the Clinic: a Grand Challenge. [Online]: 1391
www.grand-challenge2010.ca/ 1392

- [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," *European Journal of Nuclear Medicine and Molecular Imaging*, 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 head-and-neck cancer using region-based texture analysis of PET and CT images," *International Journal of Radiation Oncology Biology Physics*, vol. 75, pp. 618 – 25, 2010.
- [32] J. Daisne, T. Duprez, B. Weynand, M. Lonnew, M. Hamoir, H. Reyckler, 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?" *Journal of Nuclear Medicine*, vol. 48, p. 17611766, 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 (supplement)*, vol. 80, pp. 2505 – 2509, 1997.
- [36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 62 – 66, 1979.
- [37] J. F. Daisne, M. Sibomana, A. Bol, G. Cosnard, M. Lonnew, and V. Grégoire, "Evaluation of a multimodality image (CT, MRI AND PET) coregistration procedure on phantom and head-and-neck cancer patients: accuracy, reproducibility and consistency," *Radiotherapy & Oncology*, vol. 69, pp. 237 – 245, 2003.
- [38] ABX Advanced Biochemical Compounds, "ROVER: ROI Visualisation, Evaluation and Image Registration," 2010. [Online]: www.abx.de/rover/index.php/id-3d-regions-of-interest.html
- [39] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 641 – 647, 1994.
- [40] S. Beucher and F. Meyer, "The morphological approach to segmentation: The watershed transformation," *Mathematical Morphology in Image Processing*, pp. 433 – 482, 1993.
- [41] S. Lefèvre, "Knowledge from markers in watershed segmentation," in *Proceedings, IAPR International Conference on Computer Analysis of Images and Patterns. Lecture Notes in Computer Sciences*, vol. 4673, 2007, pp. 579–586.
- [42] 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 laparoscopic surgery," in *Proceedings, IEEE International Conference on Image Processing*, 2011.
- [43] 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. Lecture Notes in Computer Sciences*, 2010, pp. 117 – 128.
- [44] 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 Transactions on Medical Imaging*, vol. 25, pp. 417 – 434, 2006.
- [45] C. M. Li, C. Y. Kao, C. John, and Z. H. Ding, "Minimization of region-scalable fitting energy for image segmentation," *IEEE Transactions on Image Processing*, vol. 17, pp. 1940 – 1949, 2008.
- [46] T. F. Chan, S. Esedoglu, and M. Nikolova, "Algorithms for finding global minimizers of denoising and segmentation models," *SIAM Journal on Applied Mathematics*, vol. 66, pp. 1632 – 1648, 2006.
- [47] J. A. Lee, X. Geets, V. Gregoire, and A. Bol, "Edge-preserving filtering of images with low photon counts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1014 – 1027, 2008.
- [48] S. Li, "Markov random field models in computer vision," in *Computer Vision ECCV '94*, ser. Lecture Notes in Computer Science, J.-O. Eklundh, Ed. Springer Berlin / Heidelberg, 1994, vol. 801, pp. 361–370.
- [49] T. Shepherd, "Contour Evaluation," 2011, hybrid & Dynamic PET for Radiation Oncology, Turku PET Centre web pages. [Online]: www.turkupetcentre.net/PETinTreatmentPlanning/
- [50] T. T. Tanimoto, *IBM Internal Report*, 1957.
- [51] A. Tversky, "Features of similarity," *Psychological Reviews*, vol. 84, p. 327 352, 1977.
- [52] R. Real and J. M. Vargas, "The probabilistic basis of jaccard's index of similarity," *Systematic Biology*, vol. 45, p. 385 390, 1996.
- [53] 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, p. 996 1019, 2007.
- [54] 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.
- [55] 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 Transactions on Medical Imaging*, vol. 23, pp. 903 – 921, 2004.
- [56] Q. Black, I. Grills, L. Kestin, C. Wong, J. Wong, A. Martinez, and D. Yan, "Defining a radiotherapy target with pet," *International Journal of Radiation Oncology Biology Physics*, 2004.
- [57] 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 ¹¹C-choline PET scans for target volume definition in radical prostate radiotherapy," *Radiotherapy & Oncology*, vol. 99, pp. 187 – 192, 2011.
- [58] 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?" *Journal of Nuclear Medicine*, vol. 44, pp. 30 –32, 2003.
- [59] 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 Transactions on Medical Imaging*, vol. 28, pp. 881 – 893, 2009.
- [60] J. A. Swets, "ROC analysis applied to the analysis of medical imaging techniques," *Investigative Radiology*, vol. 14, pp. 109 – 121, 1979.
- [61] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145 – 1159, 1997.