Impact of consensus contours from multiple PET-segmentation methods on the accuracy of functionalvolume delineation

A. Schaefer, Maximilien Vermandel, C. Baillet, As Dewalle-Vignon, R. Modzelewski, P. Vera, L. Massoptier, C. Parcq, D. Gibon, T Fechter, et al.

▶ **To cite this version:**

A. Schaefer, Maximilien Vermandel, C. Baillet, As Dewalle-Vignon, R. Modzelewski, et al.. Impact of consensus contours from multiple PET-segmentation methods on the accuracy of functionalvolume delineation. European Journal of Nuclear Medicine and Molecular Imaging, 2015. hal-01233427

HAL Id: hal-01233427
https://hal.science/hal-01233427

Submitted on 25 Nov 2015

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.

Impact of consensus contours from multiple PET-segmentation methods on the accuracy of functionalvolume delineation

A. Schaefer[a*], M. Vermandel[b,c*], C. Baillet[c], A.S. Dewalle-Vignion[b], R. Modzelewski[d], P. Vera[d], L. Massoptier[e], C. Parcq[e] D. Gibon[e], T. Fechter[f], U. Nemer[f], I. Gardin[d**], U. Nestle[f**]

aDepartment of Nuclear Medicine, Saarland University Medical Centre, 66421 Homburg, Germany
bUniv. Lille, Inserm, CHU Lille, U1189 - ONCO-THAI - Image Assisted Laser Therapy for Oncology, F-59000 Lille, France
c CHRU de Lille, Nuclear Medicine Dpt, F-59037, Lille, France
dCentre Henri-Becquerel and LITIS EA4108, F-76000, Rouen, France
eAQUILAB, Research and Innovation Dpt, F-59120, Loos Les Lille, France
fDepartment of Radiation Oncology, University Hospital Freiburg , Germany

* These authors contributed equally to the study
** These authors contributed equally to the study

Corresponding Author:
Dr. Maximilien Vermandel
INSERM U1189 ONCO-THAI
CHRU de Lille – Université de Lille
1, avenue Oscar Lambret
59037 Lille
France

m-vermandel@chru-lille.fr

Keywords:
PET image segmentation, consensus algorithms, STAPLE, Radiation oncology,

Compliance with ethical standards

- **Conflict of interest**:None.
- **Funding**: This work was partially supported by EU project E5949 SALOME under Eurostars Program, which is powered by EUREKA and the European Community.
- **Research involving human participants**: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.However, for this type of retrospective study formal consent is not required.

Acknowledgement:
A.Schaefergreatlyacknowledges the valuable support of PhD Dr. Y.-J. Kim, department of pathology, Saarland University Medical Centre, in preparing the pathological reference database of centre-1. U.Nestle would like to thank C. Doll for assistance in analysing the data of centre-2.

Abstract:

Purpose: This study aimed to evaluate the impact of consensus algorithms on segmentation results when applied on clinical PET images. In particular, how majority vote or STAPLE algorithms could improve the final result in terms of accuracy and reproducibility when combining three semi-automatic segmentation algorithms.

Methods: Three published approaches of segmentation (contrast-oriented, possibility theory and adaptive thresholding) and two consensus algorithms, majority vote and STAPLE,were implemented in a single software platform (Artiview®). Four clinical datasets including different locations (thorax, breast, abdomen) or pathologies (NSCLC primary tumours, metastasis, lymphoma) were used to evaluate accuracy and reproducibility of the consensus approach in comparison with pathology ground truth or CT – ground truth surrogate.

Results:Our results reflect the variable performance of individual segmentation algorithms for lesions of different tumour entities that is for PET images that differ in resolution, contrast and image noise. Independent on location and pathology of the lesion, however, the consensus method displays improved volume segmentation accuracy compared to the worst performing individual method in the majority of cases and is close to the best performing method in many cases. In addition, the implementation reveals high reproducibility of the segmentation results against small changes in the respective starting conditions.No significant difference between STAPLE and majority vote algorithms was found.

Conclusion:This study shows that combining different PET-segmentation methods by application of a consensus algorithm offers robustness against the variable performance of individual segmentation methods andistherefore useful for radiation oncology purposes. It might also be relevant for other scenarios like the joining of expert recommendations in clinical routine and trials or thegeneration of multi-observer generated contours for standardisation of automatic contouring.

Keywords: ^{18}FDG PET, image segmentation, STAPLE, radiation oncology

I. Introduction

The use of molecular imaging methods in radiation oncology has become a routine procedure providing valuable information in radiotherapy treatment planning and beyond. For many malignancies, the beneficial effects of fluorodeoxyglucose (^{18}FDG) Positron Emission Tomography (PET) imaging has been shown, e.g. in the delineation of the gross tumour volume (GTV). However, due to technical and biological factors, tumours as depicted by PET appear blurred, heterogeneous, and often in a rather noisy background which hampers thesegmentation of reliable manual contours as well as the development and validation of automatic segmentation tools.

In times of increasing radiotherapy treatment precision leading to high rates of local control with minimum toxicity once reliable tumour targeting has been achieved, the correct depiction of tumour tissue is of utmost importance. However, due to the shortcomings of anatomical imaging by CT and the often superior diagnostic accuracy of molecular imaging by PET, its use is highly desirable in this context. Therefore, many groups have addressed the problem of PET segmentation in recent years proposing different segmentation approaches.The main challenging task of any segmentation algorithm in itself, however, is its validation.

Among semi-automatic PET-segmentation methods one can underline two main classes of approach: threshold based and image processing based. Threshold based segmentation methods are used for lesion delineation because of their simplicity. In this context the segmentation process relies on an intensity threshold above which all voxels are considered to belong to the tumour volume. This threshold can either be fixed [1-4] or depending on some features measured on the image (e.g.: Standard Uptake Value, SUV), background noise, signal-to-noise ratio, image contrast). In the latter case, the threshold is adaptive and needs to be determined - mostly iteratively - by specific algorithms including prior calibration of the PET device[5-13].

To tackle low contrast and heterogeneity of PET images and to avoid prior calibration of the PET system, more advanced approaches have been investigated including watershed segmentation [14-16], gradient based approach [17], clustering approach [18, 19], possibility theory [20, 21] or bayesian framework [22, 23]. Based on image processing theory and clustering approaches, these methods offer the possibility to delineate uptakes semi-automatically without prior calibration.

In different investigations including phantom studiesand/or clinical data [24, 25]many methods revealed their advantages and also their own specific weaknesses. Moreover, the accuracy of lesion segmentation by a given algorithm to a given clinical case was shown to highly depend on its software implementation, user interaction and last but not least on technical factors of the PET system in use. This may be a hint that depending on the varying clinical conditions it will never be possible to select one “perfect” automatic method.

In order to overcome those shortcomings, it may make sense to combine severalindividualautomatic and semi-automatic methods applying a consensus method as it has been proposed recently in MRI imaging [26, 27]. This may helpto exploit the advantages of the different algorithmswhile minimizing their disadvantages.

The easiest method is to apply the majority vote rule which decides if voxels belong to the lesion or not according to the results of the majority of the individual segmentation methods [28]. Recently, the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm wasproposed in the literature [29, 30], which computes a probabilistic estimate of the ground truth from a collection of segmentation results. To assure an optimal clinical workflow both, the individual segmentation methods as well as the consensus algorithmneed to be implemented on the same workstation.

A previous study by McGurk et al. [31]introduced the concept of applying consensus methods to PET segmentation. These authors investigated the use of two methods, simple majority voting and probabilistic estimation, to combine five segmentation approaches on PET phantom measurements.Both methods were found to improve the segmentation accuracy when combining volumes and to offer robustness against the variable performance of individual methods. The aim of the present paper was to validate clinically the feasibility of combining different segmentation algorithms by the concept of consensus onmulticentre clinical PET data. Investigationof the impact of the two consensus algorithms (majority vote (MV) and STAPLE) on the segmentation of ^{18}FDG PET-positive lesionswas performed in terms of accuracy of segmentation and robustness of the consensus contour. Patient data of different tumour entities representing a variety of lesions that differ in biology, size and body location were available to validate the clinical feasibility of the consensus approach. Three segmentation methods developed by the authors were used as entry of the consensus algorithms: possibility theory based approach [20], contrast oriented approach [13] and threshold oriented approach [32]. Consensus and segmentation algorithms were implemented on the same software platform.

I. Material and methods

a) Description of the software

The concept of combining several segmentation methods by a consensus algorithm within a clinical workflow was implemented as a part of the software package Artiview® (Aquilab, France). This software package allows experts to review, compare, evaluate and assess multimodality imaging and radiotherapy treatments. For PET-segmentation, three individual methods as well as two consensus algorithmswereimplemented.To process, a PET sub-volume is created by the user which roughly envelops the lesion (mask, 3D-box). The automated methods and the consensus will be applied simultaneously within this mask to calculate the resulting PET volumes.All contours can be compared visually as well as by use of different metrics (e.g. Dice similarity index (DSC; Eq. (4)) or Percent error (Eq. (5)) which were also applied in the current evaluation. The integration followed up a co-design process, that is, allend-users were fully involved in the interface design.

b) Individual segmentation methods

The following individual segmentation methods were implemented in one single softwarepackage. Because computer science may involve different implementation processes

(different programming language, code optimization achieved by computer scientist, floating precision used, stochastic formulation used) or simply differences in mathematical procedures like different points-of-origin for the definition of voxel coordinates,segmentation results might slightly differ between several software packages depending on the implementation. Thus, the implementation of different algorithms into one single software system needed to be approved in comparison with the native lab's software. This was done by phantom measurements asdescribedin [33]. For all three methods, agreement was reported in terms of the mean percent error in delineated volume(Eq. (5); <3.2% for all three methods) and/or the mean DSC ((Eq. (4); >0.92 for all three methods).

i) Contrast-oriented algorithm(COA)

The contrast-oriented algorithm is an adaptive thresholding algorithm for the FDG-PET-based delineation of tumour volumes [13] which uses two parameters to calculate the threshold for auto-contouring a volume in the FDG-PET data: (i) The mean standardized uptake value of the 70%-of- SUV_{max} –isocontour of the object to characterize the mean SUV of this object ($mSUV_{70\%}$) and (ii) the background-SUV surrounding the object (BG). The relationship between the optimal threshold, TS, and the image contrast determined by a regression analysis [13]results in the following threshold equation:

$$TS = A \cdot mSUV_{70\%} + B \cdot BG \quad (1)$$

The values of parameters A and B are known to be specific for the PET system applied in combination with the predefined imaging protocol [34]. Therefore, the use of the contrast-oriented algorithm requires a system-specific calibration by phantom measurements described in the respective original publications [13, 34].

ii) Possibility theory based method (POS)

In 2008, Dewalle et al. introduced a nearly automatic and operator independent method for volume segmentation on PET images [21]. The method relies on two key points. First, the use of the Maximum of Intensity Projections (MIP) algorithm enabled the usually poor PET image contrast to be overcome. Then, a possibility theory [35] -based algorithm was developed to take account for the gradual transition between healthy tissues and volumes of interest (VOI), partially due to the poor spatial resolution of the PET images. Application of the possibility theory framework enabled to manage fuzzy value (included in $[0;1]$) instead of binary values ($\{0;\ 1\}$). This approach, which did not require prior calibration, remained independent of PET facilities.

iii) Adaptive threshold oriented method (ADP)

The native adaptive thresholding method has been described previously [32].Briefly, the optimal threshold value, Th_{Opt}, to segment the lesion follows the mathematical expression:

$$Th_{Opt} = A/Cont_{meas} + B \quad (2)$$

Where A and B are 2 constant parameters which need to be defined during a calibration procedure and that was described in detail in [32]. $Cont_{meas}$and Th_{Opt} are obtained following an iterative process. $Cont_{meas}$corresponds to a measured local contrast between the lesion and the background. For the background region, a shell surrounding the lesion is automatically delineated. The shell has a thickness of 2 voxels, and the inner edge of the shell was chosen to be 2 voxels away from the lesion boundary to limit the partial volume effect. The average grey level in the shell, B_{avg}, is computed, as well as $Cont_{meas}$, such as:

$$Cont_{meas} = Max_{avg} / B_{avg} \quad (3)$$

whereMax_{avg} is the maximum average value of a volume of 0.5 mL within the lesion.

c) **Consensus algorithms**

i) Majority vote (MV)

The majority vote rule is a simple consensus approach[26, 28, 36]. The volume is obtained by applying the majority vote rule: a voxel is considered to belong to the lesion according to the majority of the segmentation methods results.

ii) STAPLE

The Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm is an instance of the expectation-maximization (EM) algorithm proposed in 2004 by Warfield et al [29]. From a collection of segmentation results as input, STAPLE provides a probabilistic estimate of the ground truth and a measure of the performance level of each input. Recently, Commowick et al. [30] proposed a new version of the STAPLE algorithm in which a maximum a posteriori (MAP) estimate of the true segmentation is obtained by considering a beta prior probability for the performance levels.

The algorithm was firstly developed by the Computational Radiology Laboratory (CRL). This implementation is available via the CRKit software (http://crl.med.harvard.edu/). In order to evaluate the consensus methodology in an integrated system, the algorithm was implemented in Artiview®. The implementation was performed according to the method proposed in the original paper. To model a relationship of neighbouring voxels, a Markov Random Field was incorporated[30].

STAPLE algorithm involves several parameters, which can affect the quality of the consensus estimate. Mainly, the parameters (maximum number of iterations, convergence threshold, initial performance level of each input segmentation result, global prior probability) were set to their default values [29].To optimize the remaining parameters, a prior study on PET simulated data was performed. Results of this simulation study and the implementation of STAPLE are presented in detail in [37][37].

d) **Database description**

Implementation and clinical feasibility of the consensus algorithms were evaluated on four

patient cohorts that comprised different tumour entities and were provided by four different centres in Europe. All patients underwent routine ^{18}FDG PET or PET/CT applying the centre-specific clinical protocols which are summarized in Table 1. The corresponding imaging protocols are presented in Table 2. For each cohort of patients the segmentation results were compared with a dataset-specific ground truth as a reference as stated below.

i) Lung tumours (centre-1)

Ungated PET-data sets of twelve patients (four women and eight men, ranging in age from 56 to 79 years (mean age ± SD 65±7years) with histologically proven Non-Small Cells Lung Cancer (NSCLC) were included in this evaluation. Patient characteristics with respect to tumour localization and TNM classification were described in [38]. Within three weeks after PET examination all patients were treated with lobectomy and mediastinal lymph node dissection with curative intent. Lung lobes were laminated in a standardized manner receiving slices of 4-5mm thickness. Digitized macrophotographs of each slice were recorded and evaluated as described in [38] to estimate the pathological lung volume that was used as the reference standard throughout this study.

ii) Lung tumours (centre-2)

This dataset consists of a cohort of 9 patients (14 lesions; six women and three men, mean age: 67±5 years) with primary NSCLC or pulmonary metastases who were intented to to receive stereotactic body radiotherapy. 4D-CT and 4D-PET datasets were retrospectively gated in 10 bins based on the breathing curve provided by a pressure sensor belt. On both, CT and PET images, the manual contours of 4 experts were combined with a majority vote as a ground truth surrogate for PET (manualPETvote) and CT (manualCTvote). The evaluated algorithms were applied to all PET timebins and the resulting mean volume was compared to manualCTvote, unless otherwise stated.

iii) Lymphoma (centre-3)

Eight lymphoma patients (4 men and 4 women, ranging in age from 35 to 69 years, 5 follicular lymphoma, 2 refractory Hodgkin lymphoma and 1 transformed indolent lymphoma) who underwent routine whole-body ^{18}F-FDG PET-CT before initiation of a first or new line of treatment were retrospectively included. Ten abdominal nodal lesions including bulky lesions were chosen in those 8 patients: 4 in the hepatic hilum, 3 in the lumbo-aortic area, 2 coelio-mesenteric and 1 iliac node. These lesions were chosen according to their location and if their limits were delineable in each CT slice. The manual contour of one expert nuclear medicine physician on the CT of the PET-CT was used as a surrogate of ground truth.

iv) Breast tumours (centre-4)

Ten women with confirmed mammary Invasive Ductal Carcinoma (IDC) stage T2-T3 / M0 were prospectively included. The study was declared to the ClinicalTrials.gov Protocol

Registration System (PRS) (VoSeTep study, N°RCB: 2009-A00602-55). Patient characteristics with respect to tumour localization and TNM classification were described in [39]. All patients underwent procubitus ungated PET-CT acquisition, centred on the breast region, immobilized with a device fixing the mammalian gland to avoid tumour movement.

The surgery was performed 4 ± 3 days after the PET/CT examination of the patient. Surgical respected specimen was oriented to the in-vivo geometry. The specimen was sectioned with a macrotome (EH-170T, Sofraca, France) into 5 µm thick slices at 2 mm intervals. Digitized slides of each slice were recorded and evaluated as described in [39] to estimate the pathological volume that was used as the gold standard volume of the lesion.

e) **Data analysis**

i) Accuracy evaluation

In a first step, the accuracy of segmentation was analysed for each dataset in terms of the Dice similarity coefficient (DSC)[40]and the percent error which both compare the volumes delineated by the different algorithms with the corresponding ground truth.The DSC,which provides an index of the spatial overlap [40] between the estimated volume (e.g. STAPLE or MV output), V_{estim}, and the ground truth, V_r, is defined as:

$$Dice(V_{estim}, V_r) = \frac{2|V_{estim} \cap V_r|}{|V_{estim}| + |V_r|} \quad (4)$$

Where $|X|$ represents the size of the set X.

The percent error, which compares the estimated volumes of the segmentation results (e.g. STAPLE or vote output), V_{estim}^{ml}, expressed in ml, with the gold-standard volumes in ml, V_r^{ml}, is defined as:

$$PE = \frac{\left|V_r^{ml} - V_{estim}^{ml}\right|}{\left|V_r^{ml}\right|} \quad (5)$$

ii) Ranking

In a second step a ranking approach was applied to investigate if it is favourable in clinical routine to use a consensus-contour instead of the best performing method. In the present evaluation, a pathology ground truth or CT-ground truth surrogate is available offering a reliable reference for PET-based segmentation. Therefore, it is possible to investigate whether or not it is favourable to simply use the best-performing method all the time by ranking the segmented volumes relative to the respective ground truth for the three methods, the two consensus-approaches and each patient. Ranking needs to be done twice, with respect to the best-performing and to the worst-performing method. If one individual method is observed to

consistently provide the best segmentation and simultaneously not to provide the worst segmentation, then using a consensus approach may not offer any improvement over using this best method. All segmentation methods were therefore ranked according to both, the smallest and the largest difference of volume compared to ground truth (best and worst method, respectively). Taking into account the comparatively low resolution of PET imaging, differences in segmented volumes smaller than 2% corresponding to differences in calculated diameter smaller than 1mm (smaller than half a pixel) were disregarded. The number of times, $N_{i,p}$, each method (i) was ranked best (p=1) or worst (p=0), respectively, in a comparison was recorded and this number was normalized by the total number of comparisons (N_{tot}) made.

$$Ranking_{i,p} = \frac{N_{i,p}}{N_{tot}} \quad (6)$$

iii) Reproducibility evaluation

Finally, implementation of an algorithm should always lead to reproducible segmentations results, but sensitivity of the segmentation might be affected by slight changes in the initial conditions. Thus, the sensibility of the algorithm to small changes in the starting conditions requires reproducibility testing[41] that was achieved for the implementation by repeating thedelineation procedure times for each patient. The impact of user interaction on the delineation process was simulated by modifications of the unique subvolume (mask, 3D-box)manually drawn by the user to compute contours from the different algorithms and from the consensus approaches. As a metric of accuracy the mean standard deviation of the DSC (eq. (4)) or percent error between the delineated volume and the pathological ground truth (eq. (5)) were estimated from the 5 delineations of each patient data set.

iv) Statistical analysis

The non-parametric Wilcoxon test for paired samples was applied to determine if DSC, volume percent error and/or volume normalized to pathology were significantly different between MV and STAPLE consensus volumes. In order to determine if the MV rule gives significantly different volume percent error than the individual segmentation methods, a nested analysis of variance (ANOVA) on the pooled data of the 4 centres was performed. Further, Tukey's test was applied for pairwise comparisons of respective means. Statistical analysis was carried out by use of the software package XLSTAT 2011 (Addinsoft). All values are expressed as mean and SD unless otherwise indicated. The p-values are considered statistically significant if less than 0.05.

II. Results

The patient cohorts included in this evaluation comprised different tumour entities, that is, tumours that differ in size, biology and body location. The corresponding PET datasets therefore represent a collection of clinical PET images that differ in terms of image contrast

and noise levels.

Figure 1 shows both the individually segmented volumes together with the MV and STAPLE consensus volumes for both, a patient with confluent lesions in the hepatic hilum (patient no. 2, centre-3) as an example of a high-contrast irregular object surrounded by tissues of non-negligible uptake and a breast lesion of small size and low FDG-tracer uptake (patient no.8, centre-4) exemplifying the resulting contours in case of a small, faintly accumulating lesion. Forthe lymphoma all methods produce plausible segmentations of the lesion with small differences in the resulting contours. For the breast lesion, however, the individual segmentation methods provide quite different volumes that all include large volume differences to the pathological ground truth.

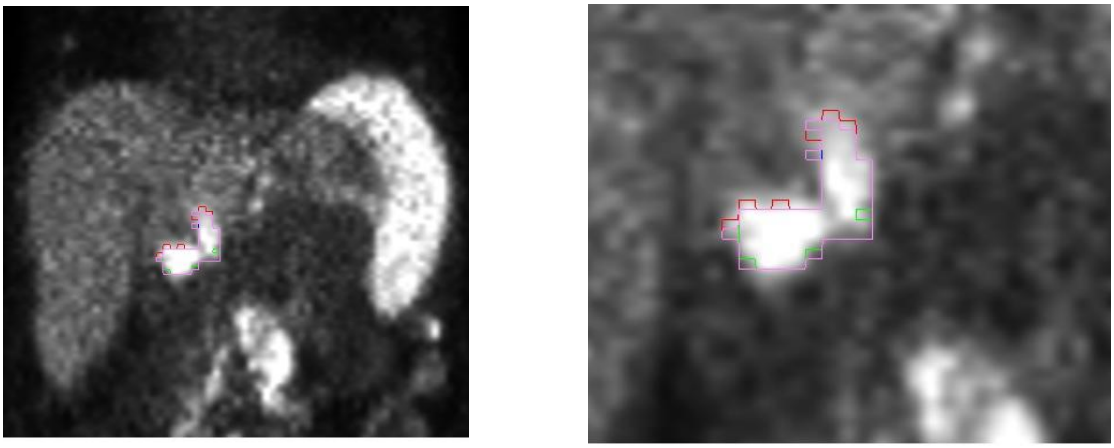


Figure 1a Coronal slice of confluent nodal lesions in the hepatic hilum with contours determined byADT (red), POS (green), COA (blue) and the MV consensus method (pink) (Lymphoma – centre-3).

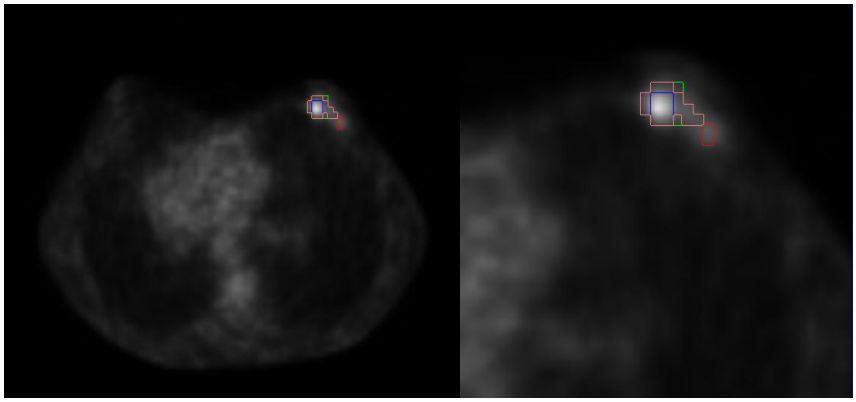


Figure 1b: Transverse PET-slice of a patient with a breast FDG positive lesion (centre-4). Contours determined by the delineation methods ADT (red), POS (green), COA (blue) and the consensus-method MV (pink) are overlaid.

a) Volume segmentation

Table 3 presents the mean delineated volumes and the mean percent- and absolute errors of both, the three individual segmentation algorithms and the two consensus methods for the

four datasets in comparison with the respective ground truth and ground truth surrogate, respectively. Very close results were found between the two consensus methods.The statistical analysis (Wilcoxon test) of all datasets confirmed that the percent errors obtained from the MV rule were not significantly different to those obtained with the STAPLE algorithm (all $p>0.201$). Therefore only MV will be considered in the following results.

For each patient of the 4 clinicaldatabases, Figure 2a-d shows the overall performance of the independent segmentation algorithms and the MV consensus method in terms of the delineated volumes normalized to the respective ground truth. Depending on tumour characteristics as lesion size and FDG-uptake, the four patient databases revealed different segmentation results applying the individual methods.

Lung tumours of centre-1 (Figure 2a) could be evaluated in comparison with pathological specimen findings. Moreover, for this evaluation the macroscopic extent of the tumours could be determined by use of a volumetric analysis method which yielded a reliable pathological ground truth [38]. Although only a few number of patients could be investigated the corresponding tumours varied with respect to parameters as FDG-uptake (range of SUVmax: 5.9 – 29.8; median SUV: 12.8), FDG-uptake heterogeneity, tumour size (mean volume 35.8±49.5ml and tumour localisation [38]. As a limitation, solely 3D-PET data could be included in this cohort. Therefore the delineation results may be influenced by respiratory motion that may vary depending on the size, location and surrounding of the individual lung tumour [38, 42, 43]. Compared with the pathological specimens slight overestimation of the volumes delineated from PET data was observed for COA for all patients of this database whereas Adaptive Threshold (ADT) and Possibility Theory (POS) underestimated the pathological volume for 1/12 and 3/12 patients, respectively.

In the second cohort of lung tumour patients (centre-2**)** comprised gated PET- and CT data that should at least diminish the impact of respiratory motion on the segmentation results.Here, the PET delineation results were compared to the majority vote (MV) of the manual CT contours of 4 experts as a ground truth surrogate. The MV consensus was chosen to reduce the well-known inter-observer variability of CT contouring [44, 45] and therefore to improve the delineation accuracy. In addition, patients with peripheral tumours are clearly delineable on CT images were included in this cohort even in casesthe tumorswere rather small (mean volume: 3.5±3.5ml, range of SUVmax: 2.9 – 28.8) compare Figure 2b).For these lesions, the PET volumes delineated by all methodswere both over- and underestimation in relation tomanualCTvotemethod. Comparing the individual methods, the volume overestimation was larger with POS and the smallest were obtained with COA.

The lymphomas of the cohort from centre 3 were bulky lesions in the abdominal area (mean volume 77.9 ± 90.5 ml) that showed high but heterogeneous accumulation of FDG (mean SUVmax 15.14 ± 6.78). In addition, these lesions were located, at least in several patients, in the neighbourhood of organs with high FDG uptake, for example the spleen and bowel. CT delineation of the lesions by an expert nuclear medicine physician was used as the ground truth surrogate. Only lesions surrounded by fat were chosen in order to provide a reasonably good surrogate as the ground truth. The delineated PET volumes were underestimated by all segmentation methods in most patients in relation to the CT delineation (Fig. 2c).

In the cohort with breast cancer from centre 4, the PET volumes were greatly overestimated by all methods in relation to the pathological volumes. These results can be explained by the fact that the lesions volumes were small (mean volume: 4.0 ± 3.29 ml) leading to partial volume effect, and thus a small SUV_{max} (metdian 4.6, range: 1.9 – 11.0). Moreover, mean percent error werehigh mainly because of the very small pathology volume of three of the ten lesions (0.76 ml, 1.30 ml and 2.07 ml) thatresulted in very large differences in the delineated volumes. For the othersevenlesions (meanpathology volume 5.14 ± 3.09 ml), the mean percent error (meanabsoluteerror) with all methodswasalsohigh but with the consensus algorithmremainedsmallerthan 67 % (2.69± 1.92 ml). As shown in Fig. 2d, the threemethods of segmentation gave differentresults, especially for theseverysmalllesions (see for example patient 8 with a pathological volume of 0.76 ml). Breast cancer waschosenbecauseconservingsurgeryis the initial step in treatment and the partial breastsurgeryprovidessamplesthatcanbeused as the gold standard. In order to avoidtumourmovement and to optimizecountingstatistics, all patients underwentprocubitusungated PET/CT acquisition centred on the breastregionimmobilizedwith a device fixing the chest.

The statistical analysis of the clinical data, comprising a nested ANOVA performed on the pooled percent errors in the tumour volumes delineated in patients from the four centres, revealed that variations mainly depended on the centre, that is, on the patient database. No significant effect of the method was found (p=0.072).

b) Segmentation agreement

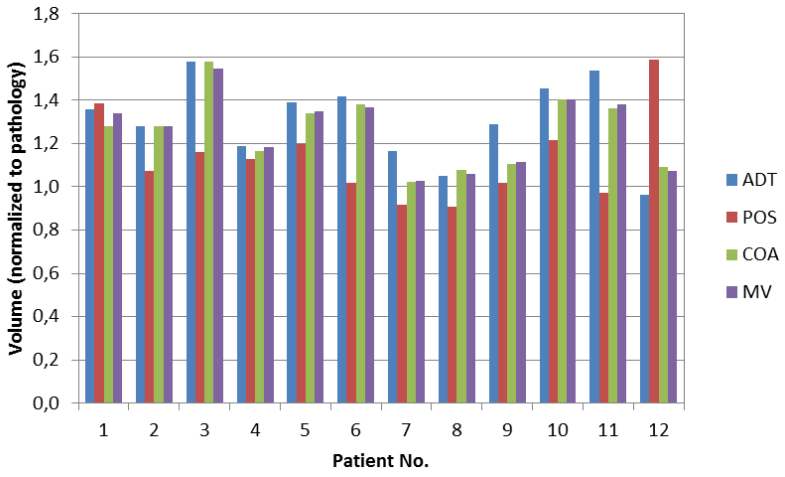
Agreement in segmentation in terms of DSC in comparison with the manual CT ground truth surrogate was determined for the databases of centre-2 (lung lesions) and -3 (lymphoma). Corresponding results are included in the last column of Table 3. Identical mean DSCswere found for the two consensus methods. The Wilcoxon testperformed on the pooled data from centre-2 and -3 confirmed that there was no significant difference in DSC between the MV rule and the STAPLE algorithm (p=0.59). Therefore only the MV rule with results presented in Figure 3 for each patient will be considered in the remainder of this paragraph.

For centre-2, the manualCTvoteof the small lung lesions was compared with the volume of the closest PET time bin. For thisdatabase, Figure 3a shows large differences in results for individual patients (minimum DSC: 0.15, maximum DSC: 0.76):in 10/14 patientsDSC of the consensus was found to be larger than 0.64(mean DSC: 0.67) indicating good segmentation quality. DSC of the individual methods was slightly smaller (mean of all methods: 0.65) with very small differences betweenthe individual methods. Small values of DSC (DSC<0.45) for both, all individual methods and the consensus were observed in 3/14 cases. These cases could be assigned to lesions located in the lower lobe of the lung exhibiting a relatively large respiratory displacement.

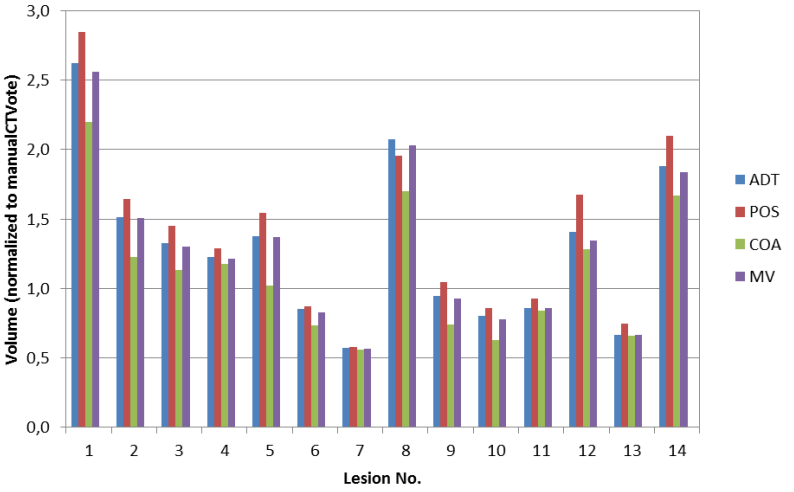
For the lymphoma lesions (centre-3)onlysmalldiffer- 561 encesamong the individualmethodswerefound(mean DSC: 0.67±0.08, minimum DSC: 0.51, maximum DSC:

0.82). The largest difference in DSC was observed for patient 4 with a DSC of 0.67 for COA while all the other methods gave DSC higher than 0.80. Here, MV either exceeds or is close to a mean DSC of 0.67 (minimum DSC:0.55, maximum DSC:0.80, compare Figure 3b) indicatinggood segmentation quality for lymphoma lesion.

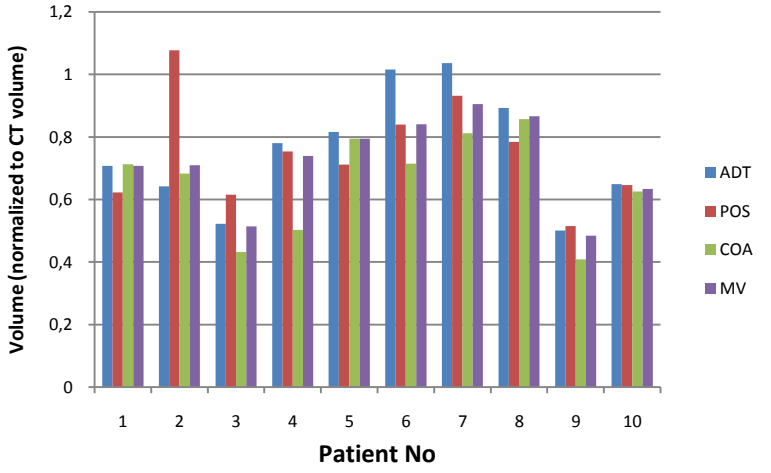
(a)



(b)



(c)



(d)

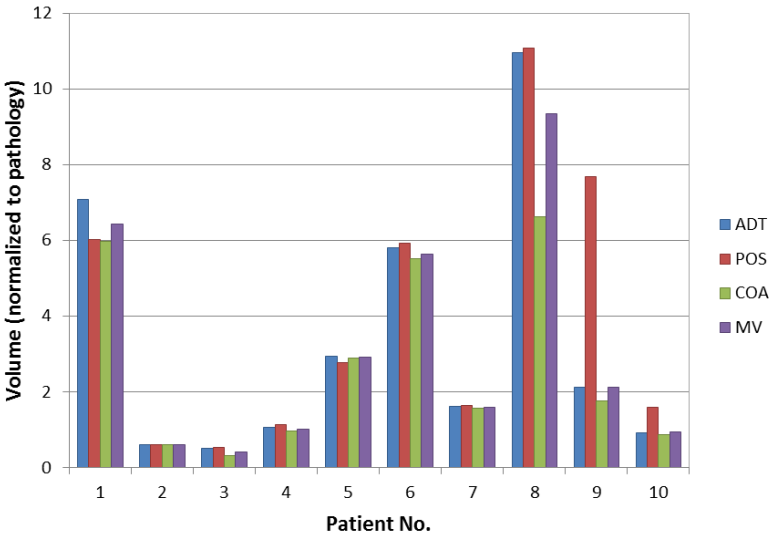
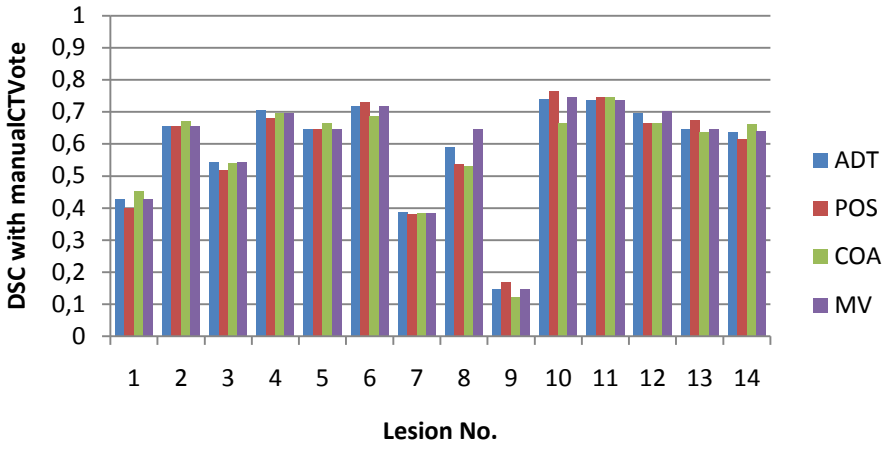


Figure 2: Volumes delineated by the three individual algorithms and the consensus-method MV and normalized to the corresponding ground truth or ground truth surrogate for each patient: (a) lung tumours of centre-1 normalized to pathology reference, (b) lung tumours of centre-2 normalized to manualCTvote, (c) lymphoma of centre-3 normalized to manual CT delineation by one expert, (d) breast cancer of centre-4 normalized to pathology reference.

(a)



(b)

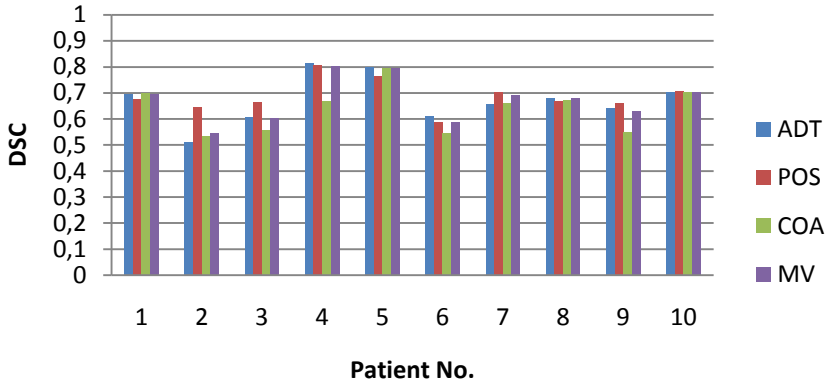


Figure 3: Mean DSC between the CT-volume and the PET-volume delineated by the three individual algorithms and the consensus-method MV for each patient. (a) Lung tumours of centre-2, (b) Lymphoma of centre-3.

c) Ranking

As described in section "data analysis" a ranking approach was applied to investigate if it is favourable in clinical routine to use a consensus-contour instead of simply using the best performing method. Compared to a reliable ground truth our results have shown that the best individual method depends on parameters like lesion size or tracer uptake resulting in PET images of different contrasts and levels of noise. In clinical routine, however, we do not generally know the reliable ground truth beforehand and therefore we do not know the best individual method before segmentation. The distribution of the resulting method ranking that was calculated according to Eq. 6 is shown in Figure 4. For the datasets of centre 1, 2 and 3 ranking demonstrates that the individual method best performing in many cases was also often the worst performing method in other cases. Compared with the individual methods the best ranking of the consensus was slightly lower but never ranked as worst method, except for one case of centre-1.. Altogether, these results demonstrate that combining different segmentation methods by application of a consensus method offers robustness against the variable performance of individual segmentation methods.

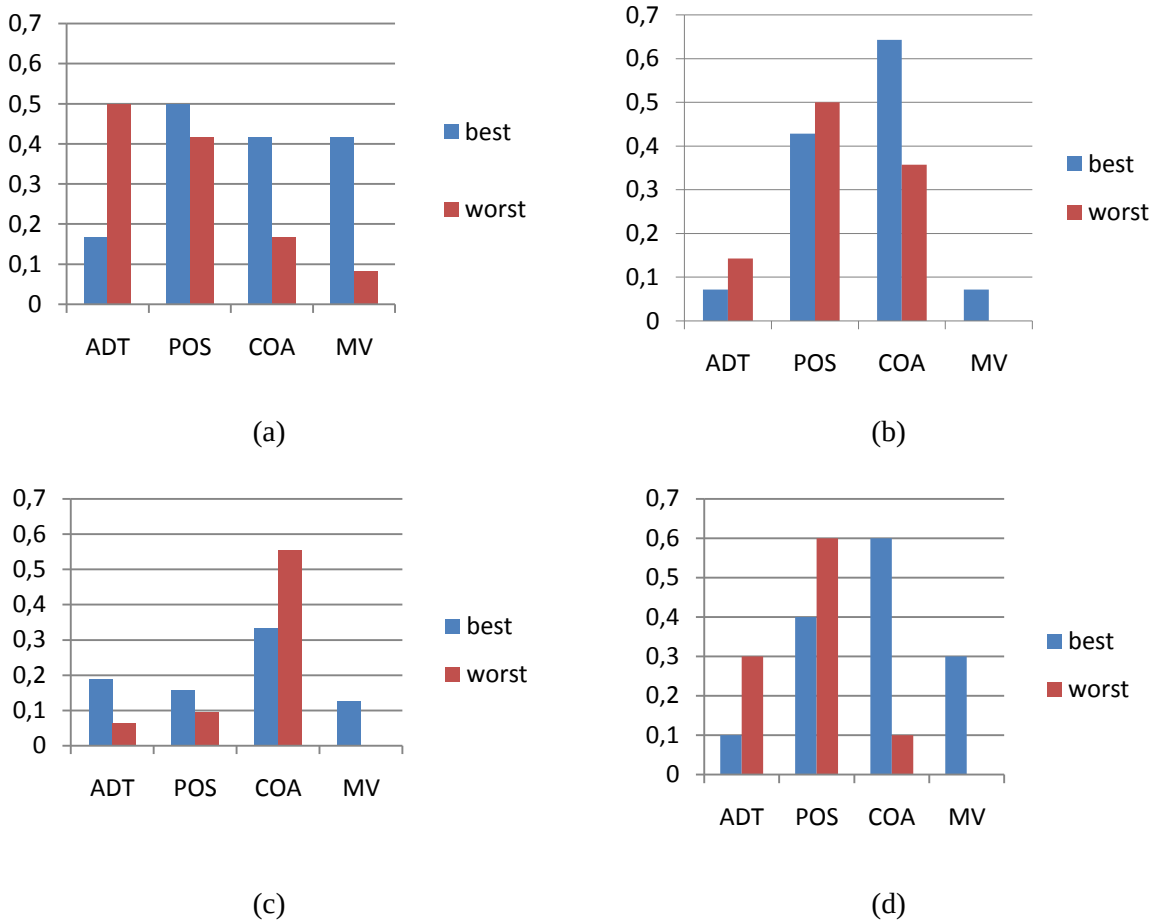


Figure 4: Rankingproportions of volume ranking of the individual segmentation algorithms and the consensus methods. (a) lung lesions (centre-1), ranking based on the volume difference to pathology (b) lung lesions (centre-2), ranking based on the volume difference to the manualCTvote; (c) lymphoma(centre-3), ranking based on the DSC;(d) breast lesions (centre-4), ranking based on the volume difference to pathology

Our results (Figures 2-4) reflect that the performance of individual segmentation algorithms can be variable for lesions of different tumour entities, that is, for PET images that differ in resolution, contrast and image noise.However, (Figures 2-4) also demonstrate that the consensus method displays improved volume segmentation accuracy compared to the worst performing individual method in all cases and is close to the best performing method in many cases. This observation was supported by the results of the statistical analysis: Application of a nested ANOVA on the pooled databases of the 4 centres followed by a further Tukey's tests revealed that the percent errors obtained with the worst performing method are significantly higher than those obtained with MV ($p=0.031$). In addition the Wilcoxon test revealed that DSC values of the worst performing method are significantly lower than those obtained with MV (centre-2: $p=0.0019$; centre-3: $p=0.0029$) confirming the robustness of the consensus-method.

d) Reproducibility

For each method the mean of the standard deviation of the percent volume error or DSC was

calculated over all patients and summarized in Table 4. The values for centre-1, centre-3 and centre-4 demonstrate that for all methods used a modification of the volume of work placed at the beginning of the segmentation process has only a minor effect on the delineation results. However, for breast lesions (centre-4), corresponding to small, faintly accumulating lesions, a change of this volume may have an important impact on the segmentation output (compare Table 4).

III Discussion

In this paper, clinical PET/CT data comprising different tumour entities were used to assess the performance of a novel PET segmentation concept that combines three individual PET segmentation methods by application of two consensus algorithms. Consensus and individual segmentation algorithms were implemented on the same software platform to allow an optimal workflow and minimize reproducibility drawbacks. Three segmentation methods developed by the authors were used as entry of the consensus algorithms: possibility theory based approach (POS)[20], contrast oriented approach (COA) [13] and adaptive threshold oriented approach (ADT)[32]. To our knowledge we provide the first multicentre clinical evaluation of combining several PET-segmentation methods by a consensus approach on different tumour entities. Our study is in line with previous work from McGurk et al [46] who evaluated application of the consensus method on phantom measurements. In addition, this group proposed to use consensus volumes to reduce the intra- and inter-observer variability of manual delineation for head-and-neck cancer patients and applied the consensus methods to assess the treatment response in radiation therapy.

Our clinical databases comprised lung tumours, breast tumours and lymphoma, that is, tumours that differ in biology, size and body location. The corresponding PET datasets therefore represent a collection of clinical PET images that differ in terms of image contrast and noise levels. In addition the images were acquired in different centres on different PET/CT systems.Overall, the present analysis was used to evaluate the feasibility and usefulness of the consensus approach to improve robustness of PET based contouring of tumour volumes.

Our results demonstrate that the consensus methods display improved volume segmentation accuracy compared to the worst performing individual methodsin the majority of cases and are close to the best performing methods for many cases of the tumour entities involved. Differences in volume percent error varied for the different tumour entities demonstrating the impact of tumour characteristics translating into PET-image characteristics on the accuracy of the individual segmentation methods. Differences in DSC were small for both, lungs and lymphoma and statistically not significant between the consensus and the individual methods. However, compared with the worst performing method both, statistically significant lower values of percent error and higher values of DSC were obtained applying the consensus algorithm. Keeping in mind that the ground truth is generally not known, these findings demonstrate higher robustness and accuracy of the consensus contour compared to application of one individual segmentation method.

Our results on ranking confirm this higher robustness of the consensus method. Even if one of the individual methods was performing best in many cases (e.g. COA for lymphoma, lungs and breast or POS for lung 1) the same method was also the worst method in other cases. Consequently, the ranking of an individual segmentation method was observed to change depending on the tumour entity, and also on a comparison of individual patients within one database. This can be explained by the variable performance of individual PET-segmentation methods which is known to depend on varying clinical conditions involving different lesion size, noise levels or radiopharmaceutical uptake heterogeneity [47, 48]. Those uncertainties of image segmentation are also known in other fields of medical imaging, especially in MRI. To overcome, combining individual algorithms by use of a consensus method has been proposed in the literature and applied in different fields [26-28, 31, 36]. In our study on PET segmentation, the consensus method showed essentially equivalent performance compared to using the best performing individual segmentation method in many cases or, respectively, improved segmentation accuracy compared to the worst performing individual method in the majority of cases. This finding demonstrates evidently that combining multiple segmentation methods provides robustness of segmentation accuracy in comparison to using one single individual method. This is very important keeping in mind that the ground truth is generally not known. Moreover, to a certain extent, the consensus methods seem to compensate the weaknesses of the individual methods. Therefore the use of the consensus method may potentially provide a more robust approach to RT planning applications.

The relatively small differences between the individual methods and the consensus algorithm might be explained by the choice of the individual methods, two of them are adaptive thresholding algorithms. As a limitation only three algorithms could be included in this evaluation (software implementation). In the recent publication by McGurk et al, the authors stated that according to [49]the accuracy in a majority vote approach is guaranteed to improve depending on the number of methods used if the individual methods have accuracies greater than 0.5. To exemplify, according to [49]combining 3 individual segmentation methods all having accuracies of 0.6 improves the consensus accuracy to 0.6480 (8%), combining 5 of them to 0.6826 (13.2%). In addition, higher accuracies of the individual methods cause higher levels of improvement. However, using three algorithms, situations might arise where the results of two methods are totally matching but are less accurate than that of the third algorithm. In these cases, the majority vote approach will not improve segmentation accuracy. Further developments of the current software should therefore involve the implementation of at least one or two additional other state-of-the-art PET-segmentation algorithms of high accuracy implying e.g. edge detection, stochastic models, and other approaches [14-20, 22, 23].

In our study we applied two consensus methods: the majority vote rule leading to decide if voxels belong to the lesion or not according to the results of the majority of the segmentation methods (MV) [28] and a probabilistic method, the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [29, 30], leading to compute a probabilistic estimate of the ground truth from a collection of segmentation results. The group of McGurk et al has investigated these two consensus methods on PET phantom measurements combining a

collection of 5 individual segmentation methods [31]. In concordance with our results on clinical data, these authors could demonstrate on phantom data that differences between MV and STAPLE were small and that both methods offer good performance when combining volumes[37]. The small differences between the two consensus methods should be explained considering the number and the respective segmentation accuracy of the individual algorithms which are used in entry of STAPLE. The number of three different algorithms included in our evaluation as well as the similarity of two of the methods might reduce the impact on the statistical estimate and therefore on the output of STAPLE. Again, further developments including several other state-of-the-art PET-segmentation algorithms of high accuracy will be necessary to investigate the impact on STAPLE. However, this was beyond the object of the current investigation.

It needs to be stated that the reproducibility of the segmentation was good for all tumour entities except breast carcinoma.The current software implementation that allowed a simultaneous application of the different segmentation methods while keeping user interaction to a minimum was surely a key point that facilitates this good reproducibility.The relatively low values for breast carcinoma (compare Table 4), however, may reflectthe impact of neighbouring DCIS components [50] and partial volume effects due to small lesion size which are typical for this tumour entity.

In the present evaluation study, only intra-user reproducibility was evaluated because of the huge amount of data to be analysed. According to our results on intra-user reproducibility we suggest that inter-user reproducibility of PET volume segmentation should also be improved when using the consensus approach instead of one individual segmentation method. This will be of high interest in those multi-centric clinical trials with targeting based on PET-CT delineation and might guaranty a more reliable, homogenous delineation approach over all the centres.

Collecting databases of different tumour entities acquired on different PET/CT system remains a challenging task and only multicentre trials can offer this variability. Nevertheless, we are aware that the ground truth and ground truth surrogates, respectively, used here (volume pathology, manual CT-delineation) were not optimal in all cases. As an alternative Monte Carlo simulation (MC) of clinical PET/CT scans could be included. Considering the current state of the art of MC offering realistic images such as databases described in Hatt et al [51]or in Papadimitrioulas et al. [52] where authors introduced heterogeneity models, demonstrates that simulating datasets with features close to real life imaging has become feasible.Nevertheless, the use of databases generated by MC was beyond the object of the current investigation.

Finally, our results on clinical data, based on different ground truth surrogates have demonstrated that combining several segmentation algorithms by a consensus method improved the segmentation accuracy in the majority of cases and, importantly, showed good robustness when comparing against the worst performing individual method for each site. Thus, this concept can be applied in clinical routine to combine different segmentation methodsor manual delineation results of several experts. This makes the use of consensus methods relevant for radiation therapy considering PET-based GTV-delineation but also for

other scenarios like the joining of expert recommendations in clinical routine and trials or the generation of multi-observer generated contours for standardisation of automatic contouring.

IV Conclusion

In this study, we determined the added value of combining PET-segmentation results with consensus methods considering different clinical scenarios, technical details and ground truths (or surrogates). Four different clinical databases comprising different tumour entities (lung, breast, lymphoma) and two consensus algorithms (MV, STAPLE) were included in this investigation. In terms of accuracy and reproducibility both consensus methods offered similar results, that is (i) consensus greatly improved volume segmentation compared to the worst performing individual method and (ii) the consensus delineation results were close to that of the best performing individual method in nearly all cases. These results were independent on tumour location (lung, breast) or pathology (lymphoma). Thus, this study demonstrates that consensus algorithms can be very useful for combining automatic segmentation results in medical imaging but also for other scenarios like the joining of expert recommendations in clinical routine and trials or the generation of multi-observer generated contours for standardisation of automatic contouring.

TABLES

Table 1: Clinical protocols usedfor routine whole-body ^{18}F-FDG PET or –PET/CT in the different centres

Centre	centre-1	centre-2	centre-3	centre-4
scanner	[1]Siemens ART	[2]Philips GEMINI TF64	[3]GE Discovery RX	[1]Siemens Biograph LSO sensation16
Fasting time (h)	6	6	6	6
Mean glucose level (mg/dl)	<150	95	129	<150
Acquisition time point (min p.i.)	90 ± 8	120	86 ± 27	69 ±11
Mean activity (MBq)	279 ± 33	263 ± 32	316 ± 53	343 ± 70

[1]CTI/ Siemens Medical Solutions, Hoffman Estates, Knoxville, TN, USA

[2]Medical Philips System, Eindhoven, Netherlands

[3]GEHC - Milwaukee, Wisconsin, USA

Table2: PET-scanner settingsas used in the different centres for acquisition of patient data..

	centre-1	**centre-2**	**centre-3**	**centre-4**
Scanner	Siemens ART	Philips GEMINI TF64	GE Discovery	Siemens Biograph 16
Matrix size PET (voxels)	128x128x92	144x144x45	128x128x47	168x168x80
Voxel size PET	5.15x5.15x3.375 mm^3	4.00x4.00x4.00 mm^3	5.46x5.46x3.27 mm^3	4.06x4.06x2.0 mm^3
Axial Field of View (FOV)	162 mm	180 mm	153 mm	152 mm
Emission scan time per bed position	10 min	15 min	2 min	3 min
Reconstruction algorithm	OSEM	BLOB-OS-TF	OSEM	AWOSEM
Algorithm settings	2 iterations, 4 subsets, 2 mm Gaussian filter	2 iterations, 21 subsets, 5 mm Gaussian filter	2 iterations, 21 subsets, 5 mm Gaussian filter	4 iterations, 8 subsets, 5 mm Gaussian filter
Attenuation correction	Transmission in Singles mode (137-Cs)	4D CT correction	CT correction	CT correction
Time bins		10		

Table 3: Mean percent-errors, absolute errors and mean DSC (if available) of PET-based delineation using the three individual segmentation algorithms and the two consensus methods in comparison with the given ground truth. Absolute values of the mean delineated volumes are also given (for comparison:; mean pathological volume lung lesions (centre-1): 35.8±49.5ml; mean CT volume lung lesions (centre-2): 3.52 ±3.54 ml mean CT volume lymphoma: 77.9 ±90.5 ml;mean pathological volume breast cancer: 4.0±3.3ml.

Method	Mean Delineated Volume (ml)	Mean percent error (%)	Mean absolute error (ml)	Mean DSC
Lung tumours (centre-1)				
ADT	44.8±61.9	30.6±18.0	8.93±13.54	
POS	36.4±48.9	13.1±19.2	0.61±3.77	
COA	41.1±53.7	25.4±16.0	5.31±5.43	
MV	41.1±53.9	26.3±16.3	5.33±5.63	
STAPLE	41.1±53.9	26.2±16.5	5.33±5.64	
Lung tumours (centre-2)				
ADT	3.44±1.73	29.45±56.43	-0.23±1.93	0.59±0.16
POS	3.69±1.77	39.50±60.32	0.02±1.92	0.58±0.16
COA	3.12±1.85	11.25±46.23	-0.55±1.85	0.58±0.16
MV	3.39±1.73	27.00±54.96	-0.28±1.92	0.59±0.16
STAPLE	3.39±1.73	27.00±54.96	-0.28±1.92	0.59±0.16
Lymphoma (centre-3)				
ADT	60.2±71.98	26.66±16.12	19.03±20.02	0.67±0.09
POS	58.48±66.78	26.57±13.69	20.08±21.55	0.69±0.06
COA	45.78±45.08	25.59±14.08	20.13±22.73	0.64±0.09
MV	56.21±65.98	27.85±14.14	21.69±23.09	0.67±0.08
STAPLE	56.23±65.98	28.08±14.05	21.67±23.09	0.67±0.08
Breast cancer (centre-4)				
ADT	6.8±2.7	255.5±333	4.4±3.3	
POS	8.5±4.6	307.4±343	5.9±4.8	
COA	6.0±2.7	196±221	3.9±2.9	
MV	6.5±2.6	231.2±285	4.2±3.1	
STAPLE	6.5±2.6	231.2±285	4.2±3.0	

Table 4: Mean of standard deviation per method for lung tumours, lymphoma and breast cancer.

Method	%Variation of standard deviation
Lung tumours (centre-1 / percent error)	
ADT	1.66±2.24
POS	5.11±8.75
COA	2.08±2.53
MV	2.10±3.13
STAPLE	2.28±3.53
Lymphoma (centre-3 / DSC)	
ADT	1.68±1.36
POS	1.25±1.09
COA	2.67±4.49
MV	1.31±1.18
STAPLE	1.44±1.23
Breast cancer (centre-4 / percent error)	
ADT	84.14±160.80
POS	35.36±67.19
COA	128.62±163.44
MV	47.55±66.93
STAPLE	47.72±66.51

References:

1. Miller TR, Grigsby PW. Measurement of tumor volume by PET to evaluate prognosis in patients with advanced cervical cancer treated by radiation therapy. International Journal of Radiation Oncology Biology Physics; 2002. p. 353-9.
2. Fernando S, Kong F, Kessler M, Chetty I, Narayan S, Tatro D, et al. Using FDG-PET to Delineate Gross Tumor and Internal Target Volumes. International Journal of Radiation Oncology Biology Physics; 2005. p. S400-S1.
3. Zasadny KR, Kison PV, Francis IR, Wahl RL. FDG-PET Determination of metabolically Active Tumor Volume and Comparison with CT. Clinical positron imaging; 1998. p. 123-9.
4. Bryant AS, Cerfolio RJ. The Maximum Standardized Uptake Values on Integrated FDG-PET/CT Is Useful in Differentiating Benign From Malignant Pulmonary Nodules. The Annals of Thoracic Surgery; 2006. p. 1016-20.
5. Black QC, Grills IS, Kestin LL, Wong CYO, Wong JW, Martinez AA, et al. Defining a radiotherapy target with positron emission tomography. International Journal of Radiation Oncology Biology Physics; 2004. p. 1272-82.
6. Daisne JF, Duprez T, Weynand B, Lonneux M, Hamoir M, Reychler H, et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: Comparison between CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology; 2004. p. 93-100.
7. Daisne JF, Sibomana M, Bol A, Doumont T, Lonneux M, Gr,goire V. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of the reconstruction algorithms. Radiotherapy & Oncology; 2003. p. 247-50.
8. Erdi YE, Mawlawi O, Larson SM, Imbriaco M, Yeung H, Finn R, et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. Cancer; 1997. p. 2505-9.
9. Erdi YE, Wessels BW, Loew MH, Erdi AK. Threshold estimation in single photon emission computed tomography and planar imaging for clinical radioimmunotherapy Cancer Research; 1995. p. S5823-S6.
10. Jentzen W, Freudenberg L, Eising EG, Heinze M, Brandau W, Bockisch A. Segmentation of PET volumes by Iterative Image Thresholding. Journal of Nuclear Medicine; 2007. p. 108-14.
11. Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welsch C, Hellwig D, Rube C, et al. Comparison of Different Methods for Delineation of PET-Positive Tissue for Target Volume Definition in Radiotherapy of Patients with Non-Small Cell Lung Cancer. Journal of Nuclear Medicine; 2005. p. 1342-8.
12. Nestle U, Schaefer-Schuler A, Kremp S, Groeschel A, Hellwig D, Rube C, et al. Target volume definition for 18 F-FDG PET-positive lymph nodes in radiotherapy of patients with non-small cell lung cancer. European Journal of Nuclear Medicine and Molecular Imaging; 2007. p. 453-62.
13. Schaefer A, Kremp S, Hellwig D, Rube C, Kirsch CM, Nestle U. A contrast-oriented algorithm for FDG-PET-based delineation of tumour volumes for the radiotherapy of lung cancer: derivation from phantom measurements and validation in patient data. European Journal of Nuclear Medicine and Molecular Imaging; 2008. p. 1989-99.

14. Riddell C, Brigger P, Carson RE, Bacharach SL. The watershed algorithm: a method to segment noisy PET transmission images. Nuclear Science, IEEE Transactions on. 1999;46:713-9.
15. Tylski P, Bonniaud G, Decenciere E, Stawiaski J, Coulot J, Lefkopoulos D, et al. 18 F-FDG PET images segmentation using morphological watershed: a phantom study. IEEE Nuclear Science Symposium Conference Record. San Diego, USA; 2006. p. 2063-7.
16. Nissen I, Yaqub M, Lammertsma A, Lee J, Geets X, Boellaard R. A novel supervised watershed method for segmentation of tumors with heterogeneous tracer uptake in PET. Society of Nuclear Medicine Annual Meeting Abstracts: Soc Nuclear Med; 2014. p. 260.
17. Geets X, Lee JA, Bol A, Lonneux M, Gr,goire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. European Journal of Nuclear Medicine and Molecular Imaging; 2007. p. 1427-38.
18. Zhu W, Jiang T. Automation segmentation of PET image for brain tumors. Nuclear Science Symposium Conference Record, 2003 IEEE: IEEE; 2003. p. 2627-9.
19. Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. Med Phys. 2010;37:1309-24.
20. Dewalle-Vignion AS, Betrouni N, Lopes R, Huglo D, Stute S, Vermandel M. A New Method for Volume Segmentation of PET Images, Based on Possibility Theory. IEEE Trans Med Imaging. 2011;30:409-23. doi:10.1109/TMI.2010.2083681.
21. Dewalle-Vignion AS, Betrouni N, Makni N, Huglo D, Rousseau J, Vermandel M. A new method based on both fuzzy set and possibility theories for tumor volume segmentation on PET images. Proceedings of the 30th IEEE EMBS Annual International Conference. Vancouver, Canada; 2008. p. 3122-5.
22. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imaging. 2009;28:881-93. doi:10.1109/TMI.2008.2012036.
23. Hatt M, Lamare F, Boussion N, Turzo A, Collet C, Salzenstein F, et al. Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET. Phys Med Biol. 2007;52:3467-91. doi:10.1088/0031-9155/52/12/010.
24. Prieto E, Lecumberri P, Pagola M, Gomez M, Bilbao I, Ecay M, et al. Twelve automated thresholding methods for segmentation of PET images: a phantom study. Phys Med Biol. 2012;57:3963-80. doi:10.1088/0031-9155/57/12/3963.
25. Shepherd T, Teras M, Beichel R, Boellaard R, Bruynooghe M, Dicken V, et al. Comparative Study with New Accuracy Metrics for Target Volume Contouring in PET Image Guided Radiation Therapy. IEEE Trans Med Imaging. 2012. doi:10.1109/TMI.2012.2202322.
26. Østergaard LR, Larsen OV. Applying voting to segmentation of MR images. Advances in Pattern Recognition: Springer; 1998. p. 795-804.
27. Artaechevarria X, Munoz-Barrutia A, Ortiz-de-Solorzano C. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans Med Imaging. 2009;28:1266-77. doi:10.1109/TMI.2009.2014372.
28. Lam L, Suen CY. Application of majority voting to pattern recognition: an analysis of its behavior and performance. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on. 1997;27:553-68.
29. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004;23:903-21. doi:10.1109/TMI.2004.828354.

30. Commowick O, Akhondi-Asl A, Warfield SK. Estimating A Reference Standard Segmentation With Spatially Varying Performance Parameters: Local MAP STAPLE. IEEE Trans Med Imaging. 2012;31:1593-606. doi:10.1109/TMI.2012.2197406.
31. McGurk RJ, Bowsher J, Lee JA, Das SK. Combining multiple FDG-PET radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods. Med Phys. 2013;40:042501. doi:10.1118/1.4793721.
32. Vauclin S, et al. Development of a generic thresholding algorithm for the delineation of 18 FDG-PET-positive tissue: application to the comparison of three thresholding models. Physics in Medicine and Biology. 2009;54:6901.
33. Doll C, Parcq C, Modzelewski R, Dewalle-Vignion AS, Christ U, Loquin K, et al. PET-based Target Volume Delineation in Radiation Therapy Planning: Are Different Implementations of the Same Automatic Delineation Method Really Equal? . Annual congress - Europena Association of Nuclear Medicine - EANM. Lyon (France); 2013.
34. Schaefer A, Nestle U, Kremp S, Hellwig D, Grgic A, Buchholz HG, et al. Multi-centre calibration of an adaptive thresholding method for PET-based delineation of tumour volumes in radiotherapy planning of lung cancer. Nuklearmedizin Nuclear medicine. 2012;51:101-10. doi:10.3413/Nukmed-0452-11-12.
35. Zadeh LA. Fuzzy Sets as the Basis for a Theory of Possibility. Fuzzy Sets ans Systems; 1978. p. 3-28.
36. Kittler J, Alkoot FM. Sum versus vote fusion in multiple classifier systems. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2003;25:110-5.
37. Dewalle-Vignion AS, Betrouni N, Baillet C, Vermandel M. Is STAPLE algorithm confident to assess segmentation methods in PET imaging? Physics in Medicine and Biology. 2015;In press (accepted august 26th, 2015).
38. Schaefer A, Kim YJ, Kremp S, Mai S, Fleckenstein J, Bohnenberger H, et al. PET-based delineation of tumour volumes in lung cancer: comparison with pathological findings. Eur J Nucl Med Mol Imaging. 2013;40:1233-44. doi:10.1007/s00259-013-2407-x.
39. Hapdey S, Edet-Sanson A, Gouel P, Martin B, Modzelewski R, Baron M, et al. Delineation of small mobile tumours with FDG-PET/CT in comparison to pathology in breast cancer patients. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2014. doi:10.1016/j.radonc.2014.08.005.
40. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol. 2004;11:178-89.
41. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2008;31:466-75. doi:10.1002/uog.5256.
42. Aristophanous M, Berbeco RI, Killoran JH, Yap JT, Sher DJ, Allen AM, et al. Clinical utility of 4D FDG-PET/CT scans in radiation treatment planning. Int J Radiat Oncol Biol Phys. 2012;82:e99-105. doi:10.1016/j.ijrobp.2010.12.060.
43. Hatt M, Cheze-le Rest C, van Baardwijk A, Lambin P, Pradier O, Visvikis D. Impact of tumor size and tracer uptake heterogeneity in (18)F-FDG PET and CT non-small cell lung cancer tumor delineation. J Nucl Med. 2011;52:1690-7. doi:10.2967/jnumed.111.092767.
44. Steenbakkers RJ, Duppen JC, Fitton I, Deurloo KE, Zijp L, Uitterhoeve AL, et al. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a 'Big Brother' evaluation. Radiotherapy and oncology :

journal of the European Society for Therapeutic Radiology and Oncology. 2005;77:182-90. doi:10.1016/j.radonc.2005.09.017.
45. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2003;21:2574-82. doi:10.1200/JCO.2003.01.144.
46. McGurk RJ. Consensus Segmentation for Positron Emission Tomography: Development and Applications in Radiation Therapy: Duke University; 2013.
47. Shepherd T, Teras M, Beichel RR, Boellaard R, Bruynooghe M, Dicken V, et al. Comparative study with new accuracy metrics for target volume contouring in PET image guided radiation therapy. IEEE Trans Med Imaging. 2012;31:2006-24. doi:10.1109/TMI.2012.2202322.
48. Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging. 2010;37:2165-87. doi:10.1007/s00259-010-1423-3.
49. Kuncheva LI. Combining pattern classifiers: methods and algorithms: John Wiley & Sons; 2004.
50. Azuma A, Tozaki M, Ito K, Fukuma E, Tanaka T, O'Uchi T. Ductal carcinoma in situ: correlation between FDG-PET/CT and histopathology. Radiation medicine. 2008;26:488-93. doi:10.1007/s11604-008-0263-6.
51. Hatt M, Cheze le Rest C, Descourt P, Dekker A, De Ruysscher D, Oellers M, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys. 2010;77:301-8. doi:S0360-3016(09)02954-X [pii] 10.1016/j.ijrobp.2009.08.018.
52. Papadimitroulas P, Loudos G, Le Maitre A, Hatt M, Tixier F, Efthimiou N, et al. Investigation of realistic PET simulations incorporating tumor patient's specificity using anthropomorphic models: creation of an oncology database. Med Phys. 2013;40:112506. doi:10.1118/1.4826162.