Is STAPLE Algorithm Confident to assess Segmentation Methods in PET imaging?

Anne-Sophie Dewalle-Vignion, Nacim Betrouni, Clio Baillet, Maximilien Vermandel

Is STAPLE Algorithm Confident to assess Segmentation Methods in PET Imaging?

Anne-Sophie Dewalle-Vignion[1], Nacim Betrouni[1],
Clio Baillet[2], Maximilien Vermandel[1, 2]

[1] Univ. Lille, Inserm, CHU Lille, U1189 - ONCO-THAI - Image Assisted Laser Therapy for Oncology, F-59000 Lille, France
[2]CHU Lille, Nuclear Medicine Dept, F-59000 Lille, France

Corresponding author

Maximilien Vermandel
INSERM U1189 ONCO-THAI
1 avenue Oscar Lambret
CHRU de Lille
59037 Lille Cedex
Tel: 33. 3. 20.44.67.21
email: m-vermandel@chru-lille.fr

Abstract

Purpose: Accurate tumor segmentation in [18F]-fluorodeoxyglucose positron emission tomography is crucial for tumor response assessment and target volume definition in radiation therapy. Evaluation of segmentation methods from clinical data without ground truth is usually based on physicians' manual delineations. In this context, the Simultaneous Truth and Performance Level Estimation algorithm could be useful to manage the multi-observers variability. In this paper, we evaluated how this algorithm could accurately estimate the ground truth in PET imaging.

Methods: Complete evaluation study using different criteria was performed on simulated data. The STAPLE algorithm was applied to manual and automatic segmentation results. A specific configuration of the implementation provided by the Computational Radiology Laboratory was used.

Results: Consensus obtained by the STAPLE algorithm from manual delineations appeared to be more accurate than manual delineations themselves (80% of overlap). An improvement of the accuracy was also observed when applying the STAPLE algorithm to automatic segmentations results.

Conclusions: The STAPLE algorithm, with the configuration used in this paper, is more appropriate than manual delineations alone or automatic segmentations results alone to estimate the ground truth in PET imaging. Therefore, it might be preferred to assess the accuracy of tumor segmentation methods in PET imaging.

Keywords

PET imaging; Tumor segmentation; STAPLE algorithm; Ground truth; Segmentation methods; Manual delineations;

I. Introduction

The role of [18F]-Fluorodeoxyglucose Positron Emission Tomography ([18F]-FDG PET) in oncology is now well established for initial staging, therapy response assessment in lymphoma (1) and solid tumors (2, 3), and for radiation treatment planning (4, 5).

FDG-PET suffers from a limited spatial resolution. A difficult issue is to accurately determine metabolically active tumor volume in order to provide a better target volume definition in radiation therapy planning and a more reliable assessment of the outcomes. Modern radiation therapy techniques such as intensity-modulated radiation therapy (IMRT) or stereotactic radiation therapy allow increasingly smaller treatment margins (ranging from 1 to 5 mm depending on the localization) and therefore require more accurate target volume definition.

Determination of the metabolic volume in PET imaging remains challenging. The limited spatial resolution and image contrast in PET lead to gradual and irregular transition between healthy and tumor tissues leading to uncertainties in tumor borders location (6). Accurate manual delineation is therefore difficult and poorly reproducible. This issue may be partially solved with semi or fully automatic segmentation methods proposed in the literature (7-12).

However, there is no framework to fully assess and compare the performances of these methods. Phantom studies allow an accurate comparison between volumes segmented on PET images and actual volumes. Results of these comparisons do not really reflect the performance observed on clinical data because of more complex images (heterogeneous tumors, complex shapes) (13). Alternatively, anatomopathological studies enable volumes segmented on PET images to be compared with volumes reconstructed from macroscopic surgical specimen after surgery. Such comparisons are tedious due to technical constraints related to both the preparation of histological specimens and to the type of tumors. Furthermore they are limited to patients undergoing surgery (14-17). For non-surgical patients, as no ground truth is available, the volumes segmented on PET images can be compared with the volumes segmented on morphological images. It relies on a strong hypothesis that anatomical and metabolic tumor volumes are identical, which is not always true (18). Another approach is to compare the segmented volume with manually delineated volume. In such cases, the Simultaneous Truth and Performance Level Estimation (STAPLE)

algorithm (19) might be very useful. This algorithm computes a probabilistic estimate of the ground truth from a collection of segmentation results. So far, the STAPLE algorithm has been used in the literature in various application domains such as zonal segmentation of prostate using multispectral magnetic resonance imaging (MRI) (20), vessel segmentation in contrast enhanced CT (21), vessel segmentation from time-of-flight magnetic resonance angiography (MRA) (21), ventilation-based segmentation of the lungs using 3He MRI (22), lymph node segmentation in CT images (23), hippocampal volume measurement using MRI (24), uterine cervix segmentation from digital cervicographic images (cervigrams) (25) to estimate unavailable ground truth.

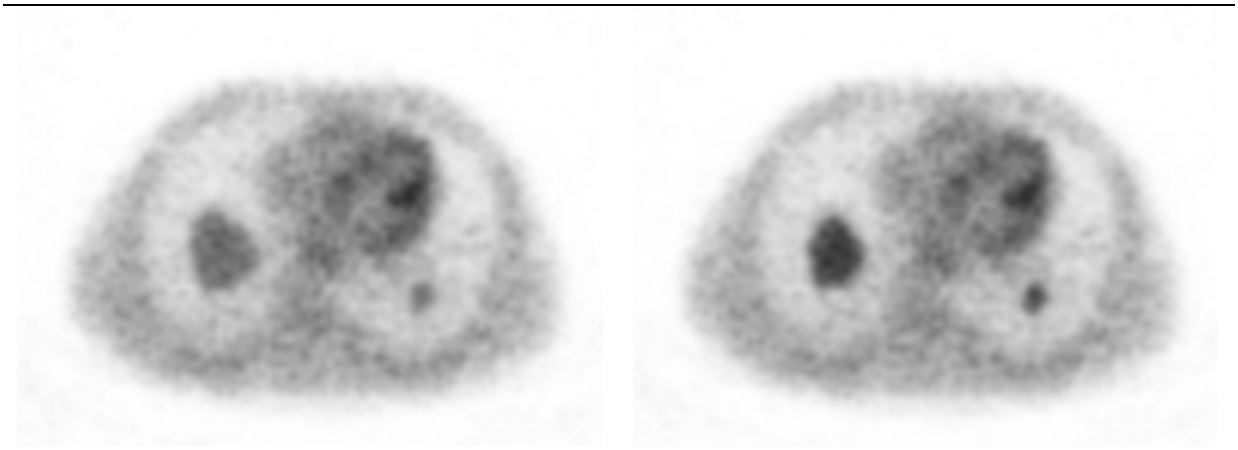
This study aimed to evaluate the ability of the STAPLE algorithm to assess segmentation methods in PET imaging. Earlier study introduced by McGurk et al. (26) investigated majority vote rule and STAPLE algorithm to combine five segmentation approaches. They aimed at reducing the impact of inconsistent performance of the individual methods in delineating regions of interest on PET images. Phantom study was achieved using the National Electrical Manufacturers Association phantom. Promising results were described since both majority vote rule and STAPLE algorithm were found to improve delineation performance. In this paper, we aimed at validating the concept of using STAPLE algorithm from manually or automatically delineated contours, either to add both accuracy and robustness to the segmentation task or to define a relevant gold standard from manual delineations when evaluating segmentation methods. Obviously, images with known ground truth were expected for STAPLE algorithm assessment which, as above-mentioned, is not easily achievable from clinical images unless under very tedious conditions (9). Simulated images (13, 27-29) which are designed as close as possible to real life imaging and for which the ground truth was actually known appeared to be the most consistent surrogate.

The rest of the manuscript is organized as follows: STAPLE algorithm mechanisms are recalled in section III. In section IV, optimization of STAPLE algorithm tuning values is achieved and optimized configuration is applied to manual segmentation results and to five automatic segmentation results. STAPLE consensuses are compared to actual ground truth in Section V. Global performances of the algorithm are discussed and presented in section VI.

II. Material

The simulated data set was obtained from patient data using the GATE simulation toolkit (27, 29, 30). Tumor contours were delineated from CT images and introduced into lungs for simulation. Datasets from 5 patients with 16 tumor volumes (Table 1) were considered (average volume = 9.25 ml; standard deviation = 14.69 ml; maximum = 56.90 ml; minimum = 1.09 ml) (Fig. 1).

For each patient, data was simulated using similar background activity and 2 different Standardized Uptake Values (SUV) of 2.5 and 4.5 in the tumor. Tumors were supposed to have a uniform activity distribution. Acquisitions from Philips GEMINI GXL PET scanner model were simulated. Random coincidences and attenuation were corrected using a delayed window and an attenuation coefficient sinogram, respectively. The simulated PET images (87 slices, matrix 144×144, voxel size of 4×4×4 mm3) were obtained using an OSEM fully 3-D algorithm, with 5 iterations and 10 subsets.



SUV= 2.5 SUV= 4.5

Fig. 1 Example of simulated images obtained from simulation with SUV= 2.5 and SUV= 4.5 (tumor volumes: 56.9 and 1.86 ml)

Table 1: List of the tumor volumes in ml, above a cut-off of 6 ml, tumors were classified as "large", elsewhere they were classified as "small".

Index of lesions	Volumes (ml)	
1	1.1	
2	1.3	
3	1.3	
4	1.4	Small lesions
4	1.5	
5	1.9	
7	2.0	

8	2.2	
9	2.8	
10	4.9	
11	6.6	
12	6.6	
13	11.2	Large lesions
14	19.3	
15	27.0	
16	56.9	

In the next sections, V_r denotes the volume delineated on CT images and used for PET simulations.

III. Methods

3-1 The STAPLE algorithm

STAPLE algorithm was proposed in 2004 by Warfield *et al.* (19) as an instance of the expectation-maximization (EM) algorithm. Based on a collection of segmentation results, this algorithm computes (1) a probabilistic estimate of the ground truth and (2) a measure of the performance level of each input segmentation result.

In this study, we used the STAPLE implementation of the Computational Radiology Laboratory (CRL) in which the STAPLE algorithm was developed. This implementation is available via the CRKit software (http://crl.med.harvard.edu/).

Several parameters may affect the result of CRL STAPLE implementation. In this paragraph, sole the parameters that were varied, are described. The other parameters (maximum number of iterations, convergence threshold, initial performance level of each input segmentation result, below mentioned global prior probability) were set to their default values, assumed to be as the optimum values (19).

- Use of a consensus region

  When this parameter is "On", all voxels for which all raters agree are assigned and the STAPLE algorithm is run only on the region of uncertainty where the raters disagree. Elsewhere, all voxels are included in the computation. Both "On" and "Off" configurations were evaluated.

- Selection of a stationary prior weight

  The prior probability of the true segmentation at voxel $i, f(T_i)$, is defined as a linear combination of a global (or identical for all voxels) prior, $g(T)$, and a spatially varying (or voxelwise) prior, $s(T_i)$, (equation 1):

$$f(T_i) = w \times g(T) + (1 - w) \times s(T_i) \quad (1)$$

Where:

- The stationary prior weight, w, shows the weight of the global prior.
- The default value of the global prior probability, $g(T)$, is defined as the sample mean of the relative proportion of each label in the input segmentation results (equation 35 of the paper Warfield *et al.* (19)).
- According to the spatially varying prior, $s(T_i)$, only majority voting based spatially varying prior has, at the moment, been implemented on the CRKit software.

In this study, three cases for the stationary prior weight, w, were studied. First we set w to 1, so that the prior probability is all global prior. Then, we set w to 0 to obtain all spatially varying prior. Finally, we set w to the middle value 0.5 to generate a combined prior.

- Use of the maximum a posteriori (MAP) formulation of the STAPLE algorithm.

  In 2012, Commowick et al. (31) proposed a new version of the STAPLE algorithm in which a MAP estimate of the true segmentation is obtained by considering a beta prior probability for the performance levels. The two configurations with and without the use of the MAP STAPLE were run. When using the MAP STAPLE, the beta distribution parameters were set to the values reported in (31).

- Use of the Markov random field (MRF) in order to account for spatial homogeneity of the true segmentation.

  In the "basic" configuration of the STAPLE algorithm, a voxelwise independence assumption (i.e. the probability of the true segmentation at any given voxel is independent of the true segmentation of the neighboring or adjacent voxels) was made (19). Aware that in practical applications the true segmentation often has an underlying spatial homogeneity, the authors proposed to introduce a MRF model for incorporating spatial homogeneity. The MRF model was applied on the output of the CRL STAPLE implementation with the CRKit software. The application of the MRF model requires an homogeneous interaction strength (which is an interaction weight between voxels in the prior probability of true segmentations). The larger the interaction strength is, the smoother and more spatially homogeneous is the estimated

true segmentation. In (19), the authors found satisfactory results with an interaction strength of 2.5 for synthetic data with strong uncorrelated random noise in the segmentations and strongly homogeneous true segmentation. In this paper, six values for the interaction strength below-mentioned as MRF weight were used: 0.01, 0.1, 1, 2.5, 5 and 10. An additional weight equal to 0 was introduced when the MRF model was unused.

Table 2: Summary of the parameters evaluated.

Consensus Region	Stationary prior weight	MAP STAPLE	MRF weight
On	0	On	0
Off	1	Off	0.01
	0.5		0.1
			1
			2.5
			5
			10

Finally, 84 configurations for the CRL STAPLE implementation (Table 2) were run using the command-line utilities in CRKit software to generate 84 ground truth probabilistic estimates (or 84 STAPLE consensuses) from a collection of segmentation results.

3-2 Lesions delineation

- **Manual delinations**

Six qualified physicians manually delineated the 16 lesions. The physicians did not act together so that the delineations were conditionally independent given the ground truth and the performance level parameters as assumed in (19). Each expert delineated all the lesions for SUV= 2.5 (in a given order) before performing the delineation for SUV= 4.5 (in the same order). Thus, 6 manual delineations were obtained for each of the 16 lesions and each SUV. The physicians had a short training phase on the software used for the delineation.

- **Automatic delineations**

Five semi-automatic segmentation methods were studied: adaptive thresholding methods from Daisne *et al.* (7) and Nestle et *al.* (11, 12), a possibility theory-based algorithm or MIP based

approach from Dewalle-Vignion *et al.* (8), a fuzzy C- means clustering algorithm (FCM) (32) and a fixed threshold of 42% of the maximum SUV was also considered.

FCM method was used here with two clusters: background and tumor lesion with two features: voxel gray level and average gray level calculated in the 3×3×3 voxel neighborhood.

The calibration steps required in the Nestle and the Daisne methods were performed using simulated data obtained on a phantom and provided by the authors of (13).

Finally, for a given lesion and SUV, let V_i, respectively $V_{Daisnes}, V_{FCM}, V_{MIP}, V_{Nestle}$ and $V_{Percent}$, be the manually delineated volume defined by the i^{th}expert (with $1 < i \leq 6$, respectively the volumes obtained by the automatic methods.

We denote $\{S_j^m\}_{j=1:84}$, respectively $\{S_j^a\}_{j=1:84}$, the 84 STAPLE consensuses generated by the application of the STAPLE configurations to the six manual delineations, respectively to the five automatic segmentations results $\{V_{Daisnes}, V_{FCM}, V_{MIP}, V_{Nestle}, V_{Percent}\}$.

Let V^m , respectively V^a, be the binary volume obtained by applying the majority vote rule to the six manual delineations $\{V_i\}_{i=1:6}$, respectively to the five automatic segmentation results, where 1 is assigned to all voxels labeled as "lesion" by a majority; elsewhere 0 is assigned.

3-4 Evaluation protocol

For each lesion and each SUV, 181 estimated volumes of the ground truth were therefore obtained: the six manual delineations $\{V_i\}_{i=1:6}$, the five automatic segmentations $\{V_{Daisnes}, V_{FCM}, V_{MIP}, V_{Nestle}, V_{Percent}\}$, the 168 STAPLE consensuses $\left\{\{S_j^m\}_{j=1:84}, \{S_j^a\}_{j=1:84}\right\}$, and the two volumes obtained by the majority vote rule $\{V^m, V^a\}$,.

These volumes were analyzed to determine if the consensus from the CRL STAPLE implementation is more accurate. In order to simplify analyses, we first determined the optimal configuration of the CRL STAPLE implementation (hereinafter called optimal STAPLE), which generates the most accurate STAPLE consensus. Then, we assessed this optimal consensus by comparison to the initial segmentations results using the Dice Similarity

Coefficient (DSC). The DSC between an estimated volume, V_{estim} , and the ground truth, V_r, is defined as:

$$DSC(V_{estim}, V_r) = \frac{2|V_{estim} \cap V_r|}{|V_{estim}| + |V_r|} \times 100\% \quad (2)$$

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

For each SUV, 16 DSC (one per lesion) were computed for each of the 181 volume estimation methods.

The absolute volume difference (VD) was also computed to evaluate volume overestimation (equation 3)

$$VD = \frac{|V_{estim} - V_r|}{V_r} \times 100\% \quad (3)$$

3-4-1 Determination of the optimal CRL STAPLE configuration

In order to determine the effect of the four CRL STAPLE implementation parameters (Table 2) on the variability of the DSC, an analysis of variance (ANOVA) with seven main factors was performed using the 5376 DSC related to the STAPLE consensus (168 STAPLE consensuses × 2 SUV × 16 lesions):

- Data type: 2 levels (data from which the STAPLE consensuses were obtained),
  - manual
  - automatic
- SUV: 2 levels,
  - 2.5
  - 4.5
- Lesion number: 16 levels
- STAPLE configuration parameters summarized Table 2: 2 x 3 x 2 x 7 levels

The ANOVA model also included 21 terms for the two-way interactions between the main factors. The null hypothesis was “there is no effect of the seven main factors and their interactions on the mean DSC”, that is, no configuration provided a STAPLE consensus

closer to the ground truth than the others. A Tukey's test for pairwise comparisons of means was also performed to rank the different levels for each factor and more particularly for each CRL STAPLE implementation parameter. The statistical level of significance was set to 0.05.

3-4-2 Comparison between the optimal CRL STAPLE results and the initial segmentations results

Let S_{best}^{m} and S_{best}^{a} be the STAPLE consensuses obtained by applying the optimal STAPLE configuration to manual and automatic segmentations, respectively.

Differences in the DSC between manual delineations, $\{V_i\}_{i=1:6}$, majority vote rule volume, V^m, and STAPLE consensus, S_{best}^{m}, were tested with three-way analysis of variance (three-way ANOVA) followed by Tukey's test for pairwise comparisons of means. The three main factors were the estimation method (levels: 8 = 6 experts + 1 majority vote rule +1 optimal STAPLE configuration), the SUV (2 levels: 2.5 and 4.5) and the lesion number (16 levels). The three two-way interactions terms were also included. The null hypothesis was " there is no difference between the means of the DSC obtained from the manual delineations, the majority vote volume and the STAPLE consensus", that is, the experts, the majority vote rule and the optimal CRL STAPLE configuration did not provide estimation closer to the ground truth than those obtained by the others. The statistical level of significance was set to 0.05.
The same statistical analysis was then performed for the automatic segmentations, the majority vote volume from automatic segmentations and the optimal STAPLE consensus, S_{best}^{a}.

IV. Results

4-1 Determination of the optimal CRL STAPLE configuration

- Fisher's test

From the ANOVA analysis, we can conclude with confidence that the seven main factors and the 21 interaction factors have a significant effect on the variability of the DSC ($p<0.0001$).
From the type III Sum of Squares (type III SS) table obtained from the ANOVA analysis, the impact significance of each main factor and of each interaction factor was assessed through a Fisher's F statistic. The higher the Fisher's F statistic corresponding to a given factor is, the stronger is the impact of the factor on the variability of the DSC.

Tables 3 and 4 summarize the impact significance with the highest value.

Table 3: Summary of the Fisher's test to evaluate the influence of a given parameters as main factor on the DSC value.

Parameters	Fisher's F statistic (p-value)
Consensus	1596.8 (< 0.0001)
MRF	95.5 (< 0.0001)
MAP	43.3 (< 0.0001)
Stationary prior weight	28.5 (< 0.0001)

Table 4: Summary of the Fisher's test to evaluate the influence of interaction factors on the DSC value.

Parameters interaction	Fisher's F statistic (p-value)
MAP and consensus	72.7 (< 0.0001)
MAP and data type	67.2 (< 0.0001)
MRF weight and data type	64.2 (< 0.0001)
MAP and prior weight	47.4 (< 0.0001)
Prior weight and data type	32 (< 0.0001)
Prior weight and consensus	27.9 (< 0.0001)

- Tukey's test

Table 5 summarizes the results on the Tuckey's test to state if the mean DSC according to the parameters values was significantly different and to evaluate how the values of the different parameters lead to higher DSC.

Table 5: Summary of the Tuckey's test. Row 1 contains the different parameters evaluated, rows 2 and 3 contain the different values assigned to a given parameter and the corresponding mean DSC, row 4 contains the Tuckey's test value.

Parameters	**Consensus region**		**MRF weight**			**MAP formulation**		**Prior weight**		
	On	Off	0	5	10	On	OFF	0	0.5	1
Mean DSC	79.5 %	67.4 %	67 %	76.2 %	77.2 %	72.5 %	74.5 %	75.1 %	72.5 %	72.5%
Tuckey's test	P<0.0001		P<0.009			P=0.51		P<0.0001		

When the consensus region was used, the DSCs were almost always significantly higher except for the largest lesion volume (Tukey's tests). These higher DSCs can be explained by the segmentation errors occurring mainly at the boundaries due to uncertainties in determining tumor borders (Fig. 1). When the consensus region was not used, resulting spatially varying distribution of errors was not taken into account and led to an overestimation. Indeed, the predictions made in regions where all raters agree induced an overestimation of the certainty of the predictions in the smaller regions where the raters disagree. The smaller the lesion volume is, the larger are the regions for which all raters agree compared to the regions of uncertainty and thus, when the consensus region is not used, the larger is the overestimation of the ground truth. Consensus region assignment is therefore advisable particularly for small lesions.

The MRF model requires the specification of a MRF weight. Seven MRF weight were tested: 0, 0.01, 0.1, 1, 2.5 (19), 5 and 10. With a value of 0, the use of the MRF model was disabled. This 0-value led to smaller DSCs than the other weights (except for the three largest lesions volumes). Among the other six weights, the two higher ones, 5 and 10, leading to the two smoothest ground truth estimates, appeared to be more effective. These results can be explained by the limited spatial resolution of the PET images and the resulting distribution of the metabolic information over voxels, which can be taken into account by efficient MRF weight.

Concerning the stationary prior weight, only three cases were considered in order to limit the number of configurations. All the Tukey's tests results included, all spatially varying prior ($w=0$) was found to be equally or more relevant than the two other priors. The stationary prior weight should therefore be set to 0 (all spatially varying prior) even if more complex combined priors could be further explored to determine the optimal one.

Regarding the last parameter, the use of the MAP formulation did not improve (and even sometimes worsen) the DSCs. The beta distribution parameters set to the values reported in (31) might not be suitable for our study. We chose to use these parameters values to avoid two additional parameters and many other ensuing configurations for the CRL STAPLE implementation. Other parameters values could be further investigated.

Considering the above-reported factors ranked in descending Fisher's F statistic order (Tables 3 and 4), the parameters summarized in table 6 should be used. Note, that to avoid a too smoother STAPLE consensus with a MRF weight of 10, the value of 5 was preferred.

Table 6: Summary of the parameters found for the optimal STAPLE configuration.

Consensus Region	On
Stationary prior weight	0
MAP STAPLE	Off
MRF weight	5

Fig. 2 shows the mean DSCs, all SUVs included, obtained from the manual delineations by all the CRL STAPLE implementation configurations that correspond to a MRF weight of 5 (black bars) among which the optimal configuration (shaded gray bar). The gray bars represent the results provided by the configurations that correspond to a MRF weight of 0 (i.e., without the use of the MRF model). Fig. 3 shows the results obtained from the automatic segmentations.

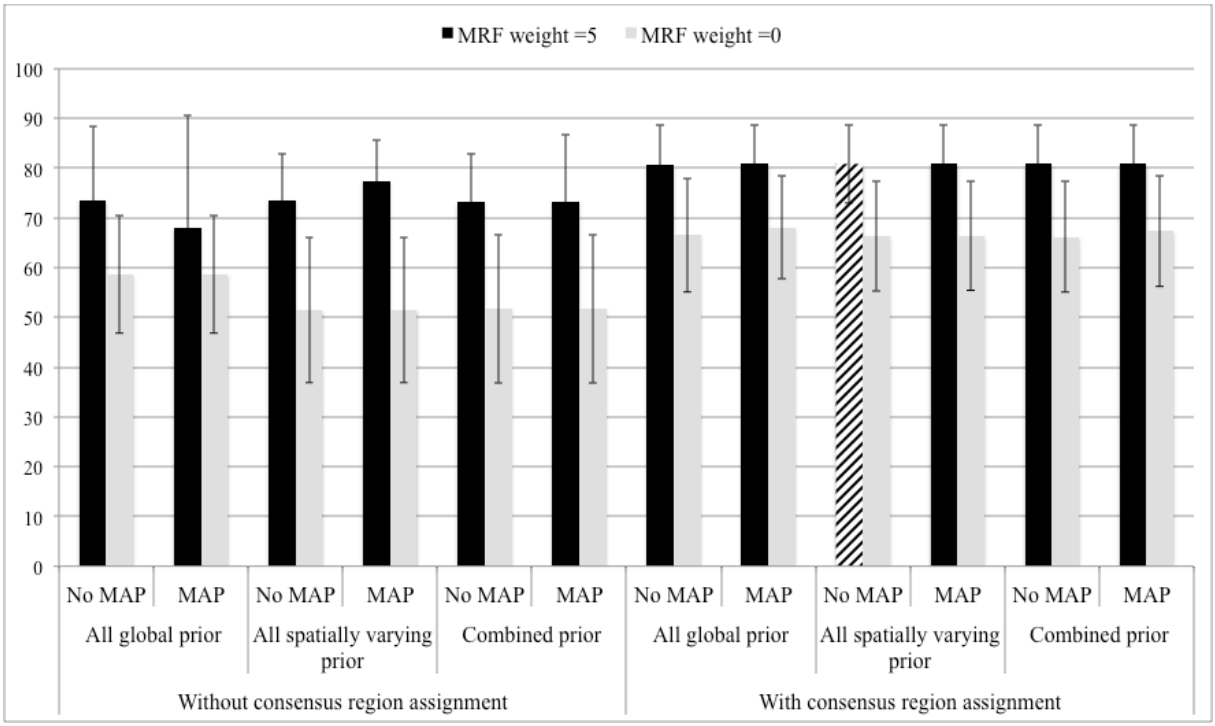


Fig. 2: Means of the DSCs obtained from the manual delineations by 12 CRL STAPLE configurations, error bars represent the mean, plus or minus one standard deviation.

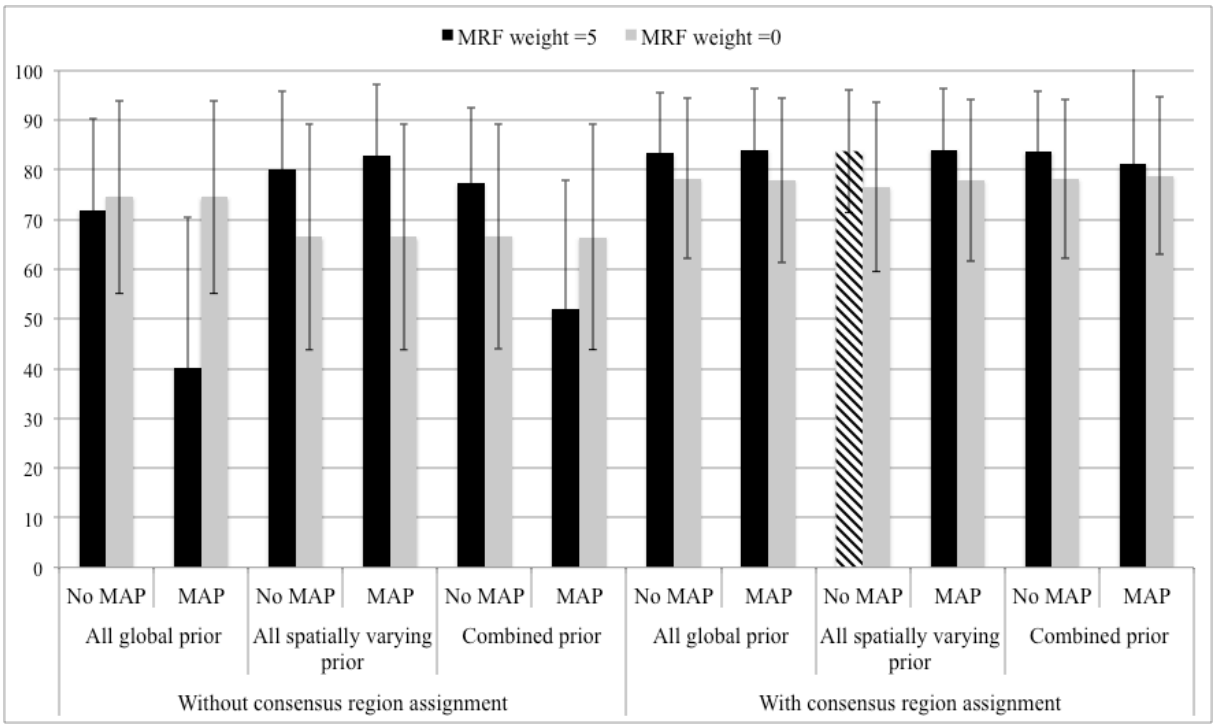


Fig. 3: Means of the DSCs obtained from the automatic segmentations by 12 CRL STAPLE configurations, error bars represent the mean, plus or minus one standard deviation.

Figs. 2 and 3 illustrate the above-mentioned highest impacts of both the consensus region assignment and the MRF model use.

4-2 Comparison between optimal STAPLE configuration and the initial segmentations results

4-2.1 Manual delineations

Fig. 4 shows an example of lesion with associated manual delineations.

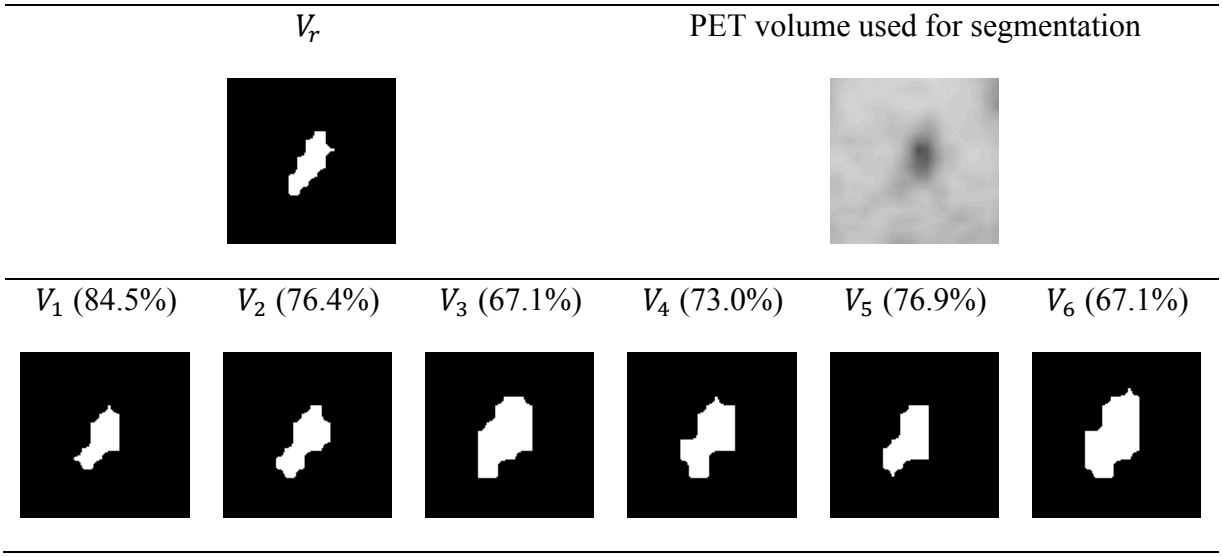
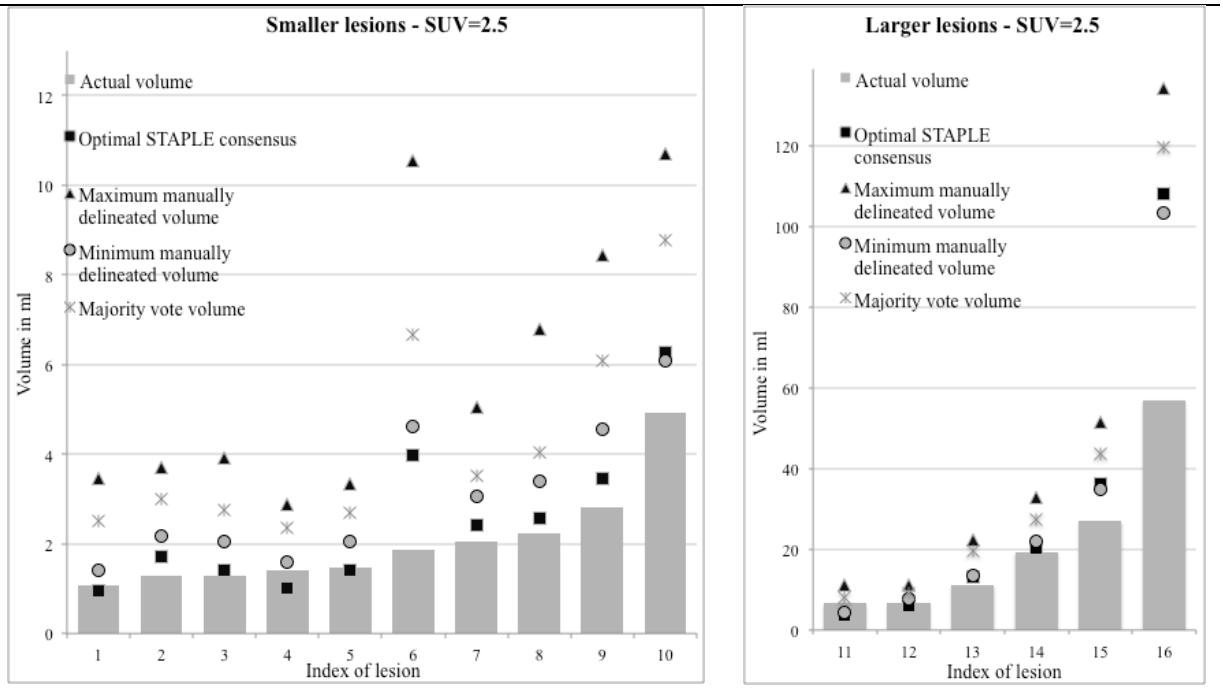


Fig. 4: Ground truth (row 1, column 1) of a given lesion (volume = 11.2 ml) and its associated simulated PET images (row 1, column 2) from which the manual delineations $\{V_i\}_{i=1:6}$ (row 2) were performed. The optimal STAPLE consensus, S^m_{Best}, and the majority vote volume V^m are displayed on row 3, column 1 and column 2, respectively. The corresponding Dice coefficients computed by comparison to V_r , the volume delineated on CT images, are indicated in brackets.

Fig. 5 gives an overview of all results as a function of the lesion size for SUV= 2.5 and for SUV= 4.5.



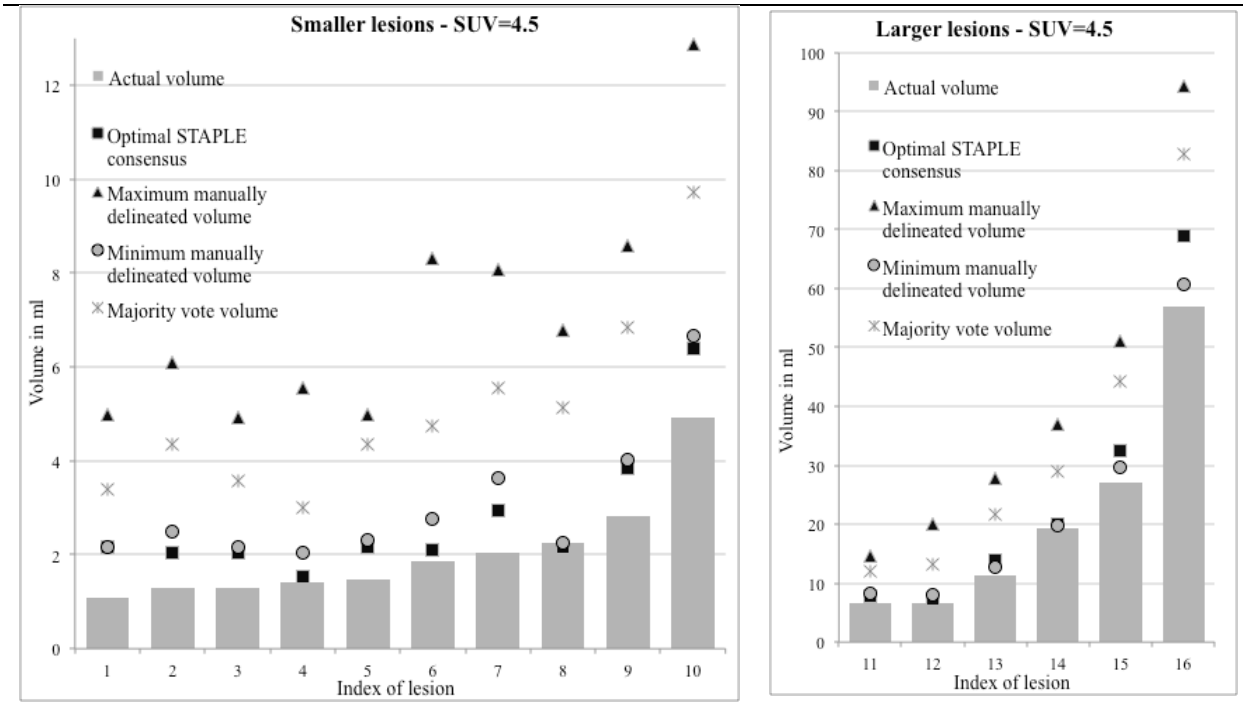


Fig. 5: Overview of the results as a function of the lesion size (in ml) for SUV= 2.5 (row 1) and SUV= 4.5 (row 2). The first column shows the results for the 10 smallest volumes and the second one for the sixth largest ones.

Fig. 5 confirms a systematic overestimation of lesion volumes by experts as already reported in an earlier study (15). This overestimation was more important for SUV= 4.5 than for SUV= 2.5.

An overestimation of lesion volumes was mostly obtained for SUV= 4.5, even with the optimal STAPLE result obtained from manual delineations (Fig. 5). Nonetheless, the optimal STAPLE result was still closer to the actual volume than the best manually delineated volume (Table 7).

Table 7: Summary of the mean absolute volume differences for optimal STAPLE consensus, majority vote volume and the best manual delineation.

	Optimal STAPLE configuration	**MV volume**	**Best manual delineation**
SUV = 2.5	30.3%	129.6%	46.4%
SUV = 4.5	31.4%	111%	40.7%

Fig. 6 and table 8 illustrate the results of the DSC when applying the optimal STAPLE configuration to manually delineated lesions.

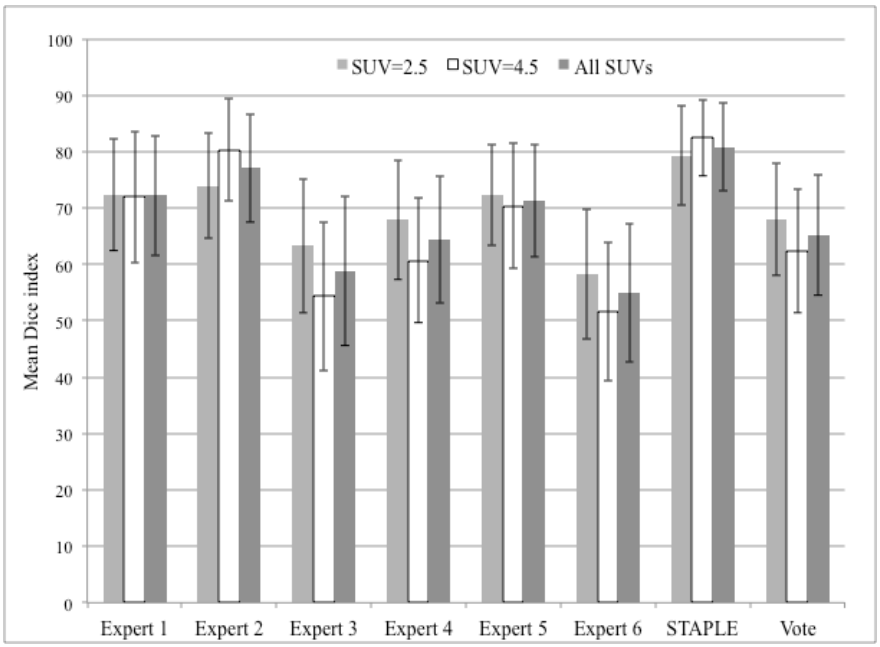


Fig. 6: Overview of the DSC means obtained between ground truth and manual delineations (first six bars), optimal STAPLE consensuses (seventh set of bars) and majority vote volumes (eighth set of bars). The error bars represent the mean, plus or minus one standard deviation.

Table 8: Summary of the Tuckey's test achieved (all SUV included) to compare optimal STAPLE consensus with overall manual delineation, best manual delineation and majority vote (MV) volume according to the mean of DSC (standard deviation). The p –value from the Tuckey's test is given to state if optimal STAPLE configuration gives significantly higher DSC than others methods.

Optimal STAPLE configuration		
80.88% (7.87%)	**Overall manual delineation** 66.44% (13.55%)	p<0.0001
	Best manual delineation 77.1% (9.59%)	p=0.001
	MV volume 65.18% (10.65%)	p<0.0001

The overall mean DSC suggests a low accuracy of manual delineations (Table 8). Moreover, a high disparity in delineation accuracy was observed on Fig. 6 between the experts.

The mean value for DSC obtained with the optimal STAPLE configuration gave higher results than all the experts. Furthermore, analyses using Tukey's test suggested that the DSCs obtained with the optimal STAPLE consensuses were significantly higher than those obtained by the six manual delineations (Tukey's tests: $p<0.0001$ except for expert 2 for which $p=0.01$) and by the majority vote volume (Tukey's test: $p<0.0001$) (Table 8).

Note that for almost all volumes below 20 ml, some manual delineations with SUV= 2.5 were actually closer to the ground truth than those obtained with SUV= 4.5 (Figs. 5 and 6). Partial volume effect may be more important for small lesions than for higher contrast levels. The impact of the partial volume effect is reduced for larger lesions and the contrast enhancement results in an improvement of the manual delineation accuracy.

4-2.2 Automatic segmentations

The results as a function of the lesion size are shown in Fig. 7 for SUV= 2.5 and SUV 4.5 Due to large volume overestimation, results from FCM method and 42% threshold-based were not plotted in order to make the graphs clearer.

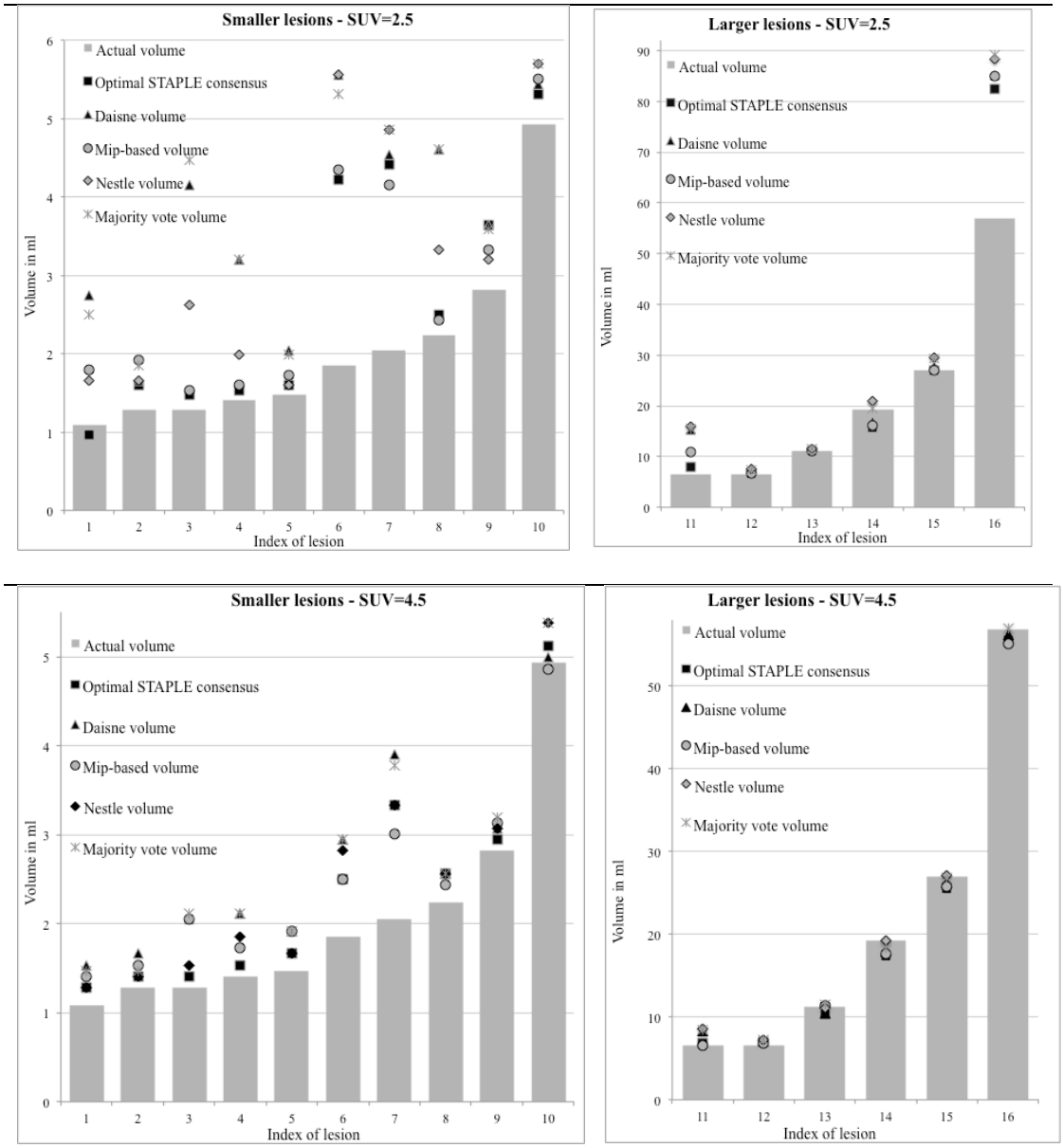


Fig. 7: Overview of the results for SUV= 2.5 (row 1) and SUV= 4.5 (row 2) as a function of the lesion size (in ml) indicated by the height of the bar, in column 1 and in column 2 for the 10 smallest volumes and the six largest ones, respectively.

Fig. 8 and table 9 represent the results in term of DSC when applying the optimal STAPLE configuration to the segmentation methods results.

Table 9: Summary of the Tuckey's test achieved (all SUV included) to compare optimal STAPLE consensuses with automatic segmentation results and majority vote volume (MV) according to the mean of DSC (standard deviation). The p –value from the Tuckey's test is

given to state if optimal STAPLE configuration gives significantly higher DSC than automatic methods.

Optimal STAPLE configuration	**Daisne**	p=0.02
	78.2% (16.46%)	
	FCM	p<0.0001
	67.69% (20.56%)	
	MIP-Based	p=0.915
	82.27% (13.7%)	
83.82% (15.86%)	**Nestle**	p=0.33
	80.88% (16.53%)	
	42 Percent	p<0.0001
	73.59% (21.79%)	
	MV volume	p=0.007
	78.84% (17.73%)	

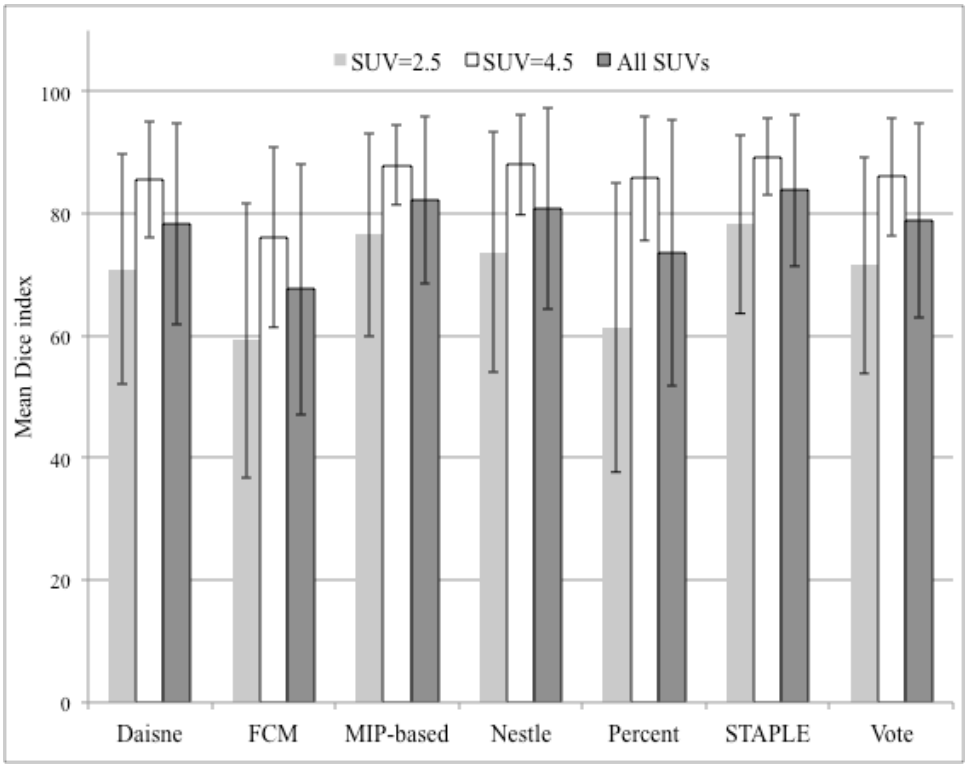


Fig.8: The means of the DSC obtained with the five automatic methods (first five sets of bars), by the optimal CRL STAPLE configuration (sixth set of bars) and by the majority vote rule (seventh set of bars). The error bars represent the mean, plus or minus one standard deviation.

The mean DSC obtained with the optimal STAPLE consensuses was 78.3% for SUV= 2.5 and 89.3% for SUV= 4.5 (Fig. 9). These high mean values, higher than those obtained with all the five automatic methods, indicated the reliability of the optimal STAPLE configuration. For eight (respectively, nine) of the 16 lesions with SUV= 2.5 (respectively, SUV= 4.5), at least one of the five DSCs obtained with the automatic methods (mainly with the MIP-based and the Nestle methods) was higher than the one obtained with the optimal STAPLE consensuses. For SUV= 2.5, no particular distribution of these eight lesions was observed while for SUV= 4.5, the nine lesions were among the 10 largest ones.

Tukey's test indicated that the DSCs obtained from optimal STAPLE configuration were not significantly different from those obtained with the MIP-based method and the Nestle method but were significantly higher than those obtained by the FCM method, the percent method, the Daisne method and the majority vote rule (Table 9).

V. Discussion

In this paper, we studied whether using the STAPLE algorithm would provide a reliable solution to assess the accuracy of tumor volume estimate in PET imaging. The STAPLE algorithm computes a probabilistic estimate of the ground truth from a set of (automatic or manual) segmentations results. The evaluation was performed using simulated data with ground truth, allowing comparison of STAPLE consensus with the actual ground truth. A wide range of lesion sizes was explored (Figs. 5 and 6).

Due to the large number of possible configurations for the CRL STAPLE implementation, a preliminary stage to determine the optimal configuration in PET imaging was performed using DSC. From this preliminary stage, the use of the consensus region and the MRF were found to have a high effect on the variability of the DSC.

For the second stage consisting of the evaluation of the optimal configuration for the CRL STAPLE implementation, we are aware that it would have been preferably performed on data other than those used for its determination. Nonetheless, due to the time consumption for generating simulated PET data and performing the manual delineations, the same data were used.

The evaluation of the optimal STAPLE was then conducted in two steps. First, the STAPLE algorithm was applied on manual delineations performed by a panel of six experts. Then, the STAPLE algorithm was applied to five automatic segmentation results (7, 8, 11, 12, 32). In each case, the resulting STAPLE consensus was compared to the ground truth using DSC. The question was whether the STAPLE consensus provided a more accurate estimate of the ground truth than those obtained by the experts or using the automatic segmentation methods.

Our results suggest that the consensus obtained from the optimal STAPLE configuration provides a more accurate estimate of the lesion volume than initial manual delineations or automatic segmentations. Although the overlap between the ground truth and the optimal STAPLE consensus is not perfect, the optimal STAPLE configuration can therefore be useful to assess tumor segmentation methods in PET imaging and can be preferred to the more common majority vote rule. Nevertheless, we can underline that the optimal STAPLE configuration applied to manual delineations leads to DSC values above 80%. This cut-off value is generally admitted to reflect an “almost perfect agreement” between ground truth and estimates (33-35).

From these promising results, different issues might be investigated. Simulated PET images used for this study exhibit insufficient heterogeneity to appear clinically realistic. More realistic images will be investigated such as database described in Papadimitrioulas et al. (36) where authors introduced heterogeneity models.

Similarly to experts, FCM and 42% threshold-based methods led to a systematic lesion volume overestimation. FCM method appeared to be the least effective automatic method (Fig. 8, Table 9). Partially due to its dependence to SUV, 42% threshold-based method that yielded about as unsatisfactory results as FCM method for SUV= 2.5, gave relevant results for SUV=4.5. With more conclusive results than FCM and 42% threshold-based methods, Daisne method however was less efficient than Nestle method and MIP based approach.

Moreover, if optimal STAPLE configuration clearly improved the manual delineations, this was less clear-cut for automatic segmentation methods. Optimal STAPLE consensus was not significantly more accurate than MIP-based volume or Nestle volume for instance. Indeed, STAPLE estimate can be viewed as an underlying weighting average of the initial segmentations results, as a result it might be outperformed by some of these initial segmentation results with higher accuracy. Thus, when excluding segmentation results with lower accuracy a better consensus might be expected. Nevertheless, this is not identifiable for clinical data of unknown ground truth and similar results we obtained on DSC of the high accuracy methods and the STAPLE CRL reveal the robustness of the STAPLE consensus result. Furthermore, when excluding segmentations results, new performance levels and new weights would be obtained such that the new STAPLE consensus might or might not be more accurate. In that context, the use of other recent segmentation methods (37-40) should be also explored to better evaluate the robustness of the approach. Last, regression without truth methods could also be explored to determine whether they are adapted to assess the accuracy of tumor segmentation methods in PET (41).

Use of consensus region included in the STAPLE algorithm should also be further explored since Van Leemput et al. (42) demonstrated the impact of the number of raters on the final result. High number of segmentations approaches or of raters could lead to a consensus result that may not reflect the individual (approach or manual rater) performances. We observed that using the consensus region assignment yielded to very small influence of the stationary prior weight. Obviously, when the consensus region is assigned, the STAPLE algorithm is run only on the region of uncertainty where the raters disagree. The influence of the stationary prior weight thus concerns only this region of uncertainty, which is much too small compared to the entire volume to significantly impact the DSCs (computed on the entire volume).

Finally, the results described in this paper were in good agreement with the study from McGurk et al. (26). Indeed, optimal CRL STAPLE configuration improved the segmentation results and, in clinical routine, might provide I) a reliable PET-volume segmented combining different segmentation results and II) a consistent ground truth out of the manual delineation by different experts. With respect to previous work (26), added values of our study mainly rely on the methodology applied where an optimization of the “tuning” values of STAPLE algorithm was achieved. Manual delineations were also used as input data. Thus more than

segmentation process, the ability of optimal STAPLE configuration to provide a relevant surrogate from manual delineations was estimated. Furthermore, the optimal configuration of STAPLE CRL achieved in this paper was a preliminary stage in a multicentric study dedicated to radiation oncology planning from PET/CT images and is being applied for segmentation purpose. Results of this multicentric study will be the subject of a further clinical oriented paper.

VI. Conclusion

In this paper, we have evaluated the ability of the STAPLE algorithm (19) to yield an accurate estimate of tumor volumes based on several estimates in PET imaging. The evaluation, which was performed using simulated data with known ground truth (29), involved a particular configuration of the STAPLE implementation of the Computational Radiology Laboratory (CRL) (31). When using the results from 6 manual segmentations as an input, the STAPLE algorithm succeeded in providing a better estimate of the tumor volumes than the initial manual delineations. This result was also obtained for the 5 considered automatic segmentation methods.

Based on these results, we conclude that the particular configuration used in this paper is an appropriate tool to assess the accuracy of tumor segmentation in PET.

Acknowledgments:

The authors would like to thank Claude Hossein-Foucher, Amandine Béron, Grégory Petyt, Pierre Lenfant and Damien Huglo from the Department of Nuclear Medicine, University Hospital, Lille, F-59000, France, for their involvement in the tumor manual delineations used in this article.

The authors would like to thank Irène Buvat and Simon Stute from the IMNC, UMR 8165 CNRS, Paris 7 and Paris 11 Universities F-91406 Orsay, France, for giving access to the simulated PET images.

The authors would like to thank L. Massoptier from Aquilab Company for his cooperation in this study.

This work was partially supported by EU project E5949 SALOME under Eurostars Programme, which is powered by EUREKA and the European Community.

REFERENCES

1. Baba S, Abe K, Isoda T, Maruoka Y, Sasaki M, Honda H. Impact of fdg-pet/ct in the management of lymphoma. Annals of nuclear medicine. 2011;25(10):701-16.
2. Delouya G, Igidbashian L, Houle A, Belair M, Boucher L, Cohade C, et al. (1)(8)f-fdg-pet imaging in radiotherapy tumor volume delineation in treatment of head and neck cancer. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2011;101(3):362-8.
3. Avril N, Sassen S, Roylance R. Response to therapy in breast cancer. J Nucl Med. 2009;50 Suppl 1:55S-63S.
4. Ashamalla H, Rafla S, Parikh K, Mokhtar B, Goswami G, Kambam S, et al. The contribution of integrated pet/ct to the evolving definition of treatment volumes in radiation treatment planning in lung cancer. Int J Radiat Oncol Biol Phys. 2005;63(4):1016-23.
5. Buijsen J, van den Bogaard J, van der Weide H, Engelsman S, van Stiphout R, Janssen M, et al. Fdg-pet-ct reduces the interobserver variability in rectal tumor delineation. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2012;102(3):371-6.
6. Greco C, Rosenzweig K, Cascini GL, Tamburrini O. Current status of pet/ct for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (nsclc). Lung Cancer. 2007;57(2):125-34.
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. 2003:247-50.
8. 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(2):409-23.
9. Geets X, Lee JA, Bol A, Lonneux M, Gregoire V. A gradient-based method for segmenting fdg-pet images: Methodology and validation. Eur J Nucl Med Mol Imaging. 2007;34(9):1427-38.
10. Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. Pet functional volume delineation: A robustness and repeatability study. Eur J Nucl Med Mol Imaging. 2011;38(4):663-72.
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. 2005:1342-8.
12. 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. 2008:1989-99.
13. Stute S, Vauclin S, Necib H, Grotus N, Tylski P, Rehfeld NS, et al. Realistic and efficient modeling of radiotracer heterogeneity in monte carlo simulations of pet images with tumors. Nuclear Science, IEEE Transactions on. 2012;59(1):113-22.
14. Buijsen J, van den Bogaard J, Janssen MH, Bakers FC, Engelsman S, Ollers M, et al. Fdg-pet provides the best correlation with the tumor specimen compared to MRI and ct in rectal cancer. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2011;98(2):270-6.

15. 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. 2004:93-100.
16. Jeganathan R, McGuigan J, Campbell F, Lynch T. Does pre-operative estimation of oesophageal tumour metabolic length using 18f-fluorodeoxyglucose pet/ct images compare with surgical pathology length? Eur J Nucl Med Mol Imaging. 2011;38(4):656-62.
17. van Baardwijk A, Bosmans G, van Suylen RJ, van Kroonenburgh M, Hochstenbag M, Geskes G, et al. Correlation of intra-tumour heterogeneity on 18f-fdg pet with pathologic features in non-small cell lung cancer: A feasibility study. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2008;87(1):55-8.
18. Dwamena BA, Sonnad SS, Angobaldo JO, Wahl RL. Metastases from non-small cell lung cancer: Mediastinal staging in the 1990s-meta-analytic comparison of pet and ct ben a. Dwamena, seema s. Sonnad, jeff o. Angobaldo, and richard l. Wahl. Radiology. 1999:530-6.
19. 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(7):903-21.
20. Makni N, Iancu A, Colot O, Puech P, Mordon S, Betrouni N. Zonal segmentation of prostate using multispectral magnetic resonance images. Med Phys. 2011;38(11):6093-105.
21. Jomier J, LeDigarcher V, Aylward SR. Comparison of vessel segmentations using staple. Med Image Comput Comput Assist Interv. 2005;8(Pt 1):523-30.
22. Tustison NJ, Avants BB, Flors L, Altes TA, de Lange EE, Mugler JP, 3rd, et al. Ventilation-based segmentation of the lungs using hyperpolarized (3)he MRI. Journal of magnetic resonance imaging : JMRI. 2011;34(4):831-41.
23. Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;77(3):959-66.
24. Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and alzheimer's disease. Neuroimage. 2010;51(4):1345-59.
25. Gordon S, Lotenberg S, Long R, Antani S, Jeronimo J, Greenspan H. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Comput Med Imaging Graph. 2009;33(3):205-16.
26. 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(4):042501.
27. Jan S, Santin G, Strul D, Staelens S, Assie K, Autret D, et al. Gate: A simulation toolkit for pet and spect. Phys Med Biol. 2004;49(19):4543-61.
28. Rehfeld NS, Stute S, Apostolakis J, Soret M, Buvat I. Introducing improved voxel navigation and fictitious interaction tracking in gate for enhanced efficiency. Phys Med Biol. 2009;54(7):2163-78.
29. Stute S, Carlier T, Cristina K, Noblet C, Martineau A, Hutton B, et al. Monte carlo simulations of clinical pet and spect scans: Impact of the input data on the simulated images. Phys Med Biol. 2011;56(19):6441-57.
30. Jan S, Benoit D, Becheva E, Carlier T, Cassol F, Descourt P, et al. Gate v6: A major enhancement of the gate simulation platform enabling modelling of ct and radiotherapy. Phys Med Biol. 2011;56(4):881-901.
31. 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(8):1593-606.

32. Dunn JC. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics. 1973:32-57.
33. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
34. Yamamoto T, Kabus S, von Berg J, Lorenz C, Chung MP, Hong JC, et al. Reproducibility of four-dimensional computed tomography-based lung ventilation imaging. Acad Radiol. 2012;19(12):1554-65.
35. 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(2):178-89.
36. 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(11):112506.
37. Willaime JM, Aboagye EO, Tsoumpas C, Turkheimer FE. A multifractal approach to space-filling recovery for pet quantification. Med Phys. 2014;41(11):112505.
38. Zeng Z, Wang J, Tiddeman B, Zwiggelaar R. Unsupervised tumour segmentation in pet using local and global intensity-fitting active surface and alpha matting. Computers in biology and medicine. 2013;43(10):1530-44.
39. Abdoli M, Dierckx RA, Zaidi H. Contourlet-based active contour model for pet image segmentation. Med Phys. 2013;40(8):082507.
40. Ballangan C, Wang X, Fulham M, Eberl S, Feng DD. Lung tumor segmentation in pet images using graph cuts. Computer methods and programs in biomedicine. 2013;109(3):260-8.
41. Lebenberg J, Buvat I, Lalande A, Clarysse P, Casta C, Cochet A, et al. Nonsupervised ranking of different segmentation approaches: Application to the estimation of the left ventricular ejection fraction from cardiac cine MRI sequences. IEEE Trans Med Imaging. 2012;31(8):1651-60.
42. Van Leemput K, Sabuncu MR. A cautionary analysis of staple using direct inference of segmentation truth. Medical image computing and computer-assisted intervention–miccai 2014: Springer; 2014. p. 398-406.