Impact of the accuracy of automatic tumour functional volume delineation on radiotherapy treatment planning

To cite this article: Amandine Le Maitre et al 2012 Phys. Med. Biol. 57 5381

View the article online for updates and enhancements.

Related content
- Feasibility of a semi-automated contrast-oriented algorithm for tumor segmentation in retrospectively gated PET images: phantom and clinical validation. Montserrat Carles, Tobias Fechter, Ursula Nemer et al.

Recent citations
- Toward a standard for the evaluation of PET-Auto-Segmentation methods following the recommendations of AAPM task group No. 211: Requirements and implementation. Beatrice Berthon et al.
- Validating PET segmentation of thoracic lesions—is 4D PET necessary? Martin S Nielsen and Jesper Carl.
Impact of the accuracy of automatic tumour functional volume delineation on radiotherapy treatment planning

Amandine Le Maitre\textsuperscript{1}, Mathieu Hatt\textsuperscript{1,4}, Olivier Pradier\textsuperscript{1,2}, Catherine Cheze-le Rest\textsuperscript{1,3} and Dimitris Visvikis\textsuperscript{1}

\textsuperscript{1} INSERM, UMR 1101 LaTIM, CHRU Morvan, Brest, France
\textsuperscript{2} Department of Radiotherapy, CHRU Morvan, Brest, France
\textsuperscript{3} Academic Department of Nuclear Medicine, CHU Poitiers, Poitiers, France

E-mail: hatt@univ-brest.fr

Received 21 February 2012, in final form 27 June 2012
Published 3 August 2012
Online at stacks.iop.org/PMB/57/5381

Abstract
Over the past few years several automatic and semi-automatic PET segmentation methods for target volume definition in radiotherapy have been proposed. The objective of this study is to compare different methods in terms of dosimetry. For such a comparison, a gold standard is needed. For this purpose, realistic GATE-simulated PET images were used. Three lung cases and three H&N cases were designed with various shapes, contrasts and heterogeneities. Four different segmentation approaches were compared: fixed and adaptive thresholds, a fuzzy C-mean and the fuzzy locally adaptive Bayesian method. For each of these target volumes, an IMRT treatment plan was defined. The different algorithms and resulting plans were compared in terms of segmentation errors and ground-truth volume coverage using different metrics ($V_{95}$, $D_{95}$, homogeneity index and conformity index). The major differences between the threshold-based methods and automatic methods occurred in the most heterogeneous cases. Within the two groups, the major differences occurred for low contrast cases. For homogeneous cases, equivalent ground-truth volume coverage was observed for all methods but for more heterogeneous cases, significantly lower coverage was observed for threshold-based methods. Our study demonstrates that significant dosimetry errors can be avoided by using more advanced image-segmentation methods.

(Some figures may appear in colour only in the online journal)
1. Introduction

The use of multimodality positron emission tomography/computed tomography (PET/CT) images has been shown to improve target volume definition for radiotherapy treatment planning (RTP) by reducing in particular inter- and intra-observer variability of the target volume delineation (Steenbakkers et al 2006, Buijsen et al 2012, Daisne et al 2005). PET images are also considered for applications such as dose redistribution (South et al 2008), dose boosting (Lee et al 2008, Chao et al 2001) or dose painting (Bentzen 2005, Sovik et al 2009). However the limited spatial resolution of PET systems (4–5 mm in the centre of the field of view) results in significant partial volume effects (PVE) (Soret et al 2007). In addition, due to the statistical nature of the PET acquisition, images are affected by a significant level of noise. Consequently, manual delineation of PET volumes is tedious, time consuming and prone to high inter- and intra-observer variability (Hatt et al 2010a, 2011a). In order to facilitate and improve functional volume delineation, several fast and semi-automatic algorithms have been proposed in the past few years (Belhassen et al 2010, Aristophanous et al 2007, Hatt et al 2009). However, most of the methods currently used in clinical practice are still based on the use of some form of binary threshold, either fixed (Erdi et al 1997, Paulino et al 2004), or adaptive using tumour-to-background (T/B) ratios (Daisne et al 2003, Nestle et al 2005). The major limitations of these algorithms are their dependence on optimization using phantom acquisitions of homogeneous spheres and the user-dependent manual determination of the background value. As a result they often fail to provide satisfactory delineation of tumours characterized by heterogeneous activity distributions and do not provide reproducible results for small tumours with low contrast (Hatt et al 2011b, Nestle et al 2005). Considering the plethora of segmentation approaches based on various advanced image processing paradigms currently available (Hatt et al 2012, Zaidi and El Naqua 2011), there is a lack of consensus regarding the automatic delineation of PET uptakes, with no clear guidelines on how to incorporate PET information into target definition.

Several studies have already compared the target volumes obtained using different automatic methods with the CT target volume in various tumour localizations (Schinagl et al 2007). However, to our knowledge the impact of the PET delineation methodology on the radiotherapy planning dosimetry has been assessed only by a few investigators (Geets et al 2006). Therefore the objective of this work was to investigate the actual impact of accurate PET uptake delineation in RTP in terms of dosimetry. In order to evaluate the potential impact, the different treatment plans have to be compared to one gold standard volume coverage. For this purpose, dosimetry was computed on simulated datasets in order to ensure knowledge and control of the necessary ground truth.

2. Materials and methods

2.1. Datasets

The data used in this work are simulated $^{18}$F-FDG PET images based on corresponding clinical PET/CT datasets (Le Maitre et al 2009), the objective being to have clinically realistic images (in terms of anatomy, radiotracer distribution, voxel sampling, texture and noise levels) with a known voxel-based ground truth. One clinical PET/CT dataset was also included in our study in order to compare the range of results with those obtained using the simulated datasets.

The simulation process consists of two major steps: the creation of the patient’s model and the simulation of the PET acquisition. We chose to focus on two different tumour localizations where radiotherapy is a major treatment regime; namely non-small cell lung cancer (NSCLC)
and head and neck (H&N) cancer, using the non-uniform rational B-spline-based cardiac torso (NCAT) (Segars 2009) and the Zubal (Zubal et al 1994) phantoms, respectively. In this work the lung cases were simulated without respiratory motion in order to improve the robustness of the analysis considering the objectives targeted in this work.

Although the NCAT phantom is based on the use of non-uniform rational B-splines allowing model flexibility, the details for H&N anatomical structures are not complete (for example, the parotid glands are not modelled). This motivated the use of the more detailed Zubal phantom for the H&N cases. In order to provide more interesting and challenging comparison cases, complex tumour shapes and activity distributions were simulated based on our previously proposed methodology (Le Maitre et al 2009). In each of these phantoms, organs are associated with a label defining an activity level and an attenuation coefficient. The activity levels were derived from the region of interest (ROI) analysis on corresponding clinical images used as a model for designing the simulated cases. Acquisitions of PET images with a Philips GEMINI PET scanner (2 min per bed position) were simulated using the Monte Carlo simulation tool Geant4 application for tomography emission (Jan et al 2004) combined with a model of the PET scanner previously developed and validated (Lamare et al 2006). The resulting simulated list-mode data were subsequently reconstructed using the one-pass list mode expectation maximization (OPL-EM) reconstruction algorithm with previously optimized parameters (Lamare et al 2006). Apart from these simulated functional images, corresponding synthetic CT datasets, necessary for the dosimetry calculations, were derived by replacing each label in the simulated phantoms voxelized maps by the corresponding Hounsfield unit.

Three localizations were considered for both the NSCLC and the H&N cases. Within some of these localizations, different tumour sizes, contrasts and heterogeneities were designed in order to compare for each of these localizations the impact of the delineation accuracy on the final dosimetry. Figure 1 shows the six simulated lung and H&N tumours with the corresponding variations in heterogeneity and contrast considered. The first NSCLC case was placed in the middle lobe of the right lung. Three sizes of intra-tumour high uptake regions were designed (representing 12%, 41% and 53% of the overall tumour volume for cases 1a, 1b and 1c, respectively). The contrast between high and low uptake areas was simulated as 2:1, 2:1 and 1.8:1 for cases 1a, 1b and 1c, respectively. The second case was placed in the upper lobe of the left lung. Case 2a is the same as case 1a, while case 2b is half the volume of case 2a (69 cm$^3$ versus 35 cm$^3$). The contrast between the high and low uptake areas was set at 2:1 and 1.8:1 for cases 2a and 2b, respectively. The third tumour was placed in the lower lobe of the left lung and simulated with a necrotic centre (volume of 19 and 30 cm$^3$ for cases 3a and 3b, respectively). For the three H&N cancer cases, both homogeneous and heterogeneous tumour activity distributions were considered. The first tumour was simulated with a homogeneous uptake and placed in the mandible with two T/B contrasts (9.5:1 and 1.8:1). For the second case the same tumour shape was simulated with heterogeneous (contrast between the two uptake areas of 1.7:1) and homogeneous activity distribution (T/B ratio of 3:1). The third tumour was simulated as heterogeneous considering two different locations of the heterogeneous sub-volumes within the tumour. In case 3a the high uptake area was placed at the outer rim of the tumour with a contrast of 2.3:1, while in case 3b the high uptake and low uptake positions were reversed and the contrast was set at 2.6:1.

One clinical H&N case was finally included in our study (see figure 7) in order to allow a comparison with the results obtained using the simulated data. The data were acquired on a GE Discovery PET/CT system. The images were reconstructed using Fourier rebinning and voxel size of 4.7 × 4.7 × 3.3 mm$^3$. The approximate measured T/B was 12:1.
Figure 1. Illustration of the six tumour cases (three lung tumours and three H&N tumours). For each case (same patient) varying configurations of contrast and heterogeneity (a)–(c) were considered.

2.2. Tumour volume definition

As already mentioned the use of simulated datasets allows knowing exactly the ground-truth volume, which was considered here to be the gross tumour volume (GTV). Within the context of this study, GTVs were defined on the PET images only, the assumption being that there was no part of the anatomical volume without elevated PET uptake. Four automatic segmentation algorithms were compared. Two are based on the use of thresholding considering both fixed and adaptive thresholds, while the other two are based on more ‘advanced’ image-segmentation paradigms; namely the fuzzy C-mean (FCM) clustering (Boudraa et al 1996) and the fuzzy locally adaptive Bayesian (FLAB) algorithm (Hatt et al 2009, 2010b, 2011a).

For the fixed threshold a value of 42% of the maximum was used based on the original work by Erdi et al (1997), denoted from here onwards as T42. The adaptive thresholding method
Dosimetry impact of functional tumour delineation (Daisne et al 2003) is based on the signal to background ratio (SBR): threshold(%) = a + b \times \sqrt{\text{SBR}}, where a and b are scanner-specific parameters obtained by linear regression. We calculated a and b with several simulations of the IEC phantom (NEMA 2-2001 IQ Phantom, T/B ratios of 4:1, 6:1, 8:1, 12:1 and 16:1) using the Philips GEMINI PET scanner model (2 min acquisition time), and reconstructed with 4 × 4 × 4 mm³ voxels using the OPL-EM algorithm. For each sphere in the IEC phantom (excluding the 10 mm sphere due to the large voxel sizes and PVE), the SBR was measured and the threshold which led to the lowest error was found by exhaustive search (all the possible cases were tested and the best one was chosen). Linear regression was conducted for all these points (threshold as a function of SBR only, as the approach does not assume any a priori information regarding object size) in order to determine parameters a and b for this particular data simulation and reconstruction configuration (a = 34.8 and b = 59.2). In order to evaluate the influence of the background region placement during adaptive thresholding segmentation, two different users were instructed to manually define a background ROI (a few centimetres away from the tumours for lung cases, and in low uptake regions for H&N), which led to two different thresholds and therefore two different segmentation results (denoted from here onwards as A1 and A2).

FCM and FLAB are both able to handle homogeneous and heterogeneous uptakes, allowing in principle to differentiate several classes within the tumour, whereas threshold-based methods only differentiate tumour and background. FCM is a clustering-based method that only considers the intensities of voxels and which has been previously used for PET segmentation (Boudraa et al 1996, Kim et al 2007). It consists in defining for each voxel a degree of membership to a cluster by minimizing the distance between the voxel value and cluster centre. Although it is based on a fuzzy model, this process does not explicitly model PVE in PET imaging. FLAB is a method based on statistical and fuzzy modelling specifically accounting for PET image characteristics such as noise and low spatial resolution (Hatt et al 2009). In contrast to FCM, it also takes into account spatial correlation between neighbouring voxels in a local fashion which makes it more robust to noise.

Although any identified tumour sub-volumes can be of interest for dose painting or dose boosting purposes, only a uniform dose prescription to the tumour volume was considered in this study in order to allow a fair comparison of the segmentation approaches considered. Therefore, for the cases where FLAB and FCM delineated two different sub-volumes within the tumour, the union of the two sub-volumes was considered as the target volume. For the clinical case, GTV was only delineated with FLAB and one adaptive threshold (denoted from here on as FLAB and A) applied separately to two ROIs, the first for the large high contrast and heterogeneous uptake, the second for the several lower uptakes close to each other (see figure 7).

The segmentation processes resulted in binary images, containing tumour and background voxels, which were transformed into DICOM datasets using the ITK DICOM library in order to import them within the treatment planning system. GTVs were then defined by thresholding these masks within the Pinnacle™ treatment planning system (Philips Healthcare, research version 8.7y). The clinical target volumes (CTVs) were derived from the GTVs by adding a 3 mm margin for microscopic extensions. Since no respiratory motion or setup errors were considered in our simulated datasets, the CTV considered is equivalent to the PTV.

2.3. IMRT treatment planning

The Pinnacle™ treatment planning system was used for IMRT planning and dose calculation. For the lung cancer cases five photon beams of 6 MV nominal energy with angles of 0°, 72°, 144°, 216° and 288° were used. For the H&N cases, seven photon beams of 6 MV nominal
Table 1. Constraints to the OARs for the lung cases.

<table>
<thead>
<tr>
<th>Organ</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lungs</td>
<td>$D_{35\text{max}} = 20 \text{ Gy}$</td>
</tr>
<tr>
<td></td>
<td>$D_{20\text{max}} = 30 \text{ Gy}$</td>
</tr>
<tr>
<td>Heart</td>
<td>$D_{100\text{max}} = 40 \text{ Gy}$</td>
</tr>
<tr>
<td>Spinal cord</td>
<td>$D_{\text{max}} = 42 \text{ Gy}$</td>
</tr>
</tbody>
</table>

Table 2. Constraints to the OARs for the H&N cases.

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Plan 50 Gy</th>
<th>Plan 70 Gy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parotid glands</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$D_{20\text{max}} = 31 \text{ Gy}$</td>
<td>$D_{20\text{max}} = 43 \text{ Gy}$</td>
</tr>
<tr>
<td></td>
<td>$D_{40\text{max}} = 20 \text{ Gy}$</td>
<td>$D_{40\text{max}} = 28 \text{ Gy}$</td>
</tr>
<tr>
<td></td>
<td>$D_{60\text{max}} = 9 \text{ Gy}$</td>
<td>$D_{60\text{max}} = 13 \text{ Gy}$</td>
</tr>
<tr>
<td>Ears</td>
<td>$D_{\text{max}} = 30 \text{ Gy}$</td>
<td>$D_{\text{max}} = 48 \text{ Gy}$</td>
</tr>
<tr>
<td>Spinal cord</td>
<td>$D_{\text{max}} = 42 \text{ Gy}$</td>
<td>$D_{\text{max}} = 42 \text{ Gy}$</td>
</tr>
<tr>
<td>Cerebral falx</td>
<td>$D_{\text{max}} = 34 \text{ Gy}$</td>
<td>$D_{\text{max}} = 48 \text{ Gy}$</td>
</tr>
</tbody>
</table>

Energy with angles of $0^\circ$, $50^\circ$, $100^\circ$, $150^\circ$, $210^\circ$, $260^\circ$, and $310^\circ$ were used. These choices were made according to usual clinical plans as routinely defined in our radiotherapy department.

A uniform dose was prescribed within the PTV. According to doses clinically used, 66 Gy was prescribed to the PTV in 2 Gy fractions for the lung cases. For the H&N we prescribed 50 Gy to the volume enclosing the tumour and the node extensions (PTV1). Then an additional dose of 20 Gy was prescribed specifically to the tumour volume (PTV2) in 2 Gy fractions, for a total of 70 Gy delivered to the tumour (PTV2). The constraints to the organs at risk (OARs) considered for the IMRT plans are summarized in tables 1 and 2. The direct machine parameter optimization algorithm was used for the dose calculation.

2.4. Plan comparison

GTVs delineated on PET images by the four approaches were compared to the ground truth. The comparison metrics used were volume error (VE), sensitivity and positive predictive value (PPV). As the CTV (= PTV) was derived from the GTV with an added 3 mm margin, the different plans were compared to the volume derived from the ground-truth volume with the addition of the same 3 mm margin (PTVGT), in order to avoid a systematic bias of volume overestimation.

Several measures can be used to assess the quality of volume coverage of a treatment plan. We chose to calculate the percentage of target volume (PTVGT) receiving 95% of the prescribed dose ($V_{95}$) and the percentage of dose received by 95% of the target volume (D95). The homogeneity of the dose within the target volume was also assessed by the homogeneity index (HI) expressed by

$$HI = \frac{D_{\text{max}} - D_{\text{min}}}{D_p} \times 100\%,$$

where $D_p$ is the prescribed dose, and $D_{\text{min}}$ and $D_{\text{max}}$ are the doses for 98% and 2%, respectively, of the target volume. The conformity of the treatment plans to the PTVGT was finally assessed using the conformity index (CI) (see figure 2):
Dosimetry impact of functional tumour delineation 5387

Figure 2. Illustration of the CI.

\[ CI = \frac{TV_{ir}}{TV} \times \frac{TV}{V_{ir}}, \]  

(2)

where \( TV_{ir} \) represents the PTV_{GT} which is within the reference isodose, \( TV \) is the target volume (PTV_{GT}) and \( V_{ir} \) is the volume of the reference isodose (here the 95% isodose). The first factor represents the target volume coverage \( (V_{95}) \), whereas the second factor represents the volume of normal tissue irradiated by the reference isodose.

The differences between the segmentation algorithms and the different measures of ground-truth volume coverage \( (V_{95}, D_{95}, HI \text{ and } CI) \) were assessed with the Kruskal–Wallis (K-W) test which is an extension of the Wilcoxon rank-sum test for three or more groups and non-paired data (we considered here five groups, FLAB, FCM, T42, A1 and A2). It allows comparison of parameters with small samples and without a Gaussian assumption which is the case here. Two statistical tests were conducted, the first on all data together and the second by differentiating homogeneous from heterogeneous uptake cases.

3. Results

3.1. Delineation performance

All segmentation results were compared to the ground truth to evaluate the delineation accuracy of the different algorithms considered. Figure 3 provided VE, sensitivity and PPV with respect to the ground truth for the different segmentation algorithms considered. Figure 3(a) shows the mean VE (which can be negative or positive) and associated standard deviation (SD). Figure 3(b) provides the mean sensitivity and PPV and their associated SD over all cases. Overall, the advanced image-segmentation approaches demonstrated higher accuracy, with a mean VE of \(-2 \pm 11\% \) and \(12 \pm 37\% \), a mean sensitivity of \(0.86 \pm 0.06 \) and \( 0.88 \pm 0.05 \) and a mean PPV of \(0.87 \pm 0.06 \) and \( 0.83 \pm 0.15 \) for FLAB and FCM, respectively. In comparison, the threshold-based methods resulted in a mean VE of \(2 \pm 107.0\% \), \(-35 \pm 27.0\% \) and \(-31 \pm 26.0\% \), a mean sensitivity of \(0.66 \pm 0.26 \), \(0.61 \pm 0.24 \) and \(0.64 \pm 0.22 \) and a mean PPV of \(0.89 \pm 0.23 \), \(0.96 \pm 0.06 \) and \(0.96 \pm 0.06 \) for T42, A1 and A2, respectively. The mean VE of T42 was quite low \((-2\% \); however, the associated SD was the highest \((107\%) \). The two adaptive thresholds led to quite similar results with underestimated volumes due to uptake heterogeneities, therefore leading to higher PPV but much smaller sensitivity.

Figures 3(c)–(f) illustrate the same segmentation results with data separated into homogeneous (H\&N cases 1 and 2b, lung case 3) and heterogeneous (H\&N cases 2a and 3, lung cases 1 and 2) cases. For homogeneous tumours, FLAB, A1 and A2 led to similar results (mean VE ~\(-13\% \), mean sensitivity and PPV of ~0.85). On the other hand, FCM and T42 overestimated tumour volume with a positive VE (18 ± 56\% and 70 ± 151\%, respectively), and sensitivity (0.88 ± 0.04 and 0.86 ± 0.12, respectively) higher than the PPV (0.83 ± 0.23 and 0.73 ± 0.32, respectively). For heterogeneous cases, threshold-based
methods underestimated the volume (mean VE, sensitivity and PPV of $-45\%$, $0.5$ and $0.98$, respectively), whereas advanced methods were able to handle these heterogeneities (mean VE of $2.3 \pm 10.7\%$ and $9.7 \pm 21.2\%$ for FLAB and FCM, respectively, and mean PPV and sensitivity of $\sim 0.85$).

3.2. Tumour volume coverage

For the lung cases 1a and 1b, PTV_{FLAB} and PTV_{FCM} could not receive 95% of the prescribed dose but GTVs did ($V_{95}$ of 89% (89%) and 83% (85%) for case 1a (1b) for PTV_{FLAB} and PTV_{FCM}, respectively). For lung case 1c none of the PTVs defined by any delineation method considered could receive 95% of the prescribed dose ($V_{95}$ of 89%, 80%, 94%, 92% and 92% for PTV_{FLAB}, PTV_{FCM}, PTV_{T22}, PTV_{A1} and PTV_{A2}, respectively) but the GTVs did. For all other cases (lung cases 2 and 3, H&N cases), PTVs received 95% of the prescribed dose. No planning was produced for the fixed threshold volumes for H&N cases 1b and 3b since they
(1) Tumour volume definition

(2) Isodose lines

(3) Dose Volume Histogram

Figure 4. Illustration of the complete procedure and results for lung case 1a: (1) target volume definition, (2) isodoses for the four different plans and (3) DVH for the four plans with dose on PTV_{GT}.

grossly overestimated the ground-truth volume by +333% and +58%, respectively. Similarly, no planning was produced for FCM in H&N case 1b (+118% overestimation relative to the ground-truth volume).

Figure 4 shows the whole procedure for one heterogeneous lung case (case 1a). The different GTVs, resulting isodoses and dose volume histograms (DVH) obtained by the four segmentation methods are illustrated. Advanced methods resulted in the largest PTV in this case, consequently leading to larger 95% isodoses ($V_p$ of 141.6 cm$^3$ and 150.0 cm$^3$ for FLAB...
and FCM, respectively) compared to the threshold-based methods ($V_a$ of 89.7, 80.3 and 95.8 cm$^3$ for T42, A1 and A2, respectively) and better PTV$_{CT}$ volume coverage ($V_{95}$ of 84.7%, 83.9%, 60.7%, 53.7% and 63.6% for FLAB, FCM, T42, A1 and A2, respectively). In addition, they also resulted in higher doses delivered to OARs ($D_{20}$ and $D_{35}$ for the lungs of 26.5, 26.5, 15.6 and 16.6 Gy, respectively) compared to the threshold-based methods ($D_{20}$ ($D_{35}$) to the lungs of 21.1 Gy (10.5 Gy), 19.1 Gy (7.6 Gy) and 20.1 Gy (8.3 Gy) for T42, A1 and A2, respectively). All these doses were however within the OARs constraints (30 and 20 Gy for $D_{20}$ and $D_{35}$, respectively). Values ($V_a$, $V_{95}$, $D_{20}$ and $D_{35}$) for this case are reported in table 3.

Figure 5(a) shows the mean and SD over all cases for $V_{95}$ and $D_{95}$, (b) HI and (c) CI.
Table 3. Volume of reference isodose ($V_{ir}$), $V_{95}$ and $D_{20}$ and $D_{35}$ for lungs for lung case 1a (the case illustrated in figure 4).

<table>
<thead>
<tr>
<th>Method</th>
<th>$V_{ir}$ (cm$^3$)</th>
<th>$V_{95}$ (%)</th>
<th>$D_{20}$</th>
<th>$D_{35}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAB</td>
<td>141.6</td>
<td>84.7</td>
<td>26.5</td>
<td>15.6</td>
</tr>
<tr>
<td>Fixed threshold</td>
<td>89.7</td>
<td>60.7</td>
<td>21.1</td>
<td>10.5</td>
</tr>
<tr>
<td>Adaptive threshold 1</td>
<td>80.3</td>
<td>53.7</td>
<td>19.1</td>
<td>7.6</td>
</tr>
<tr>
<td>Adaptive threshold 2</td>
<td>95.8</td>
<td>63.6</td>
<td>20.1</td>
<td>8.3</td>
</tr>
<tr>
<td>FCM</td>
<td>150.0</td>
<td>83.9</td>
<td>26.5</td>
<td>16.6</td>
</tr>
</tbody>
</table>

Table 4. Doses to the OARs for the lung cases.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean $D_{\text{max}}$ to spinal cord</th>
<th>Mean $D_{20}$ to both lungs</th>
<th>Mean $D_{35}$ to both lungs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAB</td>
<td>35.4 (25.5–41.9)</td>
<td>23.3 (17.8–28.0)</td>
<td>12.6 (6.8–17.1)</td>
</tr>
<tr>
<td>FCM</td>
<td>34.9 (23.4–40.9)</td>
<td>24.1 (18.5–29.1)</td>
<td>14.1 (10.0–18.1)</td>
</tr>
<tr>
<td>Fixed threshold</td>
<td>33.2 (24.6–38.8)</td>
<td>19.7 (12.6–25.1)</td>
<td>10.0 (6.5–15.1)</td>
</tr>
<tr>
<td>Adaptive threshold 1</td>
<td>32.9 (24.6–38.6)</td>
<td>19.5 (12.9–25.5)</td>
<td>9.7 (6.5–15.5)</td>
</tr>
<tr>
<td>Adaptive threshold 2</td>
<td>33.1 (24.7–39.2)</td>
<td>20.0 (13.1–25.8)</td>
<td>10.0 (6.6–16.0)</td>
</tr>
</tbody>
</table>

91.6% ± 6.3% and 90.8% ± 7.0%, respectively, mean $D_{95}$ of 91.6% ± 5.9% and 89.0% ± 9.0%, respectively than the threshold-based methods (mean $V_{95}$ and $D_{95}$ below 82.3 ± 15.0% and 79.3 ± 21.1%, respectively). Figure 5(b) shows the mean HI. This index was lower for FLAB and FCM (17.9 ± 8.3 and 23.0 ± 14.1) than for the threshold-based methods (mean >29.8 ± 24.7). Figure 5(c) shows the CI mean and associated SD. No significant differences were observed between the various delineations strategies with 0.68 ± 0.11, 0.66 ± 0.08, 0.63 ± 0.08, 0.63 ± 0.09 and 0.64 ± 0.08 for FLAB, FCM, T42, A1 and A2, respectively. Differences between delineation strategies were not significant ($p > 0.05$) for homogeneous activity distributions within tumours (H&N cases 1 and 2b) and the necrotic case (lung case 3), but were significant ($p < 0.02$) in terms of $V_{95}$ and $D_{95}$ for the heterogeneous cases (see figures 6(a) and (b)). For all heterogeneous cases, mean $V_{95}$ was 89.1 ± 6.4% and 88.5 ± 6.9% for FLAB and FCM, respectively, whereas for threshold-based methods significantly ($p < 0.05$) lower values and higher SDs were observed (73.0 ± 15.4%, 72.7 ± 16.4% and 75.4 ± 13.8% for T42, A1 and A2, respectively). By comparison, mean $V_{95}$ for homogeneous cases was globally higher (>95%) with lower SD (<2%) independently of the delineation strategy. Similar conclusions can be drawn regarding $D_{95}$ and HI. When considering heterogeneous cases, mean $D_{95}$ was 89.4 ± 6.1% and 86.3 ± 9.3% for FLAB and FCM, respectively, whereas it significantly ($p < 0.05$) dropped to 71.0 ± 19.8% for T42, and 69.1 ± 24.2% and 71.0 ± 21.5% for A1 and A2, respectively. HI associated with the FLAB and FCM delineations was 21.3 ± 8.2% and 27.9 ± 13.4%, respectively, rising to 38.8 ± 21.8%, 41.0 ± 26.0% and 39.6 ± 25.0% for T42, A1 and A2, respectively ($p > 0.05$). On the other hand, for the homogeneous cases the mean $D_{95}$ and HI were not significantly different ($p > 0.05$) across the various delineation strategies, with $D_{95} > 95%$ and HI < 13%.

3.3. OAR sparing

The highest delivered dose to the spinal cord over all the lung cases and segmentation algorithms was 41.9 Gy. The maximum dose to 100% of the heart over all cases was 0.58 Gy. Table 4 provides the mean and ranges (min and max) over all cases of the delivered doses to the spinal cord and the lungs.
Figure 6. K-W results on the heterogeneous group for (a) $V_{95}$ and (b) $D_{95}$.

For the H&N cases the maximum dose to the spinal cord and the brain stem was 49.7 and 42.8 Gy, respectively. The highest dose to the spinal cord over all the H&N cases was inferior to 50 Gy (value reached for case 2). Table 5 contains the mean and ranges (min and max) over all cases of the delivered doses to the spinal cord and the parotids.

### 3.4. Clinical case

For the clinical case no ground truth was available. Two different GTVs were obtained using the two most accurate, established on the simulated datasets, amongst each group of methods ($\text{GTV}_{\text{FLAB}}$ and $\text{GTV}_A$ for adaptive thresholding). $\text{GTV}_{\text{FLAB}}$ and $\text{GTV}_A$ were obtained by combining the delineations performed separately on the large high contrast and heterogeneous uptake on the one hand, and the small, lower contrast several uptakes on the other hand (see...
Table 5. Doses to the OARs for the H&N cases.

<table>
<thead>
<tr>
<th></th>
<th>Mean $D_{max}$ to spinal cord</th>
<th>Mean $D_{20}$ to parotids Left/Right</th>
<th>Mean $D_{40}$ to parotids Left/Right</th>
<th>Mean $D_{60}$ to parotids Left/Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRAM</td>
<td>42.4 (37.2–49.7)</td>
<td>37.7 (34.0–46.7)</td>
<td>30.0 (19.7–44.7)</td>
<td>21.9 (19.2–28.4)</td>
</tr>
<tr>
<td>FCM</td>
<td>43.8 (37.9–49.1)</td>
<td>37.6 (33.7–46.5)</td>
<td>29.7 (10.0–44.6)</td>
<td>22.2 (19.9–28.4)</td>
</tr>
<tr>
<td>Fixed threshold</td>
<td>42.5 (37.7–47.7)</td>
<td>37.0 (33.2–45.9)</td>
<td>29.5 (20.0–44.4)</td>
<td>21.7 (18.9–28.1)</td>
</tr>
<tr>
<td>Adaptive threshold 1</td>
<td>41.6 (37.5–47.7)</td>
<td>36.9 (33.4–46.2)</td>
<td>30.3 (19.9–44.3)</td>
<td>21.7 (19.7–27.9)</td>
</tr>
<tr>
<td>Adaptive threshold 2</td>
<td>41.7 (37.4–47.7)</td>
<td>36.8 (33.2–46.2)</td>
<td>30.5 (19.7–44.6)</td>
<td>21.6 (19.6–28.1)</td>
</tr>
</tbody>
</table>
yellow ROIs in figure 7). The resulting PTV volumes were 185 and 127 cm$^3$. $V_{95}$ was 99.5% and 99.9% PTV$\text{FLAB}$ and PTV$\text{A}$, respectively. The doses to parotids were equivalent for both delineations ($D_{20}$, $D_{40}$ and $D_{60}$ for the left parotid were 26, 16 and 11 Gy for FLAB and 26, 15 and 10 Gy for adaptive thresholding). The maximum dose to spinal cord was 44 and 42.8 Gy for FLAB and adaptive threshold, respectively.

4. Discussion

Several delineation approaches have been proposed in the past few years for PET uptake volume delineation on PET images. One of the objectives is to offer improved target volume delineation for radiotherapy planning in order to facilitate the incorporation of the PET information into radiotherapy treatment. Within this context, the objective of this study was to investigate the impact of the actual accuracy of such approaches on RTP in terms of dosimetry. The data used in this work were simulated PET images which allowed knowledge and control of the tumour position, size, shape and activity distribution. Four PET image-segmentation algorithms were considered and grouped into threshold-based (fixed and adaptive) and automatic methods (FLAB and FCM). The segmentation accuracy was assessed with respect to the known ground truth. For each of the obtained delineations, an IMRT plan was subsequently designed and all the plans were compared in terms of ground-truth volume coverage and OARs sparing using standard metrics.

The most significant differences in segmentation accuracy were observed for tumours exhibiting heterogeneous uptake for which the automatic approaches were able to delineate...
the entire volume, whereas the threshold-based algorithms usually significantly underestimated such volumes (high PPV and low sensitivity). The main differences between fixed and adaptive threshold methods were obtained for the lowest contrast cases in which adaptive thresholding was more accurate than the fixed threshold which highly overestimated the volume in these cases (+333% and +58% for H&N cases 1b and 2b, respectively). Similarly, FLAB provided more accurate results than FCM on these cases. The low mean VE and its high associated SD for fixed threshold can be explained by the fact that heterogeneous cases resulted in underestimation of the volume, whereas low contrast cases resulted in overestimation of the volume.

For the measures assessing the quality of the ground-truth volume coverage, significant differences in $V_{95}$ and $D_{95}$ were observed between the automatic approaches (FLAB and FCM) and threshold-based (T42, A1 and A2) but not within each of the two groups. Larger SDs were observed for the threshold-based methods compared to the automatic approaches. This can be explained by the fact that these approaches had equivalent performance for both heterogeneous and homogeneous cases, whereas threshold-based methods consistently failed in delineating heterogeneous uptakes, which resulted in insufficient volume coverage. No significant differences were observed for the CI between all the algorithms in the heterogeneous group ($p = 0.24$). This can be explained by the fact that two factors affect this index: $V_{95}$ (TV$_V$/TV) and (TV$_N$/V$_N$) which is a measure of how much normal tissue is irradiated by the reference dose (here 95% of the prescribed dose). For highly underestimated tumour volumes, $V_{95}$ is consequently low but only a few parts of normal tissue are irradiated, a factor that improves the CI.

For lung cases 1 and 2, as the threshold-based method only delineated the sub-volumes with higher uptake (see figure 4 for case 1a), the larger this sub-volume (expressed as a percentage of the overall tumour volume), the better the volume coverage when considering threshold-based methods. Lung case 1 for example was simulated with three heterogeneous sub-volumes sizes (12%, 41% and 53% of the overall tumour volume). The corresponding $D_{95}$ was 29.6%, 64.2% and 66.8% for A2. Similar results were observed for the second lung case with sub-volumes of 12% and 38% of the overall tumour volume, and corresponding $D_{95}$ values of 61.6% and 69.5%, respectively.

For the clinical case, PTV$_{FLAB}$ was larger than PTV$_A$. The PET uptake was indeed slightly heterogeneous and exhibited different levels of uptake (see the illustration in figure 7). Similar differences between the two approaches were therefore observed as in the simulated datasets. Similarly, due to higher tumour coverage, the resulting dose to spinal cord was higher for FLAB than for adaptive thresholding ($D_{max}$ of 44 and 42.8 Gy, respectively). However, both were inferior to the constraint of 45 Gy.

Our results demonstrate that there might be a significant impact on the dosimetry of IMRT plans including the PET uptake within the tumour volume and the method used to delineate this uptake. There is therefore a need for accurate and robust automatic PET heterogeneous uptake delineation in order to incorporate functional information into radiotherapy planning, especially for heterogeneous uptake tracer distributions within the tumour target volumes as well as for low contrast cases. In this work, we used FCM and FLAB; however, several other recent methods have been developed and validated against such heterogeneous or low contrast PET uptakes (Zaidi and El Naqua 2011, Hatt et al 2012) and should therefore lead to similar dosimetry results.

A limitation of the current study is the lack of respiratory motion concerning the lung cases and the lack of setup errors in all of the cases considered. However, one well-recognized result of respiratory motion in PET imaging is an overall tumour contrast reduction and as demonstrated in this study, the use of segmentation algorithms able to accurately handle low
contrast lesions can only be advantageous for dosimetry purposes. Another limitation is that our study was restricted to PET-based GTV. A future extension of the proposed framework introduced here could be the addition of anatomical/morphological imaging such as CT or MRI in the GTV delineation, by comparing results of multimodality image-segmentation approaches dedicated to multimodal treatment planning in radiotherapy (Han et al 2011) with respect to a more complete ground truth including anatomical and functional tumour volumes.

5. Conclusions

A framework was proposed to evaluate the impact of the accuracy of PET uptake volume delineation on the dosimetry of radiotherapy treatment plans. Simulated PET images and their corresponding ground truth were imported into the TPS in order to evaluate the impact of the accuracy of the different delineations on the dosimetry. The accuracy of segmentation was assessed by VEs, sensitivity and PPV with respect to the ground truth of the simulation. The corresponding quality of the treatment plans was evaluated using the same ground-truth volume and several measures ($V_{95}, D_{95}, HI$ and $CI$). Automatic advanced methods demonstrated better accuracy than threshold-based methods especially for heterogeneous tracer uptake resulting in significantly better target volume coverage (mean $V_{95}$ of 91.6%) than threshold-based methods (mean $V_{95}$ below 82.3%). On the other hand, for more homogeneous tracer uptake distribution, no significant differences were observed in terms of dosimetry between the delineation strategies. As expected, an underestimation of the true tumour uptake volume resulted in insufficient target volume coverage but in better OAR sparing as assessed by dose constraints, whereas an over-estimation of the ground-truth volume resulted in better coverage but lower OAR sparing, although still within the dose limits.

In conclusion, although for homogeneous PET uptakes, simple threshold-based methods may be sufficient to provide accurate PET GTV delineation for treatment planning, our study demonstrates that significant dosimetry errors can be avoided by using more advanced image-segmentation methods, especially when considering heterogeneous uptake volumes.

References

Bentzen S M 2005 Theragnostic imaging for radiation oncology: dose-painting by numbers Lancet Oncol. 6 112–7

Han D, Bayouth J E, Song E, Taurani A, Sonka M, Buatti J B and Wu X 2011 Globally optimal tumor segmentation in PET-CT images: a graph-based co-segmentation method Proc. of the 22nd Int. Conf. on Information Processing in Medical Imaging (Kloster Irsee, Germany) (Lecture Notes in Computer Science 6801) (Springer) pp 245–56

Hatt M, Boussion N, Cheze-Le Rest C, Visvikis D and Pradier O 2012 Metabolically active volumes automatic delineation methodologies in PET imaging: review and perspectives Cancer Radiother. 16 70–81


Soret M, Bacharach S L and Buvat I 2007 Partial-volume effect in PET tumor imaging J. Nucl. Med. 48 932–45


