Multi-observation PET image analysis for patient follow-up quantitation and therapy assessment.

Simon David, Dimitris Visvikis, Christian Roux, Mathieu Hatt

Multi-observation PET image analysis for patient follow-up quantitation and therapy assessment

Simon David , Dimitris Visvikis , Christian Roux , Mathieu Hatt *

LATIM, Laboratoire de Traitement de l'Information Medicale INSERM : U650 , Universit*é *de Bretagne Occidentale - Brest , Institut T*él*écom , T*él*écom Bretagne , CHU Brest , Universit*é *europ*é*enne de Bretagne , FR

* Correspondence should be adressed to: Mathieu Hatt <hatt@univ-brest.fr >

Abstract

In Positron Emission Tomography (PET) imaging, an early therapeutic response is usually characterized by variations of semi-quantitative parameters restricted to maximum SUV measured in PET scans during the treatment. Such measurements do not reflect overall tumour volume and radiotracer uptake variations. The proposed approach is based on multi-observation image analysis for merging several PET acquisitions to assess tumour metabolic volume and uptake variations. The fusion algorithm is based on iterative estimation using stochastic expectation maximization (SEM) algorithm. The proposed method was applied to simulated and clinical follow-up PET images. We compared the multi-observation fusion performance to threshold-based methods, proposed for the assessment of the therapeutic response based on functional volumes. On simulated datasets, the adaptive threshold applied independently on both images led to higher errors than the ASEM fusion and on the clinical datasets, it failed to provide coherent measurements for four patients out of seven due to aberrant delineations. The ASEM method demonstrated improved and more robust estimation of the evaluation leading to more pertinent measurements. Future work will consist in extending the methodology and applying it to clinical multi-tracers datasets in order to evaluate its potential impact on the biological tumour volume definition for radiotherapy applications.

MESH Keywords Algorithms ; Female ; Follow-Up Studies ; Humans ; Male ; Neoplasms ; pathology ; radionuclide imaging ; radiotherapy ; Pattern Recognition, Automated ; methods ; statistics & numerical data ; Positron-Emission Tomography ; instrumentation ; methods ; statistics & numerical data ; Radiopharmaceuticals ; diagnostic use ; Reproducibility of Results ; Sensitivity and Specificity ; Stochastic Processes

Author Keywords Bayesian classification ; image fusion ; oncology ; patient monitoring ; PET ; therapeutic response ; unsupervised segmentation.

Introduction

Positron Emission Tomography (PET) is now a widely used tool in the field of oncology, especially in applications such as diagnosis, patient follow-up studies (Krak *et al* 2005) or in radiotherapy planning (Jarritt *et al* 2006). In the context of patient follow-up, early metabolic changes detected with 2′-deoxy-2′-[^{18} F]-fluoro-D-glucose (FDG) PET imaging can occur before anatomic changes observed with computed tomography (CT) imaging. By assessing differences in several PET scans acquired before and at different times during treatment, various qualitative and quantitative methods have been proposed to characterize the therapeutic response (Weber *et al* 2007). In patient monitoring studies, qualitative methods such as visual assessment are less accurate and reproducible than quantitative measurements (Lin *et al* 2007). Furthermore, different therapeutic parameters (indexes) have been defined either on dynamic or static PET acquisitions with a similar reproducibility (Weber *et al* 1999). Being less restrictive in clinical routine, only the parameters computed in the static PET scans have been considered in our work. Most widely used in patient follow-up studies, the standardized uptake value (SUV) measures the tracer uptake in the tumour. Derived from the SUV index, two measurements, namely the maximum SUV (SUV_{max}) and the mean SUV (SUV_{mean}) were assessed in our study by computing respectively the maximum and the mean of SUV in voxels included in a region of interest defining the tumour. The reproducibility and the robustness of both SUV indexes have been previously assessed (Weber *et al* 2007 , Nahmias *et al* 2008) and compared to the reproducibility of tumour volume measurements with various automated methodologies (Hatt *et al* 2010). An early therapeutic response can be characterized by measuring relative or absolute SUV variations between pre-treatment and mid-treatment PET scans. Other quantitative parameters have been used such as the total lesion glycolysis (TLG) defined as the product of the mean SUV and the tumour volume (Larson *et al* 1999 , Hatt *et al* 2010).

The therapeutic response is usually estimated by measuring the tumour size on the CT scans, and according to guidelines such as WHO and RECIST (Therasse *et al* 2000). More recent criteria have been proposed such as PERCIST (Wahl *et al* 2009), adding the consideration of quantitative parameters extracted from PET images. However, these criteria are still limited to simple SUV measurements and do not include volumetric characterization of the tumours, and no guidelines have been established recommending the best way to characterize the therapy response according to the variation of metabolically active tumor volumes. In the current clinical practice, the therapeutic response is therefore usually assessed by considering one single value as the SUV_{max} within the primary lesion, extracted from each PET scan. Presently, the measure of SUV_{max} variation is considered as the gold standard of the treatment response definition. This method however accounts neither for the tumour volume variations nor the spatial uptake variation within the tumour volume.

Among the new methodologies developed in PET tumour delineation (Zaidi *et al* 2010), most of them have only considered the use of such delineation for static images segmentation and for diagnosis/prognosis. A few authors have recently proposed different methodologies dedicated to PET follow up, like the one by Necib *et al* in 2008 , which is aimed at assessing a response by comparing two follow-up PET images. After voxel-to-voxel registration of the two scans, a biparametric map is generated representing the tracer uptake variations within the tumour. In the context of cancer treatment prediction, Naqa *et al* have recently proposed a texture-based approach, considering texture properties of voxels within tumours as prognosis factors for the assessment of therapy response.

Regarding the use of PET in radiotherapy, the gross tumor volume (GTV) definition is usually carried out manually on fused FDG-PET/CT scans. However, imaging tumour's glucose consumption with the FDG alone may not be sufficient to determine the GTV (Mankoff *et al* 2003). Considering the measure of other features of cancer metabolism like proliferation, hypoxia and apoptosis using additional tracers may generate more complete information regarding the target tumour volume (Bentzen *et al* 2005 , Shields *et al* 2003 , Vaupel *et al* 2007). Accurate tumour volume delineation would therefore require a fusion of all available measurements obtained with these different tracers. Such a fusion could be valuable to thoroughly and potentially more accurately assess tumour volume definition as well as evolution during therapy.

The main objective of this study was to develop a fusion method derived from multi-observation approaches such as these developed in satellite and astronomical imaging (Masson *et al* 1993). Considering either patient follow-up and/or multi-tracer PET datasets, the proposed method aims at assessing a treatment response and tumour volume definition by automatically determining the different variations of tracer uptakes within the regions of the analyzed fused images. Our approach is statistical and assumes that the data can be modeled by a mixture distribution of multi observation random fields. The parameters defining the mixture distribution are estimated using a stochastic expectation maximization (SEM) algorithm (Celeux *et al* 1986) combined with a locally adaptive spatial priors estimation in order to account for voxels correlation.

Our method was applied to simulated and clinical pre and post treatment PET scans of esophageal cancer within the context of radio-chemotherapy follow-up. It was compared to current quantitative methods proposed for the assessment of the therapeutic response based on tumor volume evolution, namely the definition of the tumor volumes independently on both scans using adaptive thresholding.

Materials and methods

Multi-observation framework

The proposed method could potentially be applied for both patient follow-up applications and multi-tracer analysis using PET scans. The proposed method is aimed at merging the available PET images in order to derive a fusion of the information regarding the treatment response in patient follow-up application or/and the multi-tracer tumour volume, as illustrated in figure 1 . While the analysis for both applications might require different fusion rules or interpretation, the basics of the approach are the same and are based on the unsupervised Bayesian methods, widely used in segmentation and classification of satellite, astronomical or medical imaging (Masson *et al* 1993 , Pieczynski *et al* 2003 , Hatt *et al* 2009).

Bayesian model

Let T be a finite set corresponding to the voxels of 3D registered PET images. We consider two random processes $\mathbf{Y} = (\boldsymbol{y_t})_{t \leq T}$ and $\mathbf{X} = (x_t)_{t \leq T}$. **Y** models the observed multi-tracers or follow-up PET scans, acquired at different times during the treatment, and takes its values in $\mathbb{R}$. Each $\boldsymbol{y_t}$ is therefore a vector of real values defined as $\boldsymbol{y_t} = (y_t^{(1)}, \ldots, y_t^{(B)})$, containing the voxel values of each PET image, with B being the image number observed in the fusion. Each $\boldsymbol{y_t}$ is associated with a label x_t . **X** models the fusion map which is designated in our specific application as the therapeutic response classification. **X** takes its values in a set {1 … K }, with K being the number of classes that is usually user-dependent and defined depending on the fusion goal. The objective of the approach is therefore to estimate the distribution of (**X, Y**). Considering the Bayesian framework, the relationship between **X** and **Y** can be modeled using the joint probability:

$$p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{Y} \mid \mathbf{X}) \times p(\mathbf{X}) \quad (1)$$

where, p (**X**) is the prior knowledge about **X** and p (**Y** |**X**) is the "noise model": the likelihood of the observation **Y** conditionally to the hidden ground-truth **X** . In this Bayesian framework, the prior knowledge p (**X**) can be modeled globally, for instance by considering Markovian models (Pieczynski *et al* 2003) such as chains or trees, or locally with blind, contextual or adaptive models (Peng *et al* 1995). Different noise distributions can be used for the observation model p (**Y** |**X**) such as Gaussian or generalized Gaussian (Delignon *et al* 1997).

In our fusion method, we have used a locally adaptive prior model and the noise model has been assumed Gaussian, however other distributions could be considered in future developments. The distribution of (**X, Y**) is hence defined by the priors Π_k , the mean vectors μ_k and covariance matrices Γ_k associated to each of the K classes in the mixture. The mean parameter μ_k is a vector associated to the B images

of the fusion $\boldsymbol{\mu_k} = (\mu_k^{(1)}, \ldots, \mu_k^{(B)})$. The SEM algorithm was used here to estimate the parameters of the distribution of (**X, Y**). It is a stochastic version of the classic EM algorithm, ensuring better and faster convergence as well as higher independence on the initialization. From here onwards, our approach will be denoted ASEM.

Fusion process

Pre-processing: image deconvolution

PET images are characterized by their high level of noise and the limited spatial resolution inducing partial volume effect (PVE) (Soret *et al* 2007). The under estimation of the tissues uptakes and activity cross contamination between structures with different uptakes are two consequences of the PVE effects in PET images. When considering several co-registered PET images, the voxels most affected by PVE may not be on the same coordinates for each scan, which implies intensities distributions that might complexify both estimation and classification steps in the fusion process. In addition, without PVE correction (PVC), SUV values extracted from each scan may be significantly biased. This can lead to under or over estimation of the uptake variation between pre and post treatment scans, especially if significant tumor volume variation occurs. Indeed, PVE impact on the SUV measurement within the tumor strongly depends on the object' s size. In order to reduce the impact of these effects on the subsequent steps, a PVE correction (PVC) was applied to each image prior to their fusion. The chosen PVC method was developed by Boussion *et al* in 2008 and further improved by Le Pogam *et al* in 2009 and consists of a 3D voxel-wise correction using an iterative deconvolution improved by wavelet based optimal denoising of the residual. This preprocessing step offers two advantages: first, it reduces the size of blurred frontiers between the different regions of the images, hence reducing their impact on subsequent registration and fusion complexity. Second, it allows the extraction of corrected uptake values from fusion maps for a quantitative characterization of evolution of the activity within tumour volume and/or sub-volumes.

Local-based analysis

As our goal is to automatically determine the variation of tracer activity and position/volume of a functional tumour, we assume that the overall tumour volume has been previously automatically or manually isolated in a 3D box or volume of interest (VOI) by a clinician on the co-registered PET images. Therefore the box should be large enough to encompass the entire tumour in each scan and avoid including too many neighboring tissues with significant physiological uptake. Consequently, the definition of such a processing box should allow any shape and size, in 3D. The definition of this 3D VOI should therefore be carried out on the scan in which the tumour appears to be the largest, and automatically registered on the other volumes involved in the fusion as illustrated in figure 2 .

FKM Initialization and choice of the number of classes

In unsupervised Bayesian segmentation framework, the initialization is an important step. In our method, we used the Fuzzy K-Means (FKM) algorithm (Krishnapuram *et al* 1994) based on fuzzy logic applied to the voxels values. In his PhD thesis, Provost described an improved version of FKM, allowing the automatic estimation of the optimal number of classes in the mixture, based on the use of an entropy criterion and a user selection of the upper limit of the number of classes. This upper limit was defined as the product of the number of images and the number of classes within each image considered in the fusion. In each iteration and for all the voxels of the image, a membership coefficient associated to the *K* classes of the mixture is estimated. At the end of FKM execution, the mixture parameters associated with each class *k* (Π_k , μ_k , Γ_k) are initialized for the Bayesian estimation. The cost function of the FKM algorithm is defined as:

$$E = \sum_{k=1}^{K}\sum_{t=1}^{T}(\psi_{kt})^{q}\|x_t - c_k\|^2 + \alpha\sum_{k=1}^{K}p_k^2$$

(2)

with

$$\sum_{k=1}^{K}\psi_{kt} = 1, \forall\ t \text{ and } p_k = \sum_{t=1}^{T}\psi_{kt}$$

(3)

where *E* is the cost function to minimize, *K* is the number of classes in the mixture, *T* is the number of voxels in the volume, ψ_{kt} is the membership matrix of each voxel t to the k^{th} class, c_k is the centroid of the k^{th} class. The second term of the equation (2) is the entropic criterion where α is the parameter weighting the cost function. The minimization of the FKM cost function is performed by constraining the entropic term. This latter is high in the first iterations, leading to a reduction of the number of classes in the mixture, and it decreases exponentially in order to allow the FKM classification process.

Parameters estimation

The parameters (Π_k , μ_k , Γ_k) defining the Gaussian mixture of the (**X, Y**) distribution are estimated by the SEM algorithm by sampling several realizations of **X** according to its posterior distribution *p* (**X**| **Y**). In the adaptive framework, the global prior Π_k , associated to the k^{th} class of the mixture are re-estimated using a local neighboring 3D cube and replaced by local priors $\pi_{t,k}$ defined for each voxel and each

class. The mean vector μ_k and the covariance matrix Γ_k are finally computed for each of the *K* classes in the mixture. The details of parameters estimation with the SEM algorithm are given in the appendix section.

Decision step

In order to perform fusion on a voxel-by-voxel basis, we used a classification criterion to assign a class to each voxel. For this purpose we chose the *maximum likelihood* method. To compute a solution, this criterion requires the parameters defining the *a priori* model (priors of each class and for each voxel) as well as the observation data model (mean and covariance matrices of each class), previously estimated using the SEM algorithm (see appendix).

Simulated datasets

In order to evaluate the behavior of the fusion approach within the context of tumor evolution assessment, we considered realistic simulations of non spherical tumours. These simulated tumours were created using as models real head and neck and esophageal observed in clinical datasets. The procedure for the simulations of such data has been previously described (Le Maitre *et al* 2009). The simulated cases used in our study are composed of two simulated PET scans, one before and one after the treatment. The clinical cases used as models corresponded to patients classified as partial responders or progressive disease to the radiochemotherapy according to RECIST criteria.

In order to evaluate the robustness of the methods, three levels of noise were considered for every simulated acquisition by selecting 100, 80 and 60% of the simulated lines of response for the iterative reconstruction respectively. With 20 clinical follow-up cases and considering three levels of noise and various tumor-to-background ratios for each case, 70 different simulated cases were generated. Most of the simulated datasets, representing 15 out of the 20 cases, were generated from patients classified as partial responders. The others cases were designed to simulate progressive disease. The mean tumour volume and tumour to background ratio used in the first and second simulated follow-up cases are given in table 1 . Three examples of these simulated follow-up cases (showing only central axial slice) with their associated ground-truth are illustrated in figures 3 . Similarly to the clinical datasets they are based on, these tumours are characterized by either homogenous or heterogeneous tracer uptake. Background activity was simulated as homogeneous. The voxels were assumed to belong either to the background (BD) or to the tumour (T). In these simulated images, no registration was required.

Clinical datasets

Patient data

After preliminary studies on simulated follow-up PET scans, the fusion method was tested on real clinical datasets. Seven patients with esophageal cancer undergoing concomitant radiochemotherapy between 2005 and 2008 were considered, with one PET scan before treatment, and another PET scan after treatment, both acquisitions carried out according to the same protocol. All these patients were classified as partial responders one month after the completion of treatment, according to RECIST. Consequently, the variation of metabolically active tumor volume as well as the SUV within the tumor volume is expected to be less than 100% since residual tumor uptake is seen for all these patients, and also above 20 to 30% which is their reproducibility limits as previously demonstrated (Hatt et al, 2010). No volume or uptake increase should be measured in these cases. Visual illustrations of three clinical follow-up cases are given in figures 5 . The physiological uptake of the mediastinum around the tumour volume was much more significant in the clinical images than in simulated cased (for which background was simulated as homogeneous) and was therefore taken into account in the fusion process as an additional class. Thus, on each scan, the voxels were assumed to belong to the background (BD) of the lungs, the physiological tissues (PHY) of the mediastinum, or the tumour volume (T).

PET/PET registration

In the context of patient follow-up, the PET/CT images are acquired at several month intervals. As an important pre requisite of the proposed method which works on a voxel level, the scans must therefore be registered before the fusion method can be applied. The PET/PET registration was carried out using a method previously proposed (Ouksili *et al* 2007), in which the PET data are first registered with their associated CT scans, acquired in the same bed position. Having more landmarks and a higher resolution, the CT scans are registered using the MIPAV software. The CT/CT registration was carried out using a rigid transformation, which optimizes the least square criterion of a large volume of interest. The CT/CT transformation matrix was then applied to the PET scans for registration. A rigid registration was considered to be sufficient since the procedure was carried out only on small 3D regions of interest surrounding the lesions which were located in the mediastinum or head and neck regions. In addition, a rigid transformation avoids the deformation of tumor volumes in the PET images which would be certainly associated with the use of a deformable model. Finally the consideration of head and neck and mediastinum lesions reduces the potential influence of respiratory motion on the registration process.

In the fusion process, the use of a sliding estimation cube as described in *2.2* for the computation of the *spatial priors* is expected to reduce the impact of small registration errors of the PET datasets. The impact of the scans misregistration on the fusion process was considered in this study by shifting the second scan in a subgroup of the simulated datasets by one and two voxels (4 and 8 mm

respectively) in a random direction, and then quantifying the impact on the resulting volume error for each scan after ASEM fusion. The one to two voxel shifts used in this work corresponds to typical registration errors associated with the use of fusion algorithms in CT imaging.

Quantitative measurement of clinical datasets

In patient monitoring studies, the SUV is the most used semi-quantitative parameter and is defined as:

$$SUV = \frac{C}{A \times W} \quad (4)$$

where C is the tracer concentration, A the injected activity and W the patient weight. Among the different SUV indexes available, the most often used in clinical practice are SUV_{max} and SUV_{mean}, computed respectively as the highest value and the mean of voxels values in a given region of interest (ROI), usually defined manually. Contrary to the SUV_{max}, the computation of the SUV_{mean} depends on the volume of the ROI. Our fusion method allows identifying the variations of concentration activities. Therefore, according to the ASEM fused map, an estimation of each individual tumour volume can be carried out. Then, for each scan, a measure of SUV_{mean} can be extracted according to the metabolic volume of the tumour V_{ASEM} estimated with the multi-observation method. Note that the SUV values extracted from the fusion maps are values corrected for PVE due to the deconvolution pre processing step.

Alternative approaches used for comparison

We compared the multi-observation fusion results with methods that have been proposed for patient follow-up studies. Most clinical studies only consider SUV_{max} variation. In order to take into account full metabolically active volumes evolution as a response criterion, it has been suggested to determine them independently on both scans using threshold-based methodologies. Many studies have demonstrated that a fixed threshold value not adequate for this task and that adaptive thresholding taking into account the background uptake performs better (Nestle *et al* 2005 , Tylski *et al* 2010). Manual tumor delineation by experts was not considered in the study, due to its high inter and intra observer variability (Hatt *et al* 2010). We compared our approach to independent delineation of tumors in each scan using adaptive threshold, which value is determined from the estimated contrast between the tumour activity and the background activity and optimized for a given scanner using phantom acquisitions of spheres (Erdi *et al* 1997). Such an optimization was performed for our acquisition protocol and scanner model. An associated SUV_{mean} value was also computed.

Evaluation metric for simulated datasets

True volumes of simulated tumours are known. Therefore, the assessment of the fusion process was achieved by the estimation of volume errors (VE). For each simulated case, segmented maps of the first and second follow-up scans are obtained with the two methods, as illustrated in figure 3 . The individual segmented maps of the multi-observation method can actually be deduced from the ASEM fused map. Although volume errors may be larger than 100% in specific cases for which delineation completely fails, errors were limited to 100 %.

Quantitative variation for clinical datasets

No ground-truth was available for the clinical follow-up cases. Hence, to compare the methods, the following quantitative indexes have been considered. The variation of metabolic tumour volume (ΔV) and mean of SUV (ΔSUV_{mean}) between the pre-treatment and the follow-up PET scans for the different methods were measured with the two methods. In addition, the evolution of the original (without PVE correction) maximum SUV in the ROI was considered for comparison as it is the one currently used in clinical practice in oncology and defined as the gold standard. Contrary to the simulated datasets, the background level in clinical cases is not homogenous due to the physiological uptake of the mediastinum. The intra-observer variability of the adaptive thresholding method has been investigated for the same context of esophageal lesions in a previous study and demonstrated significant variability (Hatt *et al* 2011). Therefore the adaptive threshold segmentation was carried out here by two clinicians with similar training and experience for each follow-up case, in order to evaluate the impact on the measurement of quantitative variations. The two clinicians followed a specific protocol: they were instructed to measure the mean background value by placing manually a ROI within the mediastinum, at least a few cm away from the lesions. They were free to choose the exact location and size of the ROI.

Results

Results on simulated datasets

For the selected simulated cases, the fusion maps obtained by our multi-observation method are illustrated in figure 3(e) . In order to facilitate the interpretation of the ASEM fused maps, colors have been affected to the different uptake variations. Blue areas represent a response (negative difference in tracer uptake between the two scans) whereas green color is associated with a stable tumor (similar significant uptake in both scans at this location). Red color was used to indicate tumor progression (higher uptake in second scan with respect to first one). Note that ASEM never wrongly resulted in tumor progression or regression. The intensity associated with each voxel

in the fusion map is set as the SUV relative variation (ΔSUV) between the first and the second scan. The segmented maps of the first and second chosen follow-up scan computed the different methods are presented in figure 3(f–h) .

Mean volume errors and associated standard deviation associated with the use of ASEM fusion or independent adaptive threshold-based delineation, for all pre treatment and post treatment simulated cases are presented in figure 4 . The VE computed for the first follow-up scan was significantly (Kruskal-Wallis tests p <0.0001) lower for ASEM (−2.6±8%) than for the adaptive threshold (+28± 17%). For the second follow-up scan, however the VE were higher for both methods. Adaptive threshold led to higher overestimation of the tumor volumes than for the pre treatment image (+30±15%), whereas ASEM led to underestimation of the true post treatment tumor volume (−9±25%) with a larger variability. Adaptive threshold gave large volume overevaluation in all cases, whereas ASEM led to better results in several cases but higher errors in some other cases. In both pre and post treatment images, the adaptive threshold method tended to overestimate the tumor volume, with larger absolute errors than the ASEM method, that tended to underestimate the volume in the second scan.

Three selected cases among the simulated dataset are illustrated in the figure 3 . For the first case, no significant differences were observed between the use of the adaptive threshold applied independently to each image and the ASEM method. For the second more complex case, both methods showed different results. As illustrated in figure 3(f) , the adaptive threshold led to an underestimation of the tumors uptake, contrary to the ASEM method. As the fusion map (fig. 3(e)) shows, the disappearing lesion is correctly identified (in blue) whereas the larger lesion is shown as stable (in green) with a small blue part, indicating that this tumor uptake has indeed smaller volume in the post treatment scan. The third case illustrates the evolution of a necrotic tumour, on which ASEM correctly identified the various parts of evolution (blue, green, red).

Finally, the ASEM fusion proved robust to random mis registration of one voxel (4mm) with volume errors increasing from −2.6±8% to −9±16% and −9±25% to −13.5±30% for first and second scan respectively. A spatial shift of two voxels (8 mm) led to similar error levels regarding the first scan (−9±17% vs. −9±17%) but a higher increase regarding the second scan (−21±37% vs. −13.5±30%).

Results on clinical datasets

The fusion maps obtained by applying the ASEM method to three representative clinical follow-up cases are illustrated in figure 5(c) . Three colors were used to represent physiological uptake (in yellow), tumour response (in blue) area and stable tumour (in green), underlying the partial response status of these patients. The color intensity associated to the voxels classified as responders or stable is determined by the SUV relative variation (ΔSUV) between the first and the second follow-up scans. The segmented maps of the pre-treatment and post-treatment scans, computed with the adaptive threshold and the ASEM methods are illustrated in figure 5(d–g) .

The quantitative measurements estimated for each clinical case individually are showed in table 2 , and mean measurements, estimated for all the patients are showed in table 3 . The clinical cases were more challenging to analyze than the simulated cases, due to a combination of noisier and more heterogeneous background and tumor uptake distributions. Since the patients were classified as partial responder the tumor uptakes were expected to exhibit a significant decrease of the SUV_{max} (at least 30%) between the pre and post treatment scans. Similarly, the tumor volumes and associated mean tracer uptakes should also decrease (at least by 20 to 30%) and this can be confirmed visually.

With or without PVC, variations of SUV_{max} were higher than 30% (−52±16% without PVC and −58±23% with PVC), as expected by their partial responder status. Regarding the other quantitative measurements, there was no significant (p>0.05) difference between the mean variation of SUV_{mean} obtained with each observer using the adaptive threshold (−53±23 for Tb_1 , −55%±24% for Tb_2). SUV_{mean} variation deduced from the ASEM fusion maps was slightly, although not significantly (p>0.05) lower (−49±24%).

By contrast, the variations of tumor volumes were significantly different for all approaches. The variations measured by the two observers using the adaptive threshold were significantly (p<0.0001) different (58±222% for Tb_1 and −31±54% for Tb_2). The results obtained with the ASEM method were also significantly different from both adaptive threshold results (−66±11%). They were also much more homogeneous across the entire group of patients (11% standard deviation only). These results can be explained by analyzing the quantitative parameters of each patient individually. Among the seven clinical cases, the adaptive threshold and ASEM performed differently. On the one hand, for patient 1, 2 and 4, the variation of SUV_{mean} and tumour volume estimated with the adaptive threshold and the ASEM method were similar and pertinent with respect to the partial responder status of the patients. On the other hand, for patient 3 and 5, the tumor volume variation estimated by the two clinicians using adaptive thresholding were significantly different with a factor of 2 to 4 between the two measured variations (−81 and −40% for patient 3, −41 and −13% for patient 5). Finally, the use of adaptive threshold on patient 6 and 7 led to completely aberrant values (above +400 and +75%, and +360 and −16%) contrary to the ASEM method that produced much more consistent volume variation results.

In order to illustrated the different behavior of the adaptive threshold and ASEM method, three clinical follow-up cases corresponding to the patients 2, 5 and 7 are illustrated in figure 5 . Regarding the patient 2, both adaptive thresholding applied independently to each PET

scan and the proposed fusion method resulted in similar measurements leading to similar segmented maps. The first follow-up scan of patient 3 clearly exhibited a heterogeneous uptake within the tumor as shown in figure 5(a) . The two segmentation maps obtained with T_b clearly underestimated the overall tumor volume as shown in figure 5(d) and 5(f) by excluding the central part of the functional uptake, whereas ASEM included it. The poor reproducibility of the adaptive thresholding methodology is also emphasized for the patient 7, for which the volume variation of the first observer is clearly overestimated (+360%) contrary to the second observer which tends to under estimate this volume evolution (−16%) while the ASEM method estimated it at −60%.

Discussion

The use multiple PET scans for response to therapy assessment is rising in oncology, due to the need to assess response to therapy earlier, in order to improve patient's management in radiotherapy and/or chemotherapy. On the other hand, the use of different radiotracers to visualize processes such as cellular proliferation or hypoxia for instance is generating a large amount of research especially in radiotherapy and early therapy assessment (Shields *et al* 2003 , Vaupel *et al* 2007).

The aim of this study was to propose a fusion method based on the multi-observation framework in order to specifically address the simultaneous analysis of multiple follow-up PET scans in the context of response to therapy assessment. The use of a fusion method taking into consideration both scans at the same time was expected to produce more reliable results than independent delineations performed on both scans separately. The ASEM method demonstrated the ability to merge patient follow-up PET scans through unsupervised Bayesian estimation, with especially good results on the first scan (error −2.6±8%) and mostly good results on the second scan, with however a few cases that prove more difficult especially considering the second scan, therefore leading to a higher mean error and standard deviation of −9±25%. On simulated datasets, the adaptive threshold applied independently on both images led to higher errors than the ASEM fusion with a systematic overestimation for both the first and second scan (+28±17% and +30±15% respectively). In the real clinical datasets however, a significantly higher variability in the quantitative parameters measured with the adaptive threshold method was observed for four patients out of seven.

These results can be explained by the fact that simulated data were generated with low and uniform physiological uptake and considering homogenous tumor uptake, as well as only one user to manually determine the background region of interest. However, noisy and heterogeneous uptake in nearby healthy tissues are very common in actual clinical datasets affecting the delineation process, for instance in esophageal cancer for which the tumor is located in the mediastinum and close to lung tissues. ASEM seemed much more resilient with respect to non uniform background and reduced tumor-to-background contrasts thanks to the use of spatial (the local adaptive priors) and inter observation correlations (multi observation framework using covariance matrices in the observation model) within the ASEM fusion. It is to be emphasized that the ASEM fusion requires accurate co-registration of PET datasets with a target registration error of 1 voxel (4 mm) as demonstrated by the results obtained by shifting the second scan in the simulated datasets by one and two voxels (4 and 8 mm) in random directions. The fusion was rather robust to 1 voxel spatial shifting, however volume errors regarding the second scan became higher when shifting by 2 voxels was applied, which can be explained by the fact that as second scan tumor volumes are usually smaller, a mis registration can lead to a higher impact on the volume error of such small volumes.

Despite being more dependent on the noise and the images reconstruction, the SUV_{max} is nevertheless the parameter the most commonly used in clinical routine to assess and quantify tumor evolution and response to therapy (Weber *et al* 2007 , Nahmias *et al* 2008). However the variation of SUV_{max} only may be not sufficient to characterize the tumour response, without taking into account the information of the metabolic tumour volume, especially for early therapy assessment. The SUV_{mean} is considered more reproducible than SUV_{max} , but may depend on the definition of the tumor volume (Tylski *et al* 2010).

In this study, three quantitative indexes, namely the SUV_{max} , SUV_{mean} and volume were computed with both methodologies. The analysis of the volume and SUV_{mean} variations give additional features to characterize the tumour response. In our results, SUV measurements variations between pre and post treatment scans were similar independently on the delineation used. Tumor volume variations measured by the clinicians using the adaptive method were close for patients 1, 2 and 4, and were significantly different for patients 3 and 5, emphasizing the user dependency of such method in the presence of heterogeneous physiological uptake, which is often the case for esophageal tumours (Hatt et al, 2011). They were in addition aberrant for patients 6 and 7, demonstrating the accuracy limitation of such an approach. By contrast tumor volume variations measured by ASEM were much more homogeneous across the group of patients (−66±11%). The poor reproducibility of the adaptive method was first mentioned by Nestle *et al* in 2005 in the case of Non-Small Cell Lung Cancer. The measurements and segmented maps obtained with the ASEM method were more appropriate considering the known partial responder status of these patients. The combination of pertinent quantitative indexes such as the metabolic volume and activity concentration in the tumour, measured with a robust method could be valuable to thoroughly assess tumor response, as illustrated with the use of the ASEM method on the clinical datasets.

Conclusion

A fusion method based on the multi-observation Bayesian framework was proposed to assess multi PET scans in the context of therapy response. Using Bayesian framework, the proposed method can potentially be applied on patient follow-up and multi-tracer datasets in order to assess accurate treatment response and tumour volume definition by automatically delineating the different variations of activity within the tumour. In this study, the multi-observation method has been applied to simulated and clinical follow-up PET images and compared with current threshold-based methods used in clinical practice for the assessment of the therapeutic response. On simulated datasets, the adaptive threshold applied independently on both images led to higher errors than the ASEM fusion. The adaptive threshold proves unreliable for more than half the patients, whereas ASEM produced measurements in line with what could be expected with respect to the classification of the considered patients. Future work will also consider more than two PET scans within the context of therapy response assessment, as well as multi-tracer studies in order to adapt the proposed fusion approach for the definition of multi-tracer PET target volumes in radiotherapy, especially for dose boosting or dose painting scenarios in radiotherapy (Sovik *et al* 2009).

Acknowledgements:

We would like to thank Prof M Allard and Prof P Fernandez (Department of Nuclear Medicine, Bordeaux University Hospital) for useful discussions in the course of this study.

Appendix

Let's consider two random processes $\mathbf{Y} = (\boldsymbol{y}_t)_{t \le T}$ and $= (x_t)_{t \le T}$, modeling respectively the observations and the fusion map. Considering a mixture of multi-dimensional Gaussian density probability functions, the distribution of (**X, Y**) is hence defined by the priors Π_k, the mean vectors μ_k and covariance matrices Γ_k associated to each of the *K* classes in the mixture.

- The parameters ($\Pi_k^0, \mu_k^0, \Gamma_k^0$) defining the Gaussian mixture of the (**X, Y**) distribution and the class number K are initialized with a FKM algorithm based on entropy criterion.

- The mixture parameters are then computed with the SEM algorithm by sampling several realizations of **X** according to its posterior distribution *p* (**X** |**Y**) defined for all the voxels $t \in [1, T]$ and each class $k \in [1, K]$ as :

$$p(X_t = k \mid \mathbf{Y}_t) = \frac{\pi_{t,k} \times f(\boldsymbol{Y}_t, \boldsymbol{\mu}_k, \Gamma_k)}{\sum_{q=1}^{K} \pi_{t,q} \times f(\boldsymbol{Y}_t, \boldsymbol{\mu}_q, \Gamma_q)}$$

(6)

where $\pi_{t,k}$ is the adaptive prior of the voxel *t* and the class *k* defined at the step 4, and $f(\boldsymbol{Y}_t, \mu_k, \Gamma_k)$ is the multi-dimensional Gaussian defined for the k^{th} class by the mean vectors μ_k and covariance matrices Γ_k

- For each voxel and associated observation vector $\boldsymbol{Y}_t$ with $t \in [1, T]$, a posterior realization called $R = (r_1 \ldots r_T)$ is sampled and a partition $Q = (Q_1, \ldots, Q_K]$ is defined as:

$$Q_k = \{r_t \mid r_t = k\}$$

(7)

where Q_k is the partition associated to the *k* class. Using these realizations, the parameters of the Gaussian mixture are estimated.

- First, in the adaptive framework, priors are re-estimated using a local neighboring 3D cube, hence priors for each voxel depend on its position in the image and the current state of its neighbors in the posterior realization. Replacing global prior Π_k, local priors $\pi_{t,k}$ are defined for each voxel and each class as :

$$\pi_{t,k} = \frac{1}{Card(C_t)} \sum_{j \in C_t}' \delta(r_j, k) \text{for } k \in [1, K]$$

(8)

where C_t is the estimation cube and δ is the Dirac function. For our application we chose a cube of size (3 × 3 × 3) voxels.

- The mean vector associated to the k^{th} class can be computed for each *b* image with:

$$\mu_k^{(b)} = \frac{\sum_{t \in Q_k} Y_t^{(b)}}{Card(Q_k)}, \text{ for } k \in [1, K] \text{ for } b \in [1, B]$$

(9)

- The covariance matrix associated to the k^{th} class is defined as:

$$\Gamma_k = \frac{\sum_{t \in Q_k} [Y_t - \boldsymbol{\mu}_k]^t [Y_t - \boldsymbol{\mu}_k]}{Card(Q_k)}, \text{ for } k \in [1, K]$$

(10)

- The decision step based on the maximum likelihood criteria computes the posterior probability $p(\mathbf{X} | \mathbf{Y})$ and selects for each voxel the class that maximizes it:

$$k_{max} = \arg\max_{k \in [1,..,K]} p(X_t = k | Y_t), \forall t$$

(4)

where K_{max} is the estimated maximized class.

References:

- Bentzen SM . 2005 ; Theragnostic imaging for radiation oncology: dose-painting by numbers . Lancet Oncol . 6 : 112 - 117
- Boussion N , Cheze Le Rest C , Hatt M , Visvikis D . 2008 ; Incorporation of wavelet based denoising in iterative deconvolution for partial volume correction in whole body PET imaging . European Journal of Nuclear Medicine and Molecular Imaging . 36 : (7) 1064 - 75
- Celeux G , Diebolt J . 1986 ; L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités . Revue de statistique appliquée . 34 : (2)
- Delignon Y , Marzouki A , Pieczynski W . 1997 ; Estimation of Generalized Mixtures and Its Application in Image Segmentation . IEEE Transactions on Image Processing . 6 : (10)
- El Naqa I , Grigsby P , Apte A , Kidd E , Donnelly E , Khullar D , Chaudhari S , Yang D , Schmitt M , Laforest R . 2009 ; Exploring featurebased approaches in PET images for predicting cancer treatment outcomes . Pattern Recogn . 42 : 1162 - 1171 190
- Erdi YE , Mawlawi O , Larson SM , Imbriaco M , Yeung H , Finn R , Humm JL . 1997 ; Segmentation of Lung Lesion Volume by Adaptative Positron Emission Tomography Image Thresholding . Cancer . 80 : (S12) 2505 - 2509
- Hatt M , Cheze Le Rest C , Pradier O , Visvikis D . 2009 ; Automatic PET tumour delineation for patient's follow-up and therapy assessment . Journal of Nuclear Medicine . 50 : (S2) 182 -
- Hatt M , Cheze Le Rest C , Aboagye EO , Kenny LM , Rosso L , Turkheimer FE , Albarghach NM , Pradier O , Visvikis D . 2010 ; Reproducibility of 18F-FDG and 18F-FLT PET tumour volume measurements . Journal of Nuclear Medicine .
- Hatt M , Visvikis D , Albarghach NM , Tixier F , Pradier O , Cheze-le Rest C . 2011 ; Prognostic value of 18F-FDG PET image-based parameters in sophageal cancer and impact of tumour delineation methodology . European Journal of Nuclear Medicine and Molecular Imaging . 38 : (7) 1191 - 1202
- Jarritt H , Carson K , Hounsel AR , Visvikis D . 2006 ; The role of PET/CT scanning in radiotherapy planning . British Journal of Radiology . 79 : S27 - S35
- Krak NC , Boellaard R . 2005 ; Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial . European Journal of Nuclear Medicine and Molecular Imaging . 32 : 294 - 301
- Krishnapuram R , Keller JM . 1994 ; Fuzzy and possibilistic clustering methods for computer vision . SPIE Institute series . 12 : 133 - 159
- Larson S , Erdi Y , Akhurst T , Mazumdar M , Macapinlac H , Finn R , Casilla C , Fazzari M , Srivastava N , Yeung H . 1999 ; Tumour Treatment Response Based on Visual and Quantitative Changes in Globals Using PET-FDG Imaging, The Visual Response Score and the Change in Total Lesion Glycolysis . Clin Positron Imaging . 2 : 159 - 171 160
- Le Maitre A . 2009 ; Incorporating patient specific variability in the simulation of realistic whole body 18F-FDG distributions for oncology applications . Proceedings of the IEEE . in press
- Le Pogam A , Descourt P , Hatt M , Boussion N , Visvikis D . 2009 ; A combined 3-D wavelet and curvelet approach for edge preserving denoising in emission tomography . Journal of Nuclear Medicine . 50 : (S2) 533 -
- Lin C , Itti E , Haioun C . 2007 ; Early ^{18} F-FDG PET for prediction of prognosis in patient with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis . J Nucl Med . 48 : 1626 - 1632
- Mankoff DA , Muzi M , Krohn KA . 2003 ; Quantitative PET imaging to measure tumour response to therapy: what is the best method? . Mol Imaging Biol . 5 : 281 - 5
- Masson P , Pieczynski W . 1993 ; Adaptive Mixture Estimation and Unsupervised Local Bayesian Image Segmentation . IEEE Transactions on Geosciences and Remote sensing . 31 : (3)
- Nahmias C , Wahl L . 2008 ; Reproducibility of Standardized Uptake Value Measurements Determined by 18F-FDG PET in Malignant Tumours . J Nucl Med . 49 : 1804 - 1808 158, 159
- Necib H , Dusart M , Tylski P , Vanderlinden B , Buvat I . 2008 ; Detection the tumour changes between two FDG PET scans using parametric imaging . J Nucl Med Meeting Abstracts . 49 : 121 - 161
- Nestle U , Kremp S , Schaefer-Schuler A , Sebastian-Welch C , Hellwig D , Rübe C , Kirsch CM . 2005 ; Comparison of Different Methods for Delineation of 18F-FDG PET-Positive Tissue for Target Volume Definition in Radiotherapy of Patients with Non-Small Cell Lung Cancer . Journal of Nuclear Medicine . 46 : (8) 1342 - 8
- Ouksili Z . 2007 ; Accurate PET/PET registration of serial to assess lung tumour evolution . 4th IEEE International Symposium on Biomedical Imaging 732 - 735
- Peng A , Pieczynski W . 1995 ; Adaptive Mixture Estimation and Unsupervised Local Bayesian Image Segmentation . Graphical Models and image processing . 57 : (5) 389 - 399
- Pieczynski W . 2003 ; Modèles de Markov en traitement d'images . Traitement du Signal . 20 : (3) 255 - 277
- Provost JN . 2001 ; Classification bathymétrique en imagerie multispectrale SPOT . PhD Thesis . Université Bretagne Occidentale ;
- Shields AF . 2003 ; PET imaging with 18F-FLT and thymidine analogs: promise and pitfalls . J Nucl Med . 44 : 1432 - 1434
- Soret M , Bacharach SL , Buvat I . 2007 ; Partial-volume effect in PET tumor imaging . J Nucl Med . 48 : 932 - 945
- Sovik A , Malinen E , Olsen DR . 2009 ; Strategies for biologic dose escalation: a review . Int J Radiation Oncology Biol Phys . 73 : (3) 650 - 658
- Tylski P , Stute S , Grotus N , Doyeux K , Hapdey S , Gardin I , Vanderlinden B , Buvat I . 2010 ; Comparative assessment of methods for estimating tumor volume and standardized Uptake Value in FDG PET . J Nucl Med . 51 : 268 - 276
- Therasse P , Arbuck S , Eisenhauer E , Wanders J , Kaplan R , Rubinstein L , Verweij J , Van Glabbeke M , van Oosterom A , Christian M . 2000 ; New guidelines to evaluate the response to treatment in solid tumors . Journal of National Cancer Institute . 92 : 205 - 216
- Vaupel P , Mayer A . 2007 ; Hypoxia in cancer: significance and impact on clinical outcome . Cancer Metastasis Rev . 26 : 225 - 239
- Wahl R , Jacene H , Kasamon Y , Lodge M . 2009 ; From RECIST to PERCIST : Evolving Considerations for PET Response Criteria in Solid Tumors . J Nucl Med . 50 : 122 - 150
- Weber W , Ziegler S , Thodtmann R , Hanauske A , Schwaiger M . 1999 ; Reproducibility of metabolic measurements in malignant tumours using FDG PET . J Nucl Med . 40 : 1771 - 1777 158,159
- Weber W . 2007 ; ^{18} F-FDG PET in Non-Hodgkin's Lymphoma: Qualitative or Quantitative? . J Nucl Med . 48 : 1580 - 1582
- World Health Organization . 1979 ; WHO, Handbook for Reporting Results of Cancer Treatment . World Health Organization ; sold by WHO Publications Centre USA

- Young H , Baum R , Cremerius U , Herholz K , Hoekstra O , Lammertsma A , Pruim J , Price P . 1999 ; Measurement of clinical and subclinical tumour response using [18F]-fluorodeoxyglucose and positron emission tomography : review and 1999 EORTC recommendations . Eur J Cancer . 35 : 1773 - 1782
- Zaidi H , El Naqa I . 2010 ; PET-guided delineation of radiation therapy treatment volumes: A survey of image segmentation techniques . Eur J Nucl Med Mol Imaging . 37 : (11) 2165 - 2187

Figure 1

Multi-observation framework of multi-tracer and patient follow-up data.

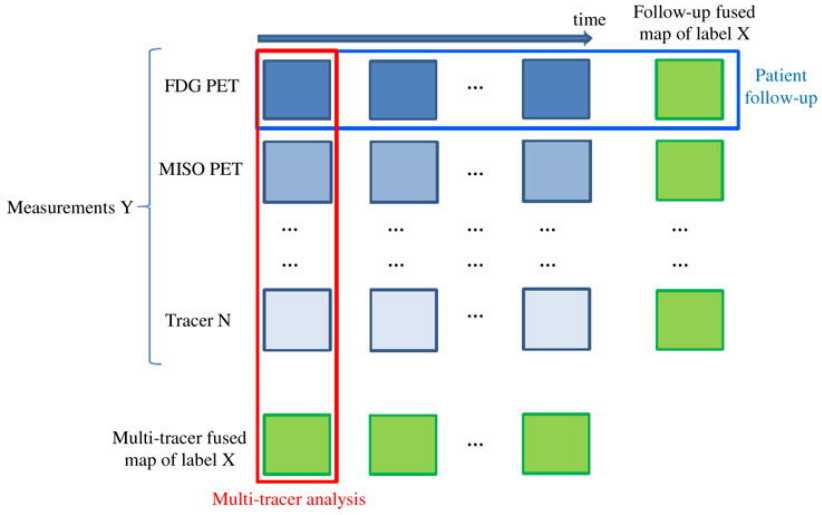

Figure 2

Illustration of a VOI definition in the pre-treatment scan (a) and automatically reported on the registered mid-treatment scan (b).

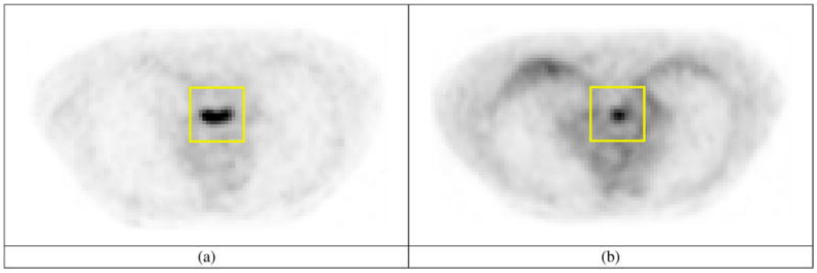

Figure 3

(b) and (d) Simulated follow-up tumours, (a) and (c) associated ground-truths, (e) ASEM fusion map, individual segmented map with the adaptive threshold (f) and (g), and the ASEM method (h) and (i) for the two simulated cases.

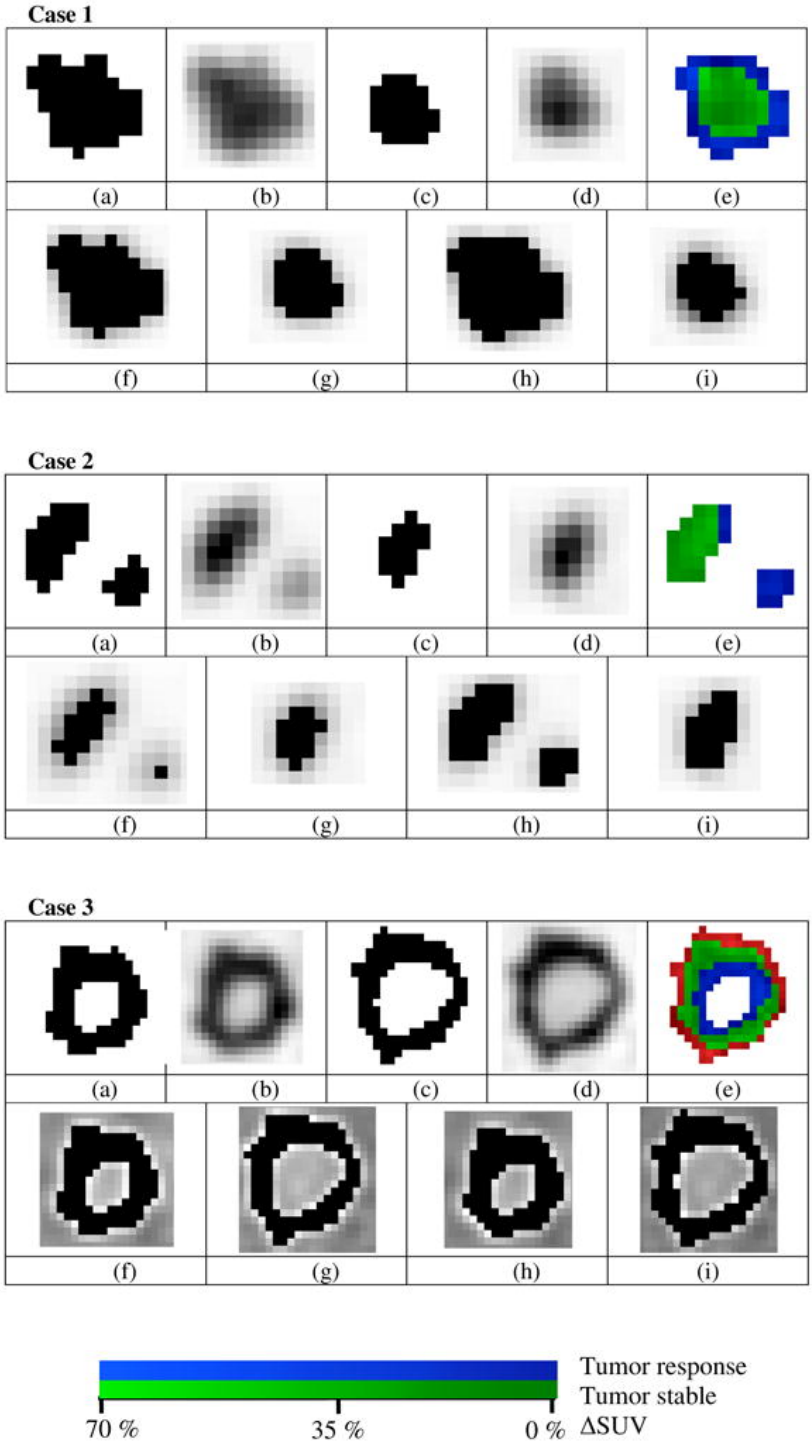

Figure 4

Mean VE (%) with standard deviation as error bars of the first and second follow-up scans for adaptive threshold and ASEM methods applied to the simulated cases.

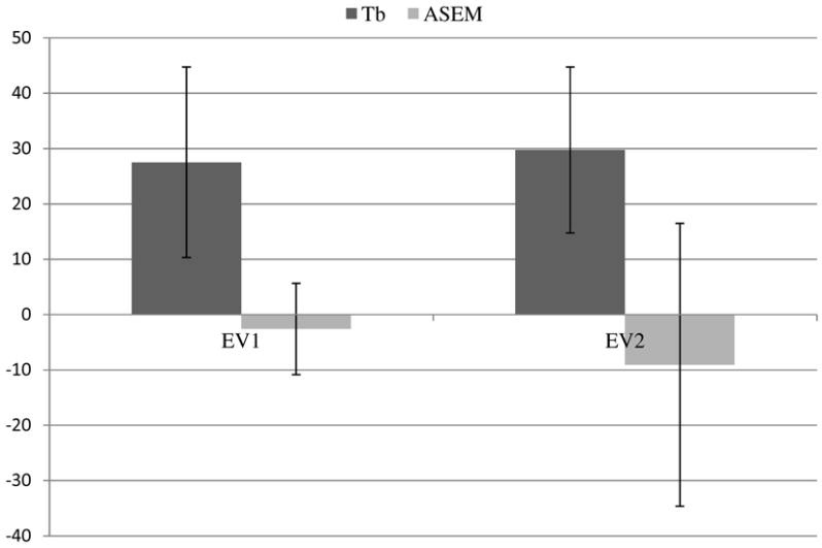

Figure 5

(a) and (b) Clinical follow-up tumours (c) ASEM fusion maps, individual segmented map of the two clinicians (d), (e) and (f), (g) with the adaptive threshold, (h) and (i), the ASEM method, for three real clinical datasets.

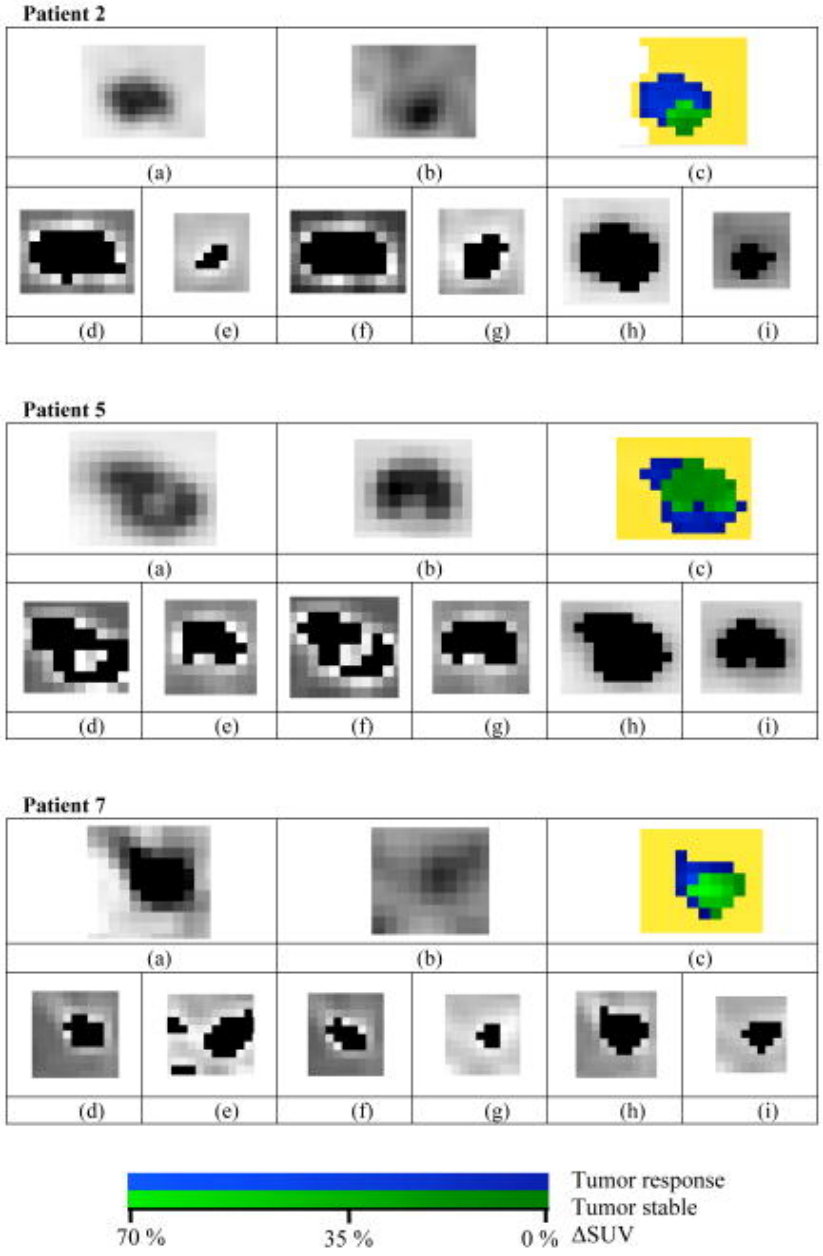

Table 1

Mean tumour to background ratios and mean tumour volumes computed for all the simulated follow-up cases.

	Tumour volume (cm^3)		T/B ratio	
PET 1	34.1 ± 27	(6 – 90)	5.7 ± 2	(2.7 – 9.3)
PET 2	17.4 ± 2.5	(1.9 – 101)	4.2 ± 1.4	(2.0 – 6.5)

Table 2

Measurements of Volume, SUV_{mean} and SUV_{max} evolution, computed with the adaptive threshold and the ASEM methods for each patients.

Patient	Method	ΔSUV_{mean} (%)	ΔV (%)	ΔSUV_{max} without PVC (%)	ΔSUV_{max} with PVC (%)
1	Tb 1	−35.7	−79.8	−51.7	−35.8
	Tb 2	−33.5	−83.2		
	ASEM	−33.1	−80.4		
2	Tb 1	−69.5	−86.3	−74.2	−76.9
	Tb 2	−74.2	−64.9		
	ASEM	−71.4	−65.2		
3	Tb 1	−59.7	−81.4	−60.2	−66.9
	Tb 2	−68.9	−40.0		
	ASEM	−62.7	−59.5		
4	Tb 1	−27.0	−66.1	−32.7	−25.7
	Tb 2	−24.2	−71.7		
	ASEM	−26.7	−59.2		
5	Tb 1	−28.4	−41.1	−28.1	−42.2
	Tb 2	−35.2	−12.9		
	ASEM	−20.6	−52.5		
6	Tb 1	−87.3	402.5	−56.1	−87.8
	Tb 2	−85.7	74.1		
	ASEM	−83.3	−81.9		
7	Tb 1	−65.9	361.0	−58.8	−66.9
	Tb 2	−60.7	−16.1		
	ASEM	−47.9	−59.8		

Table 3

Mean measurements of Volume, SUV_{mean} and SUV_{max} evolution, computed with the adaptive threshold and the ASEM methods for all the patients.

Method	ΔSUV_{mean} (%)	ΔV (%)	ΔSUV_{max} without PVC (%)	ΔSUV_{max} with PVC (%)
Tb 1	−53.3 ± 23.2	58.4 ± 221.7		
Tb 2	−54.7 ± 23.6	−30.7 ± 53.5	−51.7 ± 16.2	−57.5 ± 23.0
ASEM	−49.4 ± 23.9	−65.5 ± 11.3		