OPEN ACCESS
IOP Publishing | Institute of Physics and Engineering in Medicine Physics in Medicine & Biology
Phys. Med. Biol. **61** (2016) 4855–4869 doi:10.1088/0031-9155/61/13/4855

ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography

Beatrice Berthon[1,5], Christopher Marshall[1], Mererid Evans[2] and Emiliano Spezi[3,4]

[1] Wales Research & Diagnostic PET Imaging Centre, Cardiff University, CF14 4XN, Cardiff, UK
[2] Velindre Cancer Centre, CF14 2TL, Cardiff, UK
[3] School of Engineering, Cardiff University, Queen's Buildings, The Parade, CF24 3AA, Cardiff, UK
[4] Department of Medical Physics, Velindre Cancer Centre, CF14 2TL, Cardiff, UK

E-mail: beatrice.walker@espci.fr, MarshallC3@cardiff.ac.uk, Mererid.Evans@wales.nhs.uk and ESpezi@cardiff.ac.uk

Received 9 November 2015, revised 21 April 2016
Accepted for publication 27 April 2016
Published 7 June 2016

Abstract

Accurate and reliable tumour delineation on positron emission tomography (PET) is crucial for radiotherapy treatment planning. PET automatic segmentation (PET-AS) eliminates intra- and interobserver variability, but there is currently no consensus on the optimal method to use, as different algorithms appear to perform better for different types of tumours. This work aimed to develop a predictive segmentation model, trained to automatically select and apply the best PET-AS method, according to the tumour characteristics.

ATLAAS, the automatic decision tree-based learning algorithm for advanced segmentation is based on supervised machine learning using decision trees. The model includes nine PET-AS methods and was trained on a 100 PET scans with known true contour. A decision tree was built for

[5] Dr Berthon is now at Institut Langevin, ESPCI Paris, PSL Research University, CNRS UMR 7587, INSERM U979, 17 rue Moreau 75012 Paris, France.

each PET-AS algorithm to predict its accuracy, quantified using the Dice similarity coefficient (DSC), according to the tumour volume, tumour peak to background SUV ratio and a regional texture metric. The performance of ATLAAS was evaluated for 85 PET scans obtained from fillable and printed subresolution sandwich phantoms.

ATLAAS showed excellent accuracy across a wide range of phantom data and predicted the best or near-best segmentation algorithm in 93% of cases. ATLAAS outperformed all single PET-AS methods on fillable phantom data with a DSC of 0.881, while the DSC for H&N phantom data was 0.819. DSCs higher than 0.650 were achieved in all cases.

ATLAAS is an advanced automatic image segmentation algorithm based on decision tree predictive modelling, which can be trained on images with known true contour, to predict the best PET-AS method when the true contour is unknown. ATLAAS provides robust and accurate image segmentation with potential applications to radiation oncology.

Keywords: positron emission tomography, image segmentation, supervised machine learning, radiotherapy treatment planning

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

1. Introduction

The past decade has seen an increasing interest in the use of positron emission tomography (PET) using ^{18}F-fluorodeoxyglucose (^{18}F-FDG) in radiation oncology. This technique, allowing the visualisation of the metabolic activity of tumour tissue, shows great promise in applications such as radiotherapy treatment planning and the monitoring of response to therapy. However, the use of PET-based information for these purposes relies on the availability of an accurate and reproducible segmentation technique for defining the tumour, which is made challenging by the low resolution of PET images. The proximity of physiological high uptake areas (Blodgett *et al* 2005) and the intra-tumour heterogeneity, due to large necrotic areas and intense focal uptakes (Haberkorn *et al* 1991, Henriksson *et al* 2007), can make image segmentation particularly difficult, in particular in anatomical regions such as the head and neck (H&N). Although clinical expertise is crucial in delineating the Gross Tumour Volume, the manual delineation process is extremely time consuming. In addition, there is a large amount of evidence in the literature showing that manual PET contours are highly dependent on the operator as well as on the clinical case under consideration (Steenbakkers *et al* 2006, Breen *et al* 2007). Automatic PET-based automatic segmentation (PET-AS) was introduced to eliminate the intra- and inter-observer variability inherent in freehand outlining. Low-level PET-AS methods, such as threshold-based methods, have been shown to lack accuracy and robustness to a number of image parameters (Ford *et al* 2006)and the use of more advanced segmentation algorithms has been recommended instead (Grégoire and Chiti 2011). The number of published and validated advanced PET-AS methods is currently growing (Shepherd *et al* 2012) as more clinical centres are including PET in their planning protocols. However, the focus of the literature remains on individual experience of different centres, with methods tested on their own data, often covering only a small range of variation in image parameters. The wide range of variation in tumour characteristics observed for clinical H&N cases and the large number of segmentation methods published independently make it difficult to recommend a single delineation method. Only a small number of papers have focused on comparing automatic PET segmentation methods. Zaidi *et al* (2012) for instance determined the best image segmentation method to

use in pharyngolaryngeal squamous cell carcinoma by identifying the method which performed best overall compared to 3D biological tumour volumes defined by histology data. However, as discussed in the study, choosing a single PET-AS method to cover all clinical scenarios needs to be considered with caution, as even the best method overall can lead to large errors in some cases. It is therefore important to challenge the assumption that a single segmentation algorithm exists that yields excellent delineation accuracy for every type of tumour, and rather focus on using the wide knowledge acquired on different methods to achieve optimal segmentation. A small number of publications in the field of medical image segmentation have shown that the combination of the information from different segmentation results, using majority voting or the STAPLE algorithm (Warfield *et al* 2004), can increase the accuracy of the segmentation compared to using individual algorithms (Dewalle-Vignion *et al* 2015, Schaefer *et al* 2015). In this work we propose a different approach, which consists in selecting a single segmentation algorithm from a number of different ones, rather than combining all segmentation results. This can be done using machine learning techniques, which 'learn' from the data they encounter. Supervised machine learning allows a given algorithm to be built and optimised using an existing training dataset for which the true tumour geometry and location is known (the 'ground truth'), in order for it to make the right decisions to achieve optimal performance for cases in which the outcome is not known. This includes methods such as K Nearest Neighbours (Anbeek *et al* 2005, Lyksborg *et al* 2012), Support Vector Machine (Iordanescu *et al* 2012, Jayachandran and Dhanasekaran 2013) and Artificial Neural Networks (Bankman 2000, Reyes-Aldasoro 2000), which have been used in the literature for the segmentation of medical imaging by classifying voxels into different categories. The main advantages of such techniques are their high predictive power and their ability to adapt to any given dataset. Machine learning methods are commonly applied to a test image in order to classify the voxels into different categories (Tabakov and Kozak 2014), or for diagnostic purposes (Hassanien and Kim 2012). However, such methods could also be applied to the classification of training data into groups for which a particular segmentation algorithm would perform best. In particular, decision tree (DT) learning is a supervised learning method, which provides a graphical representation of the algorithm together with the set of classification rules learned during the training process (the tree). DT learning could be a powerful tool in the exploration of a wide range of data for the determination of optimal tumour segmentation. To our knowledge, machine learning algorithms have not yet been used for selecting the optimal segmentation algorithm from a set of available methods. In previous studies, we have investigated and compared the accuracy of a number of segmentation methods, representing the most promising approaches in the current literature, for a wide range of FDG uptake distributions (Berthon *et al* 2013). In particular, we have shown that the accuracy of different methods depends on a number of parameters describing the segmentation conditions such as the volume and tumour-to-background ratio (Berthon *et al* 2014b). Our previous work suggested that a combination, utilizing the advantages of several methods simultaneously, might improve tumour segmentation. In the present work, we describe the development and validation of a tool based on DT learning, designed to achieve this goal.

2. Methods

2.1. Description of the model

In this work, we present ATLAAS[6]: an Automatic decision Tree-based Learning Algorithm for Advanced Segmentation. The ATLAAS predictive segmentation model is designed to select the most accurate PET-AS method for optimal segmentation of a given PET image. The best

[6] Patent pending No PCT/GB2015/052981.

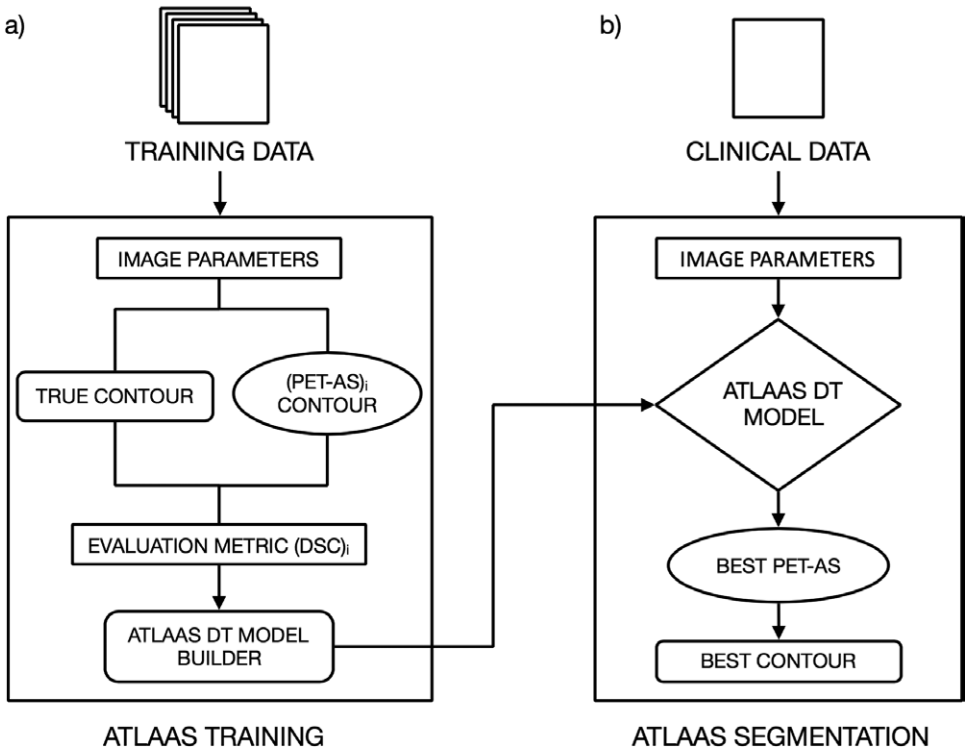

Figure 1. Workflow of (a) the ATLAAS training process (b) the clinical use of the ATLAAS predictive segmentation model.

segmentation method is chosen from a list of advanced PET-AS algorithms built into the system. A list of the PET-AS methods used for this work is given in this section. When segmenting a new image, ATLAAS computes for each PET-AS algorithm its predicted segmentation accuracy using a number of parameters extracted from the target image. The prediction is done using a model consisting of DTs built during the training of ATLAAS. The predictive model is built on a large dataset of PET images of tumours, with the parameters describing the tumour uptake varying within a certain range. The training dataset needs to be large enough to cover a large variety of clinical scenarios and provide accurate classification. Any data for which the ground truth is known can be used (i.e. phantom images, simulated images or clinical data with histopathology reference). This is known as the 'training dataset'. Figure 1 depicts the workflow used in the training phase and in the clinical application of ATLAAS.

The following parameters describing the target object uptake were identified as classifiers of the DT learning method:

- V: target object volume (ml)
- TBRpeak: Ratio between the target object's SUVpeak, calculated as the mean value in a 1cm^3 sphere centred on the maximum standardised uptake value (SUV) in the target object, and the background SUV, calculated as the mean intensity in a 0.5 cm thick extension of the object contour.
- NI: a regional texture feature related to the intensity distribution in the target object. The number NI of intensity levels in the target object is obtained from a grey level co-occurrence matrix based on the methods described by Haralick *et al* (1973) and used in

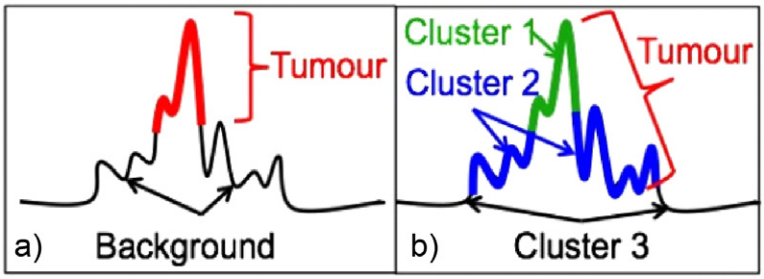

Figure 2. Illustration of the segmentation of a VOI (intensity profile shown in black) by (a) binary thresholding and (b) a segmentation method taking into account the spatial connectivity of voxels or by identifying several regions (clusters) regions within the lesion.

the work by Tixier *et al* (2011). NI is the number of different discrete intensity values in the target object, after resampling the object image to 64 discrete intensity levels as recommended by Tixier *et al*.

Clinical lesions can show a variety of patterns, ranging from globally homogeneous uptake to highly heterogeneous uptake when including, for instance, local hot spots or necrotic areas. This work was therefore based on the assumption that different PET-AS approaches may be more adequate than others for the segmentation of different types of lesions. For example, as illustrated in figure 2, highly heterogeneous lesions are expected to be more accurately segmented by methods taking into account the spatial connectivity of regions in the image or identifying multiple regions, or clusters (figure 2(a)) than by simple thresholding-based methods (figure 2(b)), which may in turn be more adequate for less complex uptake patterns.

A large range of PET-AS methods was used in this study to represent different segmentation approaches. These have been described in detail in previous publications (Berthon *et al* 2013, 2014b), and are summarised in table 1, along with references of recent publications using similar segmentation approaches. All methods were implemented in the Matlab programming language (The Mathworks, Natick USA) using the same procedure, so that the only difference between the methods was the underlying segmentation approach. Methods AT, RG and WT are binary classification methods, which previously showed high accuracy in delineating homogeneous lesions (Berthon *et al* 2013, 2014b). The clustering methods KM, FCM and GCM were each applied to the detection of 2, 3 and 4 clusters, with the resulting methods named KM2-KM4, FCM2-FCM4 and GCM2-GCM4 respectively.

2.2. Building of the statistical model

A training dataset was used to build DTs for each PET-AS method described in table 1. We used the PETSTEP simulator (Berthon *et al* 2014a, 2015a) for the rapid generation of a large training set. PETSTEP operates in the open source framework of the Matlab-based computational environment for radiotherapy research (CERR) (Deasy *et al* 2003). It uses a CT image and a map of the background FDG uptake to generate a PET image from tumour any contours drawn by the user. The training dataset was generated based on existing PET/CT data of a fillable phantom. Target tumours objects covering a range of different characteristics relevant to clinical situations were added to the background of this phantom. A total of 100 spherical tumours objects were modelled for volume and maximum uptake values in the range 0.5 ml–50 ml and 4000 Bq ml^{-1}–40000 Bq ml^{-1} respectively. For all the scans in the training

Table 1. Name and description of PET-AS methods used in this study, with references of published work using similar segmentation approaches.

Algorithm	Description	Key references
AT	3D Adaptive iterative thresholding, using background subtraction	Jentzen *et al* (2007) Drever *et al* (2007)
RG	3D Region-growing with automatic seed finder and stopping criterion	Day *et al* (2009)
KM	3D K-means iterative clustering with custom stopping criterion	Zaidi and El Naqa (2010)
FCM	3D Fuzzy-C-means iterative clustering with custom stopping criterion	Belhassen and Zaidi 2010)
GCM	3D Gaussian Mixture Models-based clustering with custom stopping criterion	Hatt *et al* (2009)
WT	Watershed Transform–based algorithm, using Sobel filter.	Geets *et al* (2007), Tylski *et al* (2006)

dataset the target outline was known since it was one of the input parameters in the synthetic image generation process. We refer to this outline as 'reference true contour' or 'ground truth'.

The workflow to build the statistical model was fully automated and computer scripts were used to (1) determine evenly spaced values of the volume and FDG uptake spanning the ranges specified, (2) generate the synthetic training scans, (3) calculate the value of corresponding tumour parameters in the target object, (4) segment the synthetic scans using the PET-AS methods in table 1, (5) calculate the segmentation accuracy for each method by comparing the result of the segmentation with the ground truth (i.e. actual target outline and location) and (6) build the corresponding DTs.

Only the PET-AS methods that performed differently across the dataset were kept in the study provided non-correlated DSC values across the dataset were kept in the study. Correlated methods were identified using SPSS 20 (IBM, Chicago, USA), as pairs of PET-AS for which the Pearson's r correlation coefficient was larger than 0.95 ($r > 0.95$) and the associated p-value smaller than 0.05 ($p < 0.05$). The best performing (highest average DSC) of each correlated pair was kept. Only the following uncorrelated methods were considered for further analysis: AT, RG, WT, KM2, KM3, KM4, FCM2, GCM3 and GCM4.

The segmentation accuracy was assessed by quantifying the conformity of the contours obtained to the reference true contour provided by the simulation template in PETSTEP. We used the DSC (Dice 1945), defined as:

$$\mathrm{DSC} = \frac{2 * |X \cap Y|}{|X| + |Y|}$$

where $|X|$ and $|Y|$ are the number of voxels in the reference and test contour, respectively, and $|X \cap Y|$ the number of voxels in their intersection. DSC above 0.7 is considered as an indicator of good overlap (Dice 1945).

The DTs predicting, for each PET-AS, the DSC score obtained according to the values of the different tumour parameters were built automatically with the Matlab statistics toolbox, based on the Classification And Regression Tree growing method (IBM Corporation 1989 2011). The Matlab method *RegressionTree.fit* was used with its default settings, including the impurity measure 'Gini', ensuring the data is classified into homogeneous groups of values. For each tree, the cross-validation error was calculated as the relative difference between the true actual DSC and the DSC predicted by the tree, averaged on a randomly chosen subset

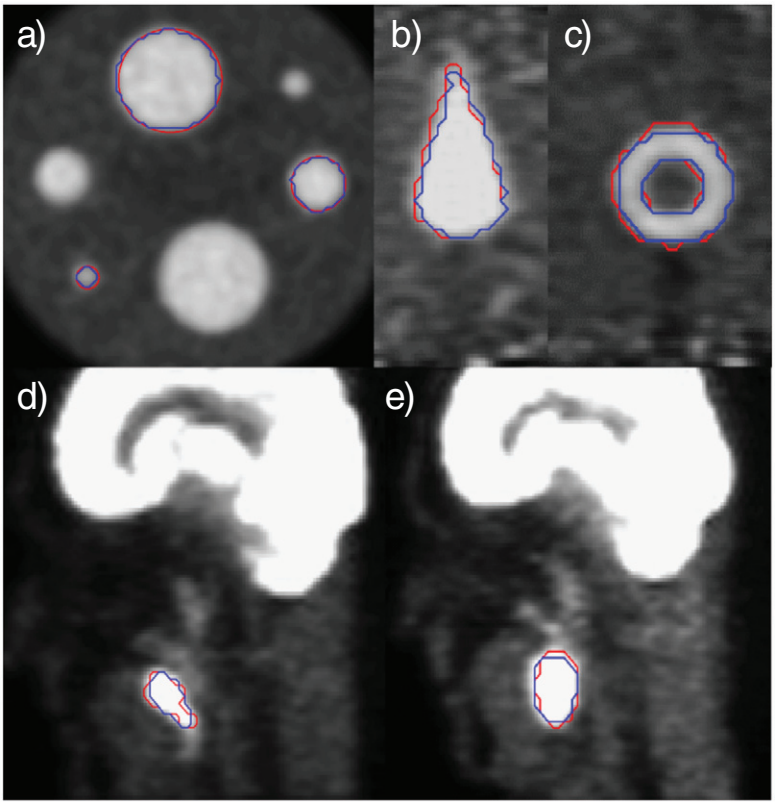

Figure 3. Examples of phantom scans used for the validation of ATLAAS, with overlaid reference true contours (in red) and contours obtained with ATLAAS (in blue). The target objects shown are (a) fillable spheres with thin plastic walls, (b) and (c) non-spherical fillable inserts with thin plastic walls, (d) spheroidal lesion on printed H&N phantom, (e) irregular lesion on printed H&N phantom.

(10%) of the training dataset. We also provide for each DT built (i.e. for each PET-AS method) the relative importance to the model of each predictor (TBRpeak, Volume and NI), which represents the increase in accuracy of the tree as a consequence of including that predictor. The final model was built by choosing the parameter with the highest importance to model across PET-AS methods, and adding the next parameters with highest importance to model among the remaining parameters until the accuracy of the model stopped increasing.

2.3. Validation of the ATLAAS predictive segmentation model

The accuracy of ATLAAS was evaluated for the segmentation of a wide range of phantom data generated for previous studies, within the range of volume and TBR_{peak} used in the training of ATLAAS. The data available included:

- 37 fillable phantom test images obtained from using thin-wall spherical (Berthon *et al* 2013) and non-spherical (Berthon *et al* 2014b) plastic inserts including ellipses, tori, tubes, pear- and drop-shaped objects,
- 17 cases of heterogeneous spheres in homogeneous background (1, 2 or 4 concentric spherical regions of different uptakes) obtained with the printed subresolution sandwich (SS) phantom (Berthon *et al* 2015b),

- 31 H&N homogeneous and heterogeneous (random and Gaussian smoothed uptake) irregular H&N cases generated with a the printed SS phantom, using a printout template representing realistic H&N background uptake derived from clinical PET/CT scans.

Example images taken from each one of these datasets are shown on figure 3.

The ATLAAS predictive segmentation model was applied to each test case of this validation dataset. The PET-AS methods in were also individually applied to each test case. The mean and minimum (min) DSC across the validation dataset were calculated for ATLAAS and for each of the different PET-AS methods. In addition, we determined for each method the percentage of cases where it returned the best DSC (P_{best}) and a DSC value within 10% of the best PET-AS algorithm (P_{10}). The Wilcoxon signed-rank test was used to determine if ATLAAS and the PET-AS generated DSC distributions were significantly different, with a significance value set to $p = 0.05$.

3. Results

3.1. The ATLAAS predictive segmentation mode

The simulated data used for training ATLAAS covered TBRpeak, V and NI ranges of 1.1–52, 0.44–52 ml and 8–38 respectively. The clinical data observed at our centre had TBRpeak, V and NI in the range 1.1–8.3, 0.44–59 ml and 13–63 respectively as shown in table 2.

A total of 9 trees were generated using all tumour parameters described previously as classifiers. Table 3 shows the performance of the different PET-AS methods on the training dataset (mean DSC and P_{best}) and the cross-validation error of the associated DTs. All trees achieved a cross-validation error smaller than 1%, except for RG. The relative importance relative to the model of each predictor is given at the bottom of table 3. TBRpeak was by far the parameter with highest importance to the model overall, followed by Volume.

Figures 4 (a) and (b) shows the DTs built for methods KM2 and AT, including the image parameter and cut-off value at each node, and the predicted DSC at each extremity (leaf) of the tree.

3.2. Evaluation of the ATLAAS predictive segmentation model

The phantom dataset built for the evaluation of ATLAAS included target objects with volumes and TBR_{peak} values in the range 0.58 ml–102 ml and 1.3–17 respectively. The results of the validation of ATLAAS with phantom data are given in table 4, showing the mean and lowest DSC obtained on the dataset as well as P_{best} and P_{10} for ATLAAS and all PET-AS methods considered in the model. ATLAAS returned the best DSC in 56% of cases, and a DSC within 10% of the best in 93% cases, compared to respectively 34% and 88% obtained when using the best single PET-AS methods on this dataset (AT). It is worthwhile noting that on average ATLAAS outperformed all the individual PET-AS methods on all the metrics used. On the phantom dataset, ATLAAS reached a minimum DSC of 0.650, which is more than twice the minimum DSC obtained using the best single PET-AS methods on this dataset (AT). Results of the Wilcoxon signed-rank test, given at the bottom of table 4, showed statistically significantly different DSC values for ATLAAS compared to WT, KM3, KM4, GCM3 and GCM4.

Table 2. Variation range of volume, TBRpeak and NI for the different datasets.

	Volume (ml)	TBRpeak	NI
Clinical data	0.44–59	1.1–8.3	13–63
Training data	0.44–52	1.1–52	8–38
Phantom data	0.58–102	1.3–17	9–57

Table 3. Performance of the PET-AS methods on the training dataset and cross-validation error of the associated tree.

Method	AT	RG	WT	KM2	KM3	KM4	FCM2	GCM3	GCM4
Mean	0.919	0.589	0.929	0.932	0.873	0.811	0.675	0.820	0.917
P_{best} (% cases)	4	22	26	39	<1	<1	<1	9	9
Cross-validation error (%)	<1	9.0	<1	<1	<1	<1	<1	<1	<1
	Relative importance to model (%)								
Vol	<1	<1	96.2	<1	92.2	86.3	<1	<1	<1
TBRpeak	100	95.2	100	100	74.2	100	100	100	100
NI	<1	<1	<1	<1	100	<1	40.1	<1	<1

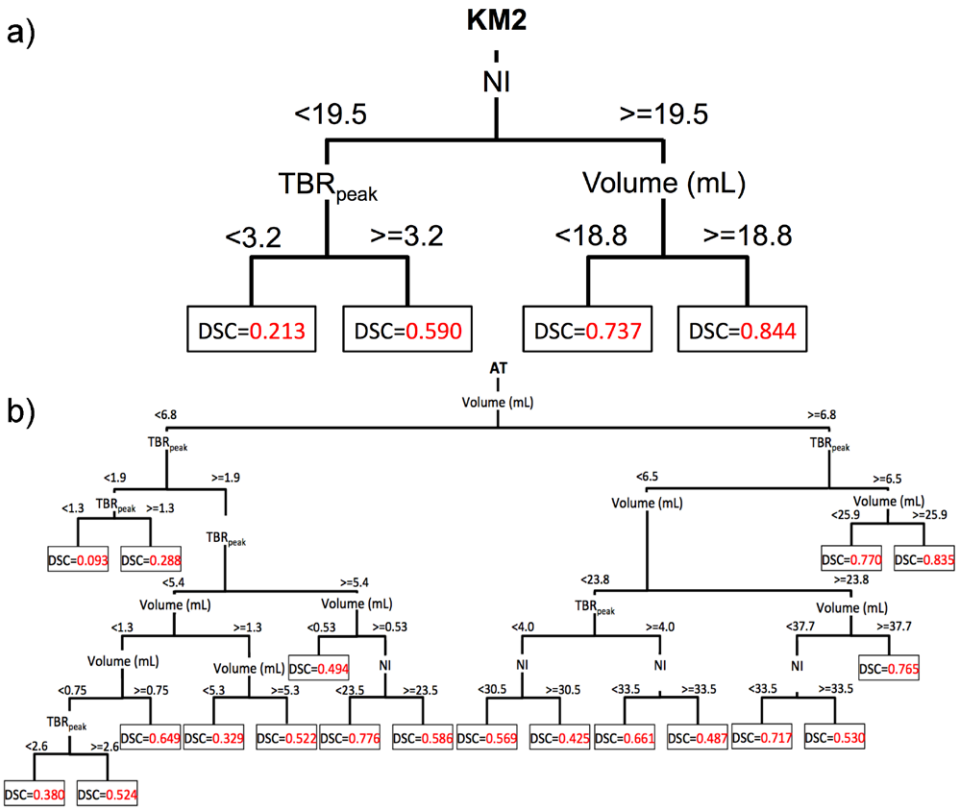

Figure 4. DTs built-in the ATLAAS predictive segmentation model for PET-AS methods KM2 (a) and AT (b).

Table 4. Performance of ATLAAS and of the different PET-AS methods for the phantom validation data, including mean, median and minimum DSC obtained, P_{best} and P_{10} values.

	AT	RG	WT	KM2	KM3	KM4	FCM2	GCM3	GCM4	ATLAAS
Mean DSC	0.841	0.846	0.800	0.836	0.660	0.537	0.587	0.793	0.752	0.858
Min DSC	0.353	0.057	0.214	0.202	0.079	0.088	0.142	0.104	0.099	0.650
P_{best} (%cases)	34	22	3	26	1	<1	2	9	3	56
P_{10} (%cases)	88	81	58	72	10	1	10	68	54	93
Wilcoxon signed-rank test results										
U	3721	3667	4690	3722	6093	6814	6786	4333	4690	
p-value	0.362	0.432	$<10^{-3a}$	0.366	$<10^{-6a}$	$<10^{-6a}$	$<1e^{-6a}$	0.012[a]	$<10^{-3a}$	

[a] Statistically significant ($p < 0.05$).
Note: The results of the Wilcoxon signed rank test are given in the bottom rows.

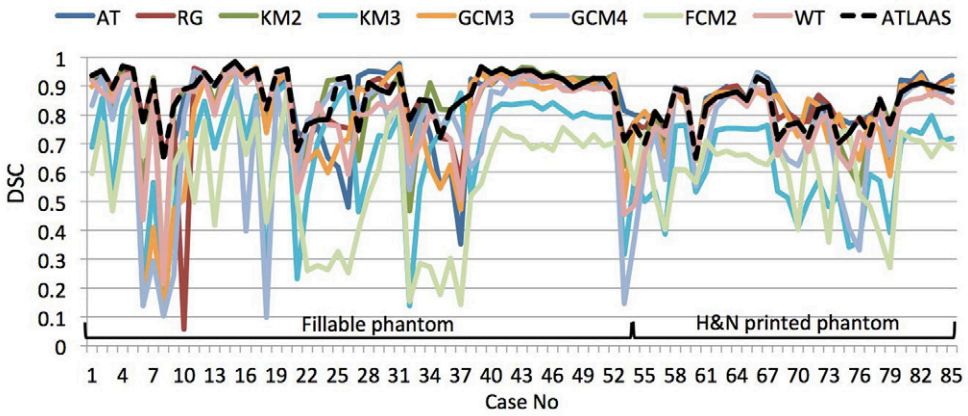

Figure 5. Accuracy of ATLAAS (black dashed line) compared to individual PET-AS methods (colour solid lines) across the validation dataset, including fillable phantom and printed H&N phantom data.

In comparison, ATLAAS built using only TBRpeak, and only TBRpeak and Volume as classifiers achieved mean DSC values across the validation data of 0.816, 0.820 respectively.

A full comparison of the segmentation accuracy (DSC) obtained for ATLAAS (black) and for the single PET-AS methods (colours), for all the cases of the validation dataset, is depicted in figure 5. It can be noted that the curve representing ATLAAS is higher than individual PET-AS curves in most cases. When ATLAAS did not reach the highest DSC, its DSC was still higher than 0.7. In addition, although the performance of other PET-AS methods such as AT, KM2 and RG was good overall, there were cases in which these methods had a low DSC score (e.g. case No. 10, 26, 32 and 38 in figure 5), while ATLAAS maintained a DSC higher than 0.7.

The average accuracy of ATLAAS across the fillable phantom data (figure 5, left) was 0.881 DSC compared to 0.852 for the best of the PET-AS methods. In the case of the printed H&N phantom data (figure 5, right), ATLAAS reached an average 0.819 DSC compared to 0.831 DSC for the best performing PET-AS method.

4. Discussion

The aim of this study was to develop a model able to predict and apply the best PET-AS algorithm for use in 18F-FDG PET imaging. We have developed a novel tree-based decision making algorithm, ATLAAS, and shown that it performs better than any of the individual PET-AS algorithm it is based on, in a range of pre-clinical conditions. Although the data used in this work represented H&N tumours, as it was available at our centre and clinically relevant, ATLAAS can potentially be used for any type of lesions, and many tumour types beyond H&N would benefit from such an optimised segmentation tool. The tool developed has potential applicability for target volume delineation in clinical practice, which can be used across different centres and clinical studies. It is important to note that ATLAAS is not designed as a segmentation method in itself, but rather a framework that can be used to select and apply the best among a pool of segmentation methods. As such, the segmentation methods included in the model have an impact on its accuracy, as the segmentation resulting from ATLAAS on a given case can only ever be as accurate as the best performing segmentation method on that case. In this work, ATLAAS was built using a variety of segmentation methods based on different mathematical approaches. The model would provide optimal segmentation accuracy if it included all the best performing methods for each approach, provided these methods are available and implemented in the same way. In this study, which aimed to provide a proof of concept for the ATLAAS model, we did not include the best existing algorithm for each approach, so an optimal segmentation method may outperform ATLAAS. However, if such a method was included in the model, which is easily done, ATLAAS would allow selecting a different segmentation algorithm in cases where this optimal method would fail, thereby outperforming it overall.

The combined use of different image segmentation algorithms has been investigated by McGurk *et al* (2013) and more recently in two more publications (Dewalle-Vignion *et al* 2015, Schaefer *et al* 2015). These works evaluated voxel-wise methods which combine the information provided by different segmentation methods into a new contour. In our approach, we do not attempt to include the results of all segmentation methods in deriving the final contour, but we predict the PET-AS method that is most appropriate for each case. The advantage of such an approach is that the information provided by low performance segmentations is not used to make the final contour and it therefore maximises the probability of producing an accurate contour. This also means that the inclusion of low performance algorithms into ATLAAS should not affect its accuracy, as all the selection process is made by the model. In addition, ATLAAS is a learning method, which can adapt to the parameters it is built on and the data used to train it. The method is built to improve with the increasing number of training cases it is given, which methods such as majority voting or STAPLE are not designed for.

The DSC was used in this study to quantify the segmentation accuracy, as it is widely used for evaluating segmentation accuracy, and provides information on the spatial conformity of the ground truth and segmented contours, combining information on the sensitivity and positive predictive value of the segmentation within a unique accuracy score. Although the model is built to include a single accuracy score, the ATLAAS framework is independent of

the output metric used and could theoretically be built with any other accuracy score. Further work could therefore focus on applying ATLAAS with other relevant accuracy metrics.

ATLAAS showed excellent accuracy across a wide range of complex phantom data and achieved a prediction of the best (Pbest) or near-best (P10) segmentation in a very large number of cases as shown in figure 5. Our data indicate that ATLAAS is a robust method that is capable of using a good segmentation algorithm when other methods fail (see figure 5).

The Wilcoxon signed-rank test showed that the distribution of DSC values obtained for ATLAAS was not significantly different overall from those obtained with AT, RG and KM2 and GCM3. This is due to the fact that these were the PET-AS methods that ATLAAS predicted and applied in many cases. The contours generated by ATLAAS are not expected to be statistically different from the contours obtained with the individual PET-AS methods, since ATLAAS is coded to use these algorithms for image segmentation. The added value of ATLAAS is the capability of predicting the best or near-best segmentation methods in a well-defined and reproducible way, thus providing optimal and robust image segmentation.

Our results also showed that ATLAAS largely outperformed all single PET-AS values on fillable phantom data but did not outperform the best PET-AS in the case of printed H&N phantom data (see figure 5 and section 3.2). This can be explained by the fact that the training dataset for ATLAAS was built using images simulated from fillable phantom templates. Although the training dataset included non-spherical target lesions, ATLAAS was not trained on highly irregular lesions, or on images with heterogeneous background, which were found in the H&N printed data. This shows precisely the influence of the training dataset on the accuracy of the predictive model. It is expected that a training dataset containing irregular ground truth data for highly irregular lesions and heterogeneous background images, more representative of clinical data may will further improve the performance of ATLAAS. This will be the aim of future work, evaluating how the complexity of the training dataset will affect the performance of ATLAAS, and taking part in collaborations using ATLAAS on clinical data from different centres.

Classification parameter NI was chosen from the set of regional grey level co-occurrence matrix which were shown by Tixier *et al* (2011) to be most promising for the prediction of response to therapy. Although such a prediction is not the aim of this study, we hypothesized that regional texture features are therefore more clinically relevant for describing tumour heterogeneity. Among all regional grey level co-occurrence texture features, we selected only one to avoid biasing the DTs towards texture, and chose NI as it was the one providing the best performance for ATLAAS. However, other texture metrics based on different calculation approaches or matrices may further increase the accuracy of ATLAAS, especially if their value can be controlled in making the training dataset. This was not the case for NI, which explains the discrepancy between clinical and training dataset NI values (see table 2).

Further work at our centre will be focused on investigating the best metric or combination of independent texture metrics to use. ATLAAS could be additionally improved by adding to the classification parameters shape metrics, which were shown by Tixier *et al* (2013) to be potentially useful for the classification of segmentation methods. The training dataset was generated using PETSTEP, a fast and flexible PET simulation tool. PETSTEP has been previously calibrated and validated for the simulation of 18F-complex FDG uptake in the H&N (Berthon *et al* 2015a). However, the current version of PETSTEP does not include axial filtering, and therefore handles all PET image slices individually, which does not allow accurate modelling of the acquisition process. In addition, Time-of-Flight correction, which was indeed used in for all scans acquired in our centre, is not currently modelled in PETSTEP. Since the quality of the ATLAAS predictive model depends largely on the quality of the training dataset we expect

that any improvement in the quality of the images simulated with PETSTEP will result in an improved performance of ATLAAS.

It is worth noting that as part of this work, we have also developed an automated procedure for building the ATLAAS training dataset. This allows building and using the model with data from any centre or any PET imaging system. Indeed ATLAAS could be built for any tracer or imaging modality (e.g. SPECT), as long as corresponding data with reference true contours are available. Furthermore, there is no limit to the number or type of PET-AS algorithms that can be included in the ATLAAS predictive segmentation model.

5. Conclusion

We have developed ATLAAS, an advanced and automatic image segmentation algorithm, based on the DT predictive modelling method. We have shown that ATLAAS can be trained to predict the best PET-AS method when the ground truth is unknown and demonstrated that ATLAAS provides robust and accurate image segmentation that can potentially have wide applicability in radiation oncology, across multiple tumour types.

This article presents a novel method for optimised segmentation of PET images which, to our knowledge is the first to combine existing segmentation algorithms into a machine learning framework. We developed ATLAAS and advanced and automatic image segmentation algorithm based on the DT predictive modelling method. We have shown that ATLAAS can be trained to predict the best PET image segmentation method when the ground truth is unknown. ATLAAS provides a robust and accurate image segmentation that can prove very useful in radiation oncology.

Acknowledgments

This work was funded by Cancer Research Wales grants No. 2476 and 7061.

References

Anbeek P, Vincken K L, van Bochove G S, van Osch M J P and van der Grond J 2005 Probabilistic segmentation of brain tissue in MR imaging *Neuroimage* **27** 795–804

Bankman I 2000 *Handbook of Medical Image Processing and Analysis. Part II. Segmentation* ed I Bankman (Baltimore, MD: Academic)

Belhassen S and Zaidi H 2010 A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET *Med. Phys.* **37** 1309–24

Berthon B, Häggström I, Apte A, Beattie B J, Kirov A S, Humm J L, Marshall C, Spezi E, Larsson A and Schmidtlein C R 2014a PETSTEP: a fast positron emission tomography simulator for synthetic lesion simulation *Eur. J. Nucl. Med. Mol. Imaging* **41** S367

Berthon B, Häggström I, Apte A, Beattie B J, Kirov A S, Humm J L, Marshall C, Spezi E, Larsson A and Schmidtlein C R 2015a PETSTEP: generation of synthetic PET lesions for fast evaluation of segmentation methods *Phys. Medica* **31** 969–80

Berthon B, Marshall C, Edwards A, Evans M and Spezi E 2013 Influence of cold walls on PET image quantification and volume segmentation *Med. Phys.* **40** 1–13

Berthon B, Marshall C, Evans M and Spezi E 2014b Evaluation of advanced automatic PET segmentation methods using non-spherical thin-wall inserts *Med. Phys.* **41** 022502

Berthon B, Marshall C, Holmes R B and Spezi E 2015b A novel phantom technique for evaluating the performance of PET auto-segmentation methods in delineating heterogeneous and irregular lesions *Eur. J. Nucl. Med. Mol. Imaging Phys.* **2** 13

Blodgett T M, Fukui M B, Snyderman C H, Branstetter B F, McCook B M, Dave W and Meltzer C C 2005 Combined PET-CT in the head and neck. Part 1. Physiologic, altered physiologic and artifactual FDG uptake *Radiographics* **25** 897–912

Breen S L *et al* 2007 Intraobserver and interobserver variability in GTV delineation on FDG-PET-CT images of head and neck cancers *Int. J. Radiat. Oncol. Biol. Phys.* **68** 763–70

Day E, Betler J, Parda D, Reitz B, Kirichenko A, Mohammadi S and Miften M 2009 A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients *Med. Phys.* **36** 4349–58

Deasy J O, Blanco A I and Clark V H 2003 CERR: a computational environment for radiotherapy research *Med. Phys.* **30** 979–85

Dewalle-Vignion A-S, Betrouni N, Baillet C and Vermandel M 2015 Is STAPLE algorithm confident to assess segmentation methods in PET imaging? *Phys. Med. Biol.* **60** 9473–91

Dice L 1945 Measures of the amount of ecologic association between species *Ecology* **26** 297–302

Drever L, Roa W, McEwan A and Robinson D 2007 Iterative threshold segmentation for PET target volume delineation *Med. Phys.* **34** 1253–65

Ford E C, Kinahan P E, Hanlon L, Alessio A, Rajendran J, Schwartz D L and Phillips M 2006 Tumor delineation using PET in head and neck cancers: threshold contouring and lesion volumes *Med. Phys.* **33** 4280–8

Geets X, Lee J A, Bol A, Lonneux M and Grégoire V 2007 A gradient-based method for segmenting FDG-PET images: methodology and validation *Eur. J. Nucl. Med. Mol. Imaging* **34** 1427–38

Grégoire V and Chiti A 2011 Molecular imaging in radiotherapy planning for head and neck tumors *J. Nucl. Med.* **52** 331–4

Haberkorn U, Strauss L G, Reisser C, Haag D, Dimitrakopoulou A, Ziegler S, Oberdorfer F, Rudat V and van Kaick G 1991 Glucose uptake, perfusion, and cell proliferation in head and neck tumors: relation of positron emission tomography to flow cytometry *J. Nucl. Med.* **32** 1548–55

Haralick R M, Shanmugam K and Dinstein I 1973 Textural features for image classification *IEEE Trans. Syst. Man Cybern.* **SMC-3** 610–21

Hassanien A E and Kim T 2012 Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks *J. Appl. Log.* **10** 277–84

Hatt M, Cheze C, Turzo A, Roux C and Oncology I T 2009 A fuzzy locally advanced bayesian segmentation approach for volume determination in PET *IEEE Trans. Med. Imaging* **28** 881–93

Henriksson E, Kjellen E, Wahlberg P, Ohlsson T, Wennerberg J and Brun E 2007 2-Deoxy-2-[18F] fluoro-D-glucose uptake and correlation to intratumoral heterogeneity *Anticancer Res.* **27** 2155–9

Iordanescu G, Venkatasubramanian P N and Wyrwicz A M 2012 Automatic segmentation of amyloid plaques in MR images using unsupervised support vector machines *Magn. Reson. Med.* **67** 1794–802

Jayachandran A and Dhanasekaran R 2013 Brain tumor detection and classification of MR Images using texture features and fuzzy SVM classifier *Res. J. Appl. Sci. Eng. Technol.* **6** 2264–9

Jentzen W, Freudenberg L, Eising E G, Heinze M, Brandau W and Bockisch A 2007 Segmentation of PET volumes by iterative image thresholding *J. Nucl. Med.* **48** 108–14

Lyksborg M, Larsen R, Soelberg S, Per Blinkenberg M, Garde E, Siebner H R and Dyrby T B 2012 Segmenting multiple sclerosis lesions using a spatially constrained K-nearest neighbour approach *Image Analysis and Recognition* (Berlin: Springer) vol 7325 pp 156–63

McGurk R J, Bowsher J, Lee J A and Das S K 2013 Combining multiple FDG-PET radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods *Med. Phys.* **40** 042501

Reyes-Aldasoro C C 2000 Image segmentation with kohonen neural network *Int. Conf. on Telecommunications* (Acapulco: IEEE)

Schaefer A, Vermandel M and Baillet C 2015 Impact of consensus contours from multiple PET segmentation methods on the accuracy of functional volume delineation *Eur. J. Nucl. Med. Mol. Imaging* **43** 911–24

Shepherd T, Berthon B, Galavis P, Spezi E, Apte A, Lee J A, Visvikis D, Hatt M, De Bernardi E, Das S, El Naqa I and Nestle U 2012 Design of a benchmark platform for evaluating PET-based contouring accuracy in oncology applications *Eur. J. Nucl. Med. Mol. Imaging* **39** S264

Steenbakkers R J *et al* 2006 Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis *Int. J. Radiat. Oncol. Biol. Phys.* **64** 435–48

Tabakov M and Kozak P 2014 Segmentation of histopathology HER2/neu images with fuzzy decision tree and Takagi-Sugeno reasoning *Comput. Biol. Med.* **49** 19–29

Tixier F, Hatt M, Valla C, Visvikis D and Cheze le Rest C 2013 Shape indices derived from baseline 18F-FDG PET images can predict response to concomitant radio-chemotherapy in esophageal cancer *Eur. J. Nucl. Med. Mol. Imaging* **40** S210

Tixier F, Le Rest C C, Hatt M, Albarghach N, Pradier O, Metges J-P, Corcos L and Visvikis D 2011 Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer *J. Nucl. Med.* **52** 369–78

Tylski P, Bonniaud G, Decencière E, Stawiaski J, Coulot J and Lefkopoulos D 2006 F-FDG PET images segmentation using morphological watershed: a phantom study *IEEE Nuclear Science Symp. Conf. Rec. 2006* vol 4 (IEEE) pp 2063–7

Warfield S K, Zou K H and Wells W M 2004 Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation *IEEE Trans. Med. Imaging* **23** 903–21

Zaidi H, Abdoli M, Fuentes C L and El Naqa I M 2012 Comparative methods for PET image segmentation in pharyngolaryngeal squamous cell carcinoma *Eur. J. Nucl. Med. Mol. Imaging* **39** 881–91

Zaidi H and 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** 2165–87