CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/121958/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Parkinson, Craig , Evans, Mererid, Guerrero-Urbano, Teresa, Michaelidou, Andriana, Pike, Lucy, Barrington, Sally, Jayaprakasam, Vetri, Rackley, Thomas, Palaniappan, Nachi, Staffurth, John , Marshall, Christopher and Spezi, Emiliano 2019. Machine-learned target volume delineation of 18F-FDG PET images after one cycle of induction chemotherapy. Physica Medica, European Journal of Medical Physics 61 , pp. 85-93. 10.1016/j.ejmp.2019.04.020

Publishers page: http://dx.doi.org/10.1016/j.ejmp.2019.04.020

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

Machine-learned target volume delineation of 18F-FDG PET images after one cycle of induction chemotherapy

Craig Parkinson[a], Mererid Evans[b], Teresa Guerrero-Urbano[c], Andriana Michaelidou[c], Lucy Pike[e], Sally Barrington[e], Vetri Jayaprakasam[f], Thomas Rackley[b], Nachi Palaniappan[b], John Staffurth[b,f], Christopher Marshall[g], Emiliano Spezi[a,b]

[a]School of Engineering, Cardiff University, Queen's Buildings, 14-17 The Parade, Cardiff, CF24 3AA, UK

[b]Velindre Cancer Centre, Velindre Rd, Cardiff, CF14 2TL, UK

[c]Clinical Oncology, Guy's & St Thomas' NHS Foundation Trust, London, UK

[e]King's College London and Guy's and St Thomas' PET Centre, School of Biomedical Engineering and Imaging Sciences, King's College London, King's Health Partners, London, UK

[f]School of Medicine, UHW Main Building, Heath Park, Cardiff, CF14 4XN

[g]Wales Research & Diagnostic PET Imaging Centre, Cardiff University, School of Medicine, Ground Floor, C Block, UHW Main Building, Heath Park, Cardiff, CF14 4XN, UK

Corresponding Author:

Craig Parkinson (parkinsonc3@cardiff.ac.uk)

S3.17d, School of Engineering, Cardiff University, Queen's Buildings, 14-17 The Parade, Cardiff, CF24 3AA, UK

Abstract*—Biological tumour volume (GTV_{PET}) delineation on ^{18}F-FDG PET acquired during induction chemotherapy (ICT) is challenging due to the reduced metabolic uptake and volume of the GTV_{PET}. Automatic segmentation algorithms applied to ^{18}F-FDG PET (PET-AS) imaging have been used for GTV_{PET} delineation on ^{18}F-FDG PET imaging acquired before ICT. However, their role has not been investigated in ^{18}F-FDG PET imaging acquired after ICT. In this study we investigate PET-AS techniques, including ATLAAS a machine learned method, for accurate delineation of the GTV_{PET} after ICT. Twenty patients were enrolled onto a prospective phase I study (FiGaRO). PET/CT imaging was acquired at baseline and 3 weeks following 1 cycle of induction chemotherapy. The GTV_{PET} was manually delineated by a nuclear medicine physician and clinical oncologist. The resulting GTV_{PET} was used as the reference contour. The ATLAAS original statistical model was expanded to include images of reduced metabolic activity and the ATLAAS algorithm was re-trained on the new reference dataset. Estimated GTV_{PET} contours were derived using sixteen PET-AS methods and compared to the GTV_{PET} using the Dice Similarity Coefficient (DSC). The mean DSC for ATLAAS, 60% Peak Thresholding (PT60), Adaptive Thresholding (AT) and Watershed Thresholding (WT) was 0.72, 0.61, 0.63 and 0.60 respectively. The GTV_{PET} generated by ATLAAS compared favourably with manually delineated volumes and in comparison, to other PET-AS methods, was more accurate for GTV_{PET} delineation after ICT. ATLAAS would be a feasible method to reduce inter-observer variability in multi-centre trials.

***Index Terms*—Automated Segmentation, Head & Neck, Machine learning, PET/CT, Target delineation**

I. INTRODUCTION

Head and neck cancer (H&N) is the sixth most common tumour worldwide [1]. In the UK, H&N cancers account for 3% of all new cases and the rate of incidence has increased by 30% since the early 1990s [2]. Over the last 10-20 years, Human Papillomavirus (HPV) has become an increasing cause of a subset of oropharyngeal squamous cell carcinomas (OPSCC) in the H&N [3]. Radiation therapy (RT) is often used to treat OPSCC and fluorine-18 fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) aids target volume delineation (TVD) in RT planning. ^{18}F-FDG PET discriminates the biological tumour volume (GTV_{PET}) from healthy tissue with higher sensitivity and specificity compared to conventional anatomical imaging [4–11]. TVD is usually performed manually [6,7] and thus is prone to inter and intra-observer variability [12]. However, multidisciplinary approaches to TVD in H&N cancer have been shown to improve TVD accuracy. A multi-disciplinary TVD approach in H&N consists of a radiation oncologist (clinical oncologist, UK) , a neuro-radiologist, and if required, a head-and-neck surgeon [13]. However, it is time-consuming [14–16].

Inaccuracies during TVD can lead to higher organ at risk (OAR) doses and increased toxicity [17], and to tumour under-dose. This is critical for intensity-modulated radiotherapy (IMRT), which is characterised by steep dose gradients, further decreasing the margin for error [18]. Combined these errors could increase the rates of clinically significant geographical miss [18–20], local [19] and local regional recurrence [18].

A variety of PET automated segmentation (PET-AS) methods have been proposed to reduce variability in TVD [21]. Standardised uptake value max (SUV_{max}) thresholding is commonly used, especially in H&N, lung and cervical cancer [6]. SUV_{max} and SUV_{peak} are considered to be suitable for GTV_{PET} auto-segmentation as they are less dependent upon accurate volume definition than other PET measures [22].

More complex PET-AS such as adaptive iterative thresholding (AT) [23,24] have been shown to correlate with the GTV_{PET}. Other methods have been shown to be promising including region growing (RG) [25], Watershed Transform (WT) [26,27] and clustering techniques, such as Fuzzy C-means (FCM) [28], Gaussian Mixture Model based Fuzzy C-means (GCM) as described by Hatt et al [29] and K-means (KM) [6]. However, current recommendations state that no validated quantitative approach for PET contouring will result in idealistic TVD in all cases [30] and that no single PET-AS method can be recommended for TVD [22].

Machine learned PET-AS methods show promise for accurate TVD. However, comparing machine learned PET-AS is challenging due to the lack of a standardised validation imaging dataset [22]. A PET-AS methodology called Automatic decision Tree-based Learning Algorithm for Advanced Image Segmentation (ATLAAS) is an algorithm designed to select the most accurate PET-AS technique for a given PET image [31]. ATLAAS has been evaluated on pre-treatment ^{18}F-FDG PET scans [32]. However, there is increasing use of multimodality therapy including surgery and chemotherapy [33] and this requires expanding the original statistical model and re-training the ATLAAS algorithm on a new reference dataset.

Typically, RT planning is performed on CT imaging acquired before RT. Induction chemotherapy (ICT) before RT can lead to tumour downstaging and reduced GTV_{PET} volume [34]. Therefore, tumour shape, size and volume

delineated on pre-ICT imaging may not be accurately represented when RT treatment starts. RT planning on PET/CT imaging acquired during ICT could improve the accuracy of RT planning. However, TVD during treatment with chemotherapy and/or radiotherapy is challenging due to the reduced metabolic activity and GTV_{PET} volume. The aims of this study were to evaluate the feasibility of re-training a machine learned PET-AS method and assess the performance of PET-AS TVD on ^{18}F-FDG PET imaging acquired after one cycle of ICT.

II. Materials and Methods

A. Clinical Data

An ongoing phase I, multi-centre, feasibility trial called FiGaRO (Chief Investigator: TGU) involving Guy's and St Thomas' Hospitals (London, UK) and Velindre Cancer Centre (Cardiff, UK) is investigating dose escalation to the residual ^{18}F-FDG signal following 1 cycle of Cisplatin and 5-Fluorouracil (5FU) chemotherapy, in patients with primary OPSCC. The aim of the study is to improve tumour control rates whilst delivering acceptable toxicity levels. Ethical approval for the trial by the research ethics committee was granted in July 2012 (REC: 12/LO/1724). Written informed consent was obtained from all patients.

Twenty-three patients were enrolled between October 2013 and March 2017. One patient was excluded as ICT resulted in a GTV_{PET} volume too small for effective dose escalation. Two other patients were excluded due to technical and unrelated medical problems. Twenty patients with OPSCC proceeded to have dose-escalated radiotherapy. Of the twenty scans, one scan had two separate GTV_{PET} volumes, providing twenty-one GTV_{PET} volumes for analysis. Patients had histologically confirmed OPSCC, assessed as either HPV negative by p16 immunohistochemistry and in-situ hybridization for high-risk subtype DNA, or 'intermediate' or 'high' risk HPV positive, defined as having a greater than 10 pack/year smoking history and advanced N stage (TNM v7 N2b, N2c, N3) [3]. All patients were over 18 years old, staged with at least T2 tumours and planned for treatment with 2 cycles of ICT followed by primary radical IMRT to the primary and bilateral neck nodes, with concurrent Cisplatin chemotherapy (chemo-IMRT). Exclusion criteria included previous radiotherapy to the H&N region, previous malignancy except for non-melanoma skin cancer, as well as previous or concurrent illness which could interfere with the completion of the radiotherapy plan or follow-up. The clinical pathway for recruited patients is shown in Supplementary data.

B. ^{18}F-FDG PET/CT imaging

Patients underwent a planning ^{18}F-FDG/CT scan in the treatment position using a H&N immobilisation shell, 3 weeks following the first cycle of ICT. Planning ^{18}F-FDG/CT scan consisted of a low dose CT for attenuation correction, then the ^{18}F-FDG-PET, followed immediately by a contrast-enhanced CT. Scans were acquired with a 3 mm slice thickness. Patients were injected with 350+/- 10% MBq of ^{18}F-FDG and rested for 90 minutes before PET/CT scanning (GE Discovery 710 and GE Discovery 690, General Electric Medical Systems, Waukesham WI). The PET scan was acquired with a field of view of 700 mm and a matrix of 256 x 256, resulting in a voxel dimension of 2.7 mm x 2.7 mm with a slice thickness of 3.2 mm. PET imaging was acquired using 4 minutes per bed position, with 3 beds in total. PET data were reconstructed with the ordered subset expectation maximisation (OSEM) reconstruction algorithm using time of flight information with 2 iterations and 24 subsets. A 6.4 mm full

width at half maximum (FWHM) Gaussian post filter was applied to the data. Point spread function modelling techniques were not used.

C. Manual GTV_{PET} delineation after 1 cycle of chemotherapy

Manual GTV_{PET} delineation in this study was performed by a nuclear medicine physician and a radiation oncologist (clinical oncologist, UK). Delineation of the GTV_{PET} was performed independently by each viewer using an TVD protocol. The initial GTV_{PET} delineation was guided by a seed-based method and differences in delineation were resolved by an intermediary clinician. PET images for GTV_{PET} TVD were displayed in SUV and visualised using an inverse linear colour scale with a fixed window level of SUV 0 – 10. Software used for GTV_{PET} delineation at Guy's & St Thomas' PET Centre was Hermes Hybrid Viewer (Hermes Medical Solutions, Sweden) versions 2.2C and 2.6H and at Velindre Cancer Centre was Velocity AI version 2.7 (Varian Medical Systems, Palo Alto, USA) and ProSoma version 3.2 (OSL Oncology Systems Limited, UK). The GTV_{PET} was used for RT planning and as the reference GTV in subsequent analyses. Whilst, consensus techniques can be used to reduce intra and inter-observer variability, GTV_{PET} volumes obtained in the clinical environment cannot be considered a ground truth (GT), even with histopathological reference. GT volumes can only be obtained from phantom-based and simulated PET scans. Therefore, for these reasons and following the guidance of the Task Group 211 report comparisons of delineated GTV_{PET} volumes were performed using Dice Similarity Coefficient (DSC) as DSC does not favour one delineation over the other [22]. Additionally, the Sensitivity and Positive Predictive Value (PPV) are also reported. DSC is calculated as twice the intersection of two defined GTV_{PET}'s divided by the union of the two defined GTV_{PET} volumes. A result of 1 means the delineated volumes are identical and a result of 0 means a complete volume mismatch. Sensitivity is calculated as the number of true positives divided by the sum of the true positives and the false negatives. A result of 1 means the PET-AS identifies all voxels as tumour, whereas a result of 0 means the PET-AS identifies no voxels as tumour. PPV is calculated as the intersection of the reference GTV_{PET} and the comparison GTV_{PET} volume divided by the comparison GTV_{PET} volumes. A result of 1 means the PET-AS method only identifies tumour voxels as tumour and a result of 0 means the PET-AS method does not identify tumour voxels. Tumour characteristics extracted from the manually defined GTV_{PET}'s are presented in Supplementary Data (A).

D. Development of $ATLAAS_{ICT}$

ATLAAS is a machine learned, decision tree-based, PET-AS methodology, optimised with a training dataset for which the segmentation outcome is known, for optimal performance in new clinical cases in which the outcome is not known. ATLAAS has been previously described in detail [31] and is incorporated into the Computational Environment for Radiotherapy Research (CERR) [35].

The ATLAAS statistical model was originally developed on pre-treatment H&N ^{18}F-FDG PET imaging data and is built using the following image and tumour parameters: tumour to background ratio (TBR), metabolic tumour volume (MTV) and the Number of discrete Intensities (NI) within the tumour volume. NI is therefore a measure of tumour homogeneity. Hyper-parameters are parameters which are defined by the user, which can inform the training of a statistical model. Potential hyper-parameters for the ATLAAS statistical model include lymph node

size, the number of involved lymph nodes, the total number of distant metastases as well as patient characteristics including weight and age. However, hyper-parameters are not used in the development of the statistical model as the ATLAAS segmentation methodology is designed to be applied to PET images with limited user interaction and with no prior knowledge of the PET scan other than primary tumour location. The training database was built using the PETSTEP simulator [36]. PETSTEP uses a CT image and a map of the FDG uptake to generate a PET image from tumour contours that can be drawn by the user, regular in shape (e.g. spheroids) or automatically generated. The ATLAAS database was generated based upon existing PET/CT data of a fillable phantom with target tumour objects covering a range of differing characteristics working in the setting of diagnostic PET scans. The ATLAAS training database consists of a total of 100 regular spherical tumour objects with volumes and maximum uptake values in the range of 0.5 mL – 50 mL and 4000 Bq mL^{-1} – 40000 Bq mL^{-1} [31] ($ATLAAS_{ORIG}$). The mean TBR [range] of the $ATLAAS_{ORIG}$ training dataset was 3.37 [1.55 – 4.78]. The $ATLAAS_{ORIG}$ training dataset was reconstructed using optimised subset expectation maximization reconstruction using 2 iterations and 24 subsets with a point spread function of 4.9.

Table I shows that ICT reduced the MTV, SUVmax, TBR and NI values when contoured using 42% of the SUV_{max} fixed thresholding and a one-tailed T-test showed a significant reduction ($P < 0.05$) in the extracted characteristics after ICT. A threshold of 42% of the SUVmax was chosen to standardise extraction of the tumour characteristics on the pre-ICT and post-ICT PET imaging, as manually defined GTV_{PET} volumes were only delineated on post-ICT PET imaging, which was used for RT planning. Following a previously published experiment [31,32], a new statistical model was developed by simulating an additional set of 100 synthetic target tumour objects with MTV's, NI and TBR covering the range of values obtained from 10 FiGaRO clinical scans, with GTV_{PET} contours used as a basis for target tumour simulation. The simulation process has been described previously [36]. Resulting synthetic PET scans had a matrix resolution of 256 x 256 with a voxel size of 2.7 mm x 2.7 mm and a slice thickness of 3.3 mm. Simulated PET scans were reconstructed using optimised subset expectation maximization reconstruction using 2 iterations and 24 subsets with a point spread function of 4.9. These reconstruction parameters were chosen to match the $ATLAAS_{ORIG}$ training dataset, whilst demonstrating ATLAAS'S robustness to PET imaging acquired from different centres. Simulated PET scans were combined with the $ATLAAS_{ORIG}$ training dataset to create the $ATLAAS_{ICT}$ dataset. FiGaRO clinical scans were used for tumour characteristic acquisition and as a basis for PET simulation only and were not incorporated into the $ATLAAS_{ICT}$ training dataset. These ranges were 1.59-21.25 mL, 28-63 and 0.57-3.5 for MTV, NI and TBR respectively. Figure 1 shows the range of MTV, TBR and NI for the FiGaRO trial PET data and $ATLAAS_{ORIG}$ (Figures 1a, 1c and 1e) and $ATLAAS_{ICT}$ datasets (Figures 1b, 1d and 1f). The post-ICT GTV_{PET} have smaller MTV, lower TBR and lower NI than the $ATLAAS_{ORIG}$ dataset. The $ATLAAS_{ICT}$ dataset has MTV, TBR and NI values that overlap with the FiGaRO values and this was used to develop a new decision tree-based statistical model for ATLAAS. This model was used to delineate the $GTV_{PET\text{-}AS}$ for our study.

TABLE I
MEAN [RANGE] MTV, TBR, SUV_{MAX} AND NI ON ^{18}F-FDG PET IMAGING ACQUIRED BEFORE AND AFTER ICT WHEN CONTOURED USING 42% SUV_{MAX} FIXED THRESHOLDING.

Parameter	Before ICT ^{18}F-FDG PET	After ICT ^{18}F-FDG PET
Mean MTV (mL)	9.67 [2.79 – 36.18]	7.43 [3.81 – 15.11]
Mean TBR	2.16 [1.77 – 2.69]	1.79 [1.32-2.31]
Mean SUV_{max}	16.05 [6.96 – 32.96]	10.93 [4.73 – 25]
Mean NI	59.75 [45 – 65]	54.38 [63-42]

E. Automatic $GTV_{PET\text{-}AS}$ delineation after 1 cycle of ICT

$GTV_{PET\text{-}AS}$ was defined as the biological tissue delineated by the PET-AS methods well as the $ATLAAS_{ICT}$ statistical model. For threshold-based methods, a relative threshold ranging from 20-80%, in increments of 10%, of the SUV_{peak} (PT20-PT80) was used. AT is an adaptive iterative thresholding-based PET-AS method implemented in 3D, using background subtraction [23,24]. RG is a 3D Region-growing implementation with an automatic seed finder. In our implementation, RG stops when there is less than 5% change in the number of voxels included as tumour between one iteration and the next [25]. KM is a 3D K-means iterative clustering with custom stopping criterion [6]. KM2/KM3 use two and three clusters respectively. FCM is a 3D Fuzzy-C-means iterative clustering implementation with custom stopping criterion [28]. FCM2 uses two clusters. GCM is a 3D Gaussian Mixture Models based clustering with custom stopping criterion [37]. GCM3/GCM4 uses three and four clusters respectively. WT is a Watershed Transform-based implementation using a Sobel filter [26]. A total of 336 $GTV_{PET\text{-}AS}$ volumes were delineated, 16 for each patient.

Figure 1: The range of MTV, TBR and NI for the FiGaRO trial PET data and ATLAAS$_{ORIG}$ (a, c and e) and ATLAAS$_{ICT}$ datasets (b, d and f)

F. Statistical analysis

All GTV$_{PET-AS}$ contours were compared to the GTV$_{PET}$, using the DSC calculated using Matlab 2016b (The MathWorks Inc., Natick, Massachusetts, US). The mean and standard deviation (SD) for each metric were also calculated. Kruskal-Wallis tests were used to analyse the results. P values less than 0.05 were considered statistically significant.

III. RESULTS

The median MTV [IQR] of the reference GTV$_{PET}$'s contoured using the ^{18}F-FDG PET scans acquired after 1 cycle of chemotherapy was 4.25 [2-7.16] mL; the median SUV$_{max}$ [IQR] was 10.20 [5.23-12.59].

A. Validation of ATLAAS$_{ICT}$ PET-AS method

Figure 2 shows the mean DSC (+/- SD), Supplementary Data (B) shows the mean Sensitivity (+/- SD) and mean PPV (+/- SD) for all GTV$_{PET-AS}$ compared to the GTV$_{PET}$ delineated manually. The ATLAAS$_{ORIG}$ and ATLAAS$_{ICT}$ statistical models had a mean DSC [range] of 0.42 [0.00 – 0.80] and 0.72 [0.54 – 0.92], respectively. The mean Sensitivity [range] and PPV for ATLAAS$_{ORIG}$ was 0.59 [0.00 – 1.00] and 0.81 [0.00 – 1.00] respectively. Whereas, the mean Sensitivity [range] and PPV for ATLAAS$_{ICT}$ was 0.73 [0.37 – 1.00] and 0.80 [0.43 – 1.00]. A Kruskal-Wallis test was used to compare all of the MTVs delineated by the GTV$_{PET-AS}$ and the GTV$_{PET}$ delineated and showed a significant difference (P = 0.0003) in the MTV between at least two of the delineation techniques. Table II shows the mean MTV, SUV$_{max}$, DSC, Sensitivity and PPV of the four best performing PET-AS methods, all with a mean DSC ≥ 0.6. A Kruskal-Wallis test compared the DSC, Sensitivity and PPV obtained for ATLAAS$_{ICT}$, PT60, AT and WT. No significant difference was found between the DSC values (p = 0.07). A significant difference was found between the Sensitivity values (P = 0.02) and the PPV (p = 0.04) values. Further, SUV$_{max}$ values obtained from the ATLAAS$_{ICT}$, PT60, AT and WT volumes were compared using a Kruskal-Wallis test and no significant difference was found (p = 1). Figure 3 compares ATLAAS$_{ICT}$ derived contours with GTV$_{PET}$ volumes that were delineated manually.

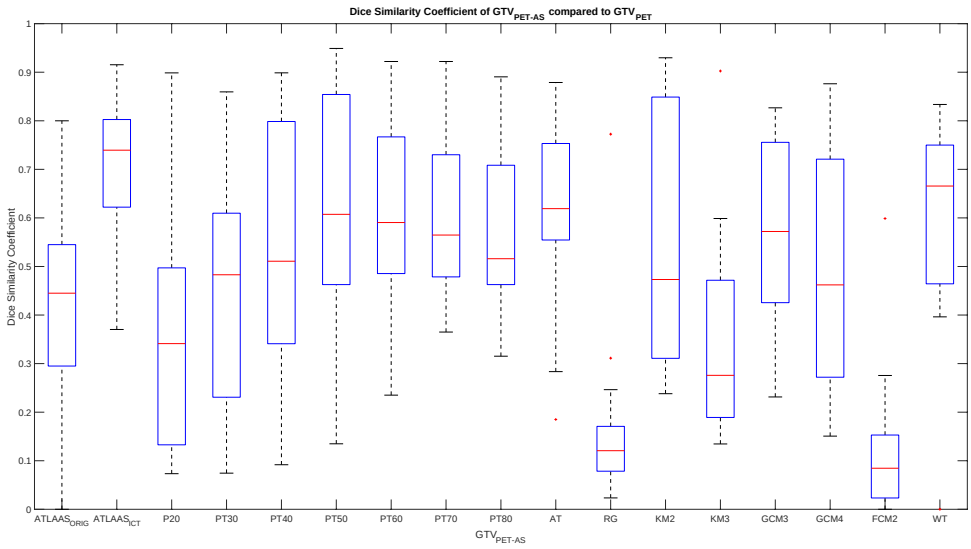

Fig 2: The DSC of GTV$_{PET-AS}$ contours compared to the GTV$_{PET}$

TABLE II
THE MEAN MTV [RANGE] AND MEAN SUV$_{MAX}$ [RANGE] OF THE BEST FOUR PET-AS METHODS

PET-AS	Mean MTV [range] mL	Mean SUV$_{MAX}$ [range]	Mean DSC (+/- STD)	Mean Sensitivity (+/- STD)	Mean PPV (+/- STD)
ATLAAS$_{ICT}$	6.01 [1.3 - 24]	10.19 [3.88 – 25.00]	0.72 (+/- 0.10)	0.73 (+/- 0.19)	0.80 (+/- 0.19)
PT60	8.66 [3.28 – 44.01]	10.17 [3.18 – 25.00]	0.61 (+/- 0.20)	0.81 (+/- 0.22)	0.64 (+/- 0.35)
AT	3.85 [1.30 – 8.75]	10.28 [4.12 – 25.00]	0.63 (+/- 0.15)	0.61 (+/- 0.22)	0.79 (+/- 0.26)
WT	7.20 [0.54 – 27.55]	10.20 [3.39 – 25.00]	0.60 (+/- 0.21)	0.75 (+/- 0.24)	0.60 (+/- 0.31)

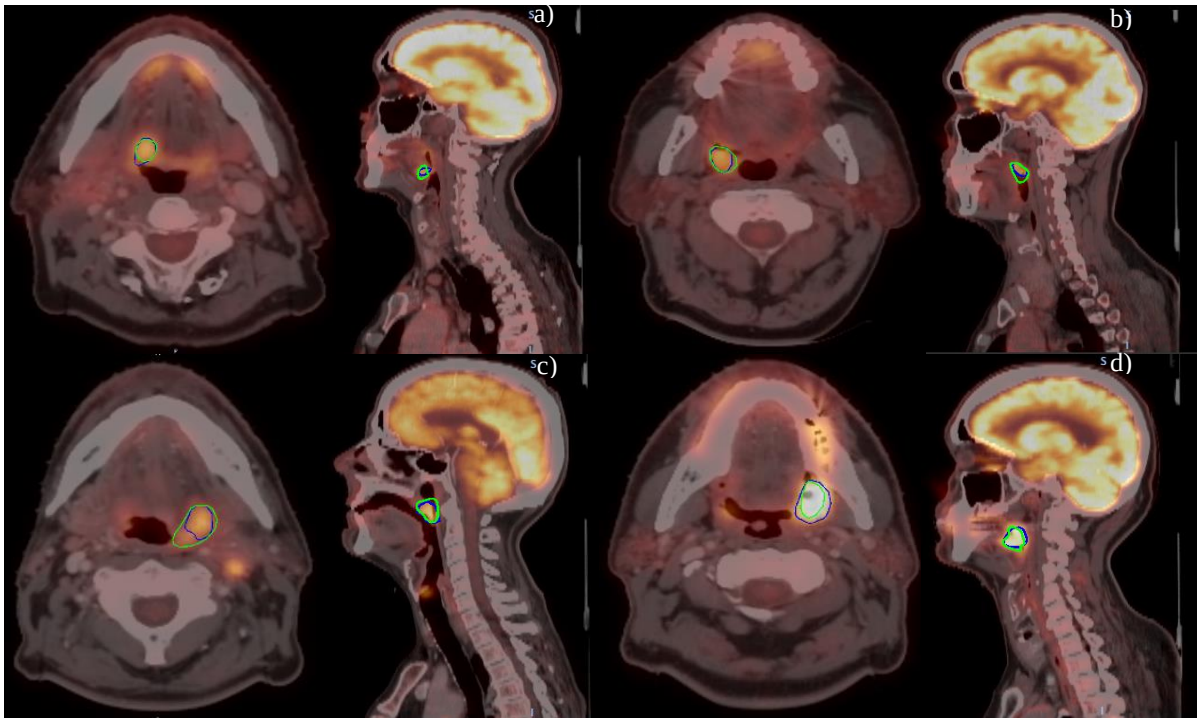

Fig 3: ATLAAS$_{ICT}$ derived contours (Blue) and the GTV$_{PET}$ contours (Green) delineated jointly and by consensus in four of the FiGaRO patients. Good agreement is seen between the two contouring methods. However, slight variations exist. A) Patient 1, B) Patient 2, C) Patient 3, D) Patient 4

B. Comparison between GTV$_{PET-AS}$ and GTV$_{PET}$

Figure 4 shows the percentage increase in GTV$_{PET-AS}$ volume obtained from ATLAAS$_{ICT}$, PT60, AT and WT PET-AS methods when compared to the GTV$_{PET}$. In this study, the four best performing PET-AS methods delineated GTV$_{PET-AS}$ smaller than the GTV$_{PET}$ in 8 patients and larger in 6 patients. In 7 patients, there was a variation with some methods outlining a larger volume and some a smaller volume. Further, in patients 2,4,8, 10 and 19, PT60 delineated larger GTV$_{PET-AS}$ volumes compared to the other PET-AS methods. A 651% and a 522% increase in patients 4 and 17 respectively. In these cases, the GTV$_{PET}$ MTV was less than or equal to 2 mL with a SUV$_{max}$ of 4.73 and 5.23 respectively, suggesting GTV$_{PET-AS}$ delineation with PT60 is limited in MTV's smaller than 2 mL. The two best performing methods (ATLAAS$_{ICT}$ and AT) agreed in the delineation of GTV$_{PET-AS}$ smaller or larger than the GTV$_{PET}$ in 19 patients. Current recommendations to reduce TVD variability are to standardise window width and levels for all patients [14,38]. In tumours with lower metabolic activity, standardised delineation parameters limit TVD due to reduced constrast compared to the background. ATLAAS$_{ICT}$

and AT operate independently of window width and levels, therefore the agreement between the two methods to delineate smaller or larger volumes in 19 of the patients may highlight additional information in RT planning.

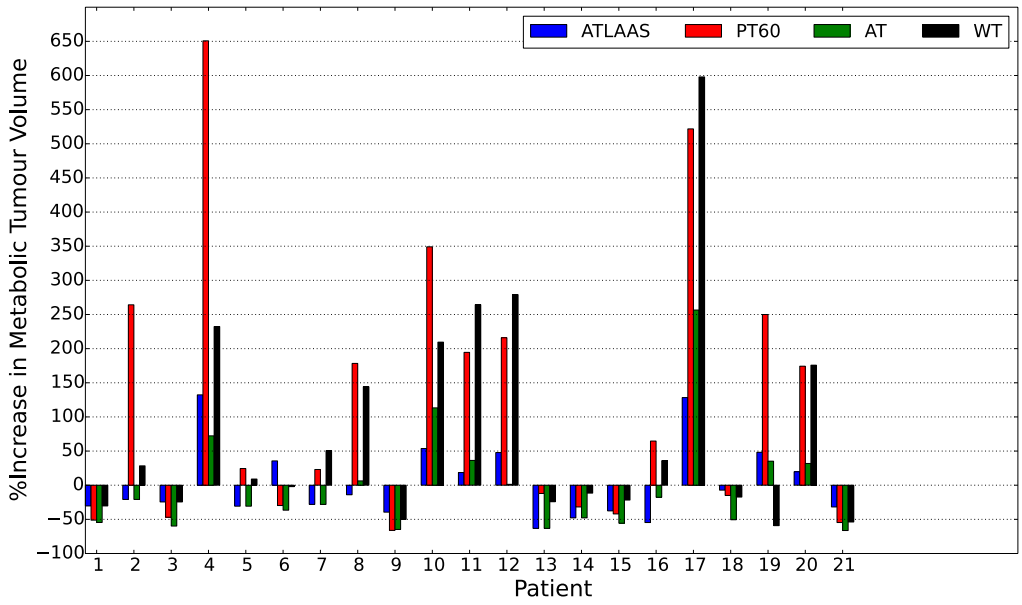

Fig 4: The percentage increase in MTV delineated by PET-AS compared to the MTV of the GTV_{PET}

IV. Discussion

This study highlights challenges in TVD with ^{18}F-FDG PET of the H&N, acquired after one cycle of ICT. We have shown that 1 cycle of ICT reduces the GTV_{PET} volume and metabolic activity, making manual and automated delineation harder. In our study, GTV_{PET} volumes defined using 42% SUV_{max} thresholding on diagnostic pre-treatment and planning intra-treatment imaging resulted in mean volumes of 9.67 mL and 7.43 mL respectively. Ranging between 2.79 mL and 36.18 mL for diagnostic pre-treatment imaging and 3.81 mL and 15.11 mL for planning intra-treatment imaging. In IMRT planning, accurate TVD is critical as errors in delineation may increase geographical miss [18–20]. ^{18}F-FDG PET discriminates biologically active tumour tissue with higher sensitivity and specificity compared to anatomical imaging and therefore, is increasingly being used for TVD [4–11]. However, manual TVD on PET/CT imaging in clinical practice is challenging [6,7] with relatively high levels of intra and inter-observer variability.

Arens et al [38] in 2013 investigated the role of semi-automated PET-AS methods for TVD in sequential FLT PET/CT imaging in H&N carcinomas and their relationship with clinical outcome. Eighteen patients were treated with 3D-conformal RT and 28 patients were treated with IMRT. In total, 46 patients with a total of 48 GTV_{PET}'s were treated with chemo-radiotherapy. FLT PET/CT imaging was acquired prior to treatment delivery (n = 46) and in the 2nd (n = 44) and 4th (n=28) weeks of treatment. In comparison, we analysed ^{18}F-FDG PET imaging data acquired from 20 patients recruited to a prospective phase I feasibility study (FiGaRO), with PET/CT imaging acquired pre-ICT delivery and post one cycle of ICT. The GTV_{PET}'s delineated by Arens et al were delineated by a singular trained radiation oncologist (clinical oncologist, UK), whereas in our study we reduced intra-observer variability with the use of a TVD protocol and joint delineation techniques. Arens et al also found that the

SUV_{PEAK} derived on sequential FLT PET/CT scans significantly decreased during treatment, and the results presented in this study affirm their findings in ^{18}F-FDG PET/CT imaging. Additionally, the GTV_{PET}'s derived by PET-AS methods on the FLT based PET/CT imaging showed significant differences between the derived volumes ($P < 0.05$) and our study produced equivalent results. The FLAB method investigated in Arens et al [38] correlated with the manually derived GTV_{PET}, whereas in our study the ATLAAS segmentation methodology was the best performing PET-AS method. A direct comparison of these two methods from the literature is impossible due to the different metrics used to assess the performance, the different scan acquisition parameters and the differing radiotracers. This highlights the requirement for a standardised evaluation dataset for assessing PET-AS method performance as highlighted by the AAPM Task Group No 211 [22] and by Berthon et al[39]. There have been several studies using deep learning to deal with PET or PET/CT segmentation, including the work from Czakon et al, that won the 2018 MICCAI PET segmentation challenge [40], based on a convolutional neural network. Work is in progress in our institution to implement similar technology in the ATLAAS segmentation workflow,

In this study, we compared twenty-one manually delineated GTV_{PET}'s with contours derived by PET-AS methods. Our results show that our machine-learned method, ATLAAS, has the potential to be trained and applied to a variety of imaging data whilst being the best performing PET-AS method. In previous work, ATLAAS has been described [31], trained upon and validated in diagnostic PET imaging [32]. However, this work demonstrates that new scans can be developed for which ATLAAS can be trained upon when the imaging characteristics of clinically obtained PET scans are significantly different from the default training dataset. Further to this, this work demonstrates that ATLAAS outperforms advanced PET-AS methods in low tumour to background ratio scenarios due to this additional training. Compared to non-machine learned PET-AS, ATLAAS had the highest DSC. Whilst, adaptive thresholding techniques were found to be more robust for accurate TVD compared to fixed thresholding techniques, the robustness of the machine-learned method ATLAAS was greater in comparison to all the other PET-AS methods included in this study, suggesting that it is an excellent candidate for ongoing multicentre studies using ^{18}F-FDG PET for radiotherapy planning.

Our results show that the performance of individual automated segmentation methodologies can be enhanced through the use of machine-learning techniques. Studies have shown how imaging parameters such as reconstruction settings [41] as well as tumour features impact TVD [31,42–44]. The ATLAAS statistical model, however, was developed using tumour characteristics which have been demonstrated to be classifiers for accurate MTV delineation. These characteristics are MTV, NI and TBR [31,43,45]. Potentially, there are parameters which could impact the accuracy of the ATLAAS segmentation methodology, including morphological features [46] and the reconstruction modality. The results in this study, however, demonstrate that without the addition of the reconstruction modality as a classifier, and without the standardisation of the PET imaging, ATLAAS outperforms all of the included PET-AS methods. The standardisation of PET imaging acquisition and reconstruction is important for the accurate quantification of PET imaging in multicenter studies [47]. It has also been proposed that the use of consensus techniques, could improve the accuracy and the robustness of TVD [48] and we have shown that machine learning techniques can be used to select the most appropriate PET-AS method for a given FDG-PET image, without the requirement for multiple delineations of the GTV_{PET}.

Our study also shows how ATLAAS is adaptable to differing situations and has been successfully trained on PET imaging data acquired after chemotherapy. This is a clear advantage of the ATLAAS machine learning-based approach compared to other PET-AS methods. Our study also supports published literature [22] in that no one single PET-AS method can be recommended for TVD after one cycle of ICT. However, machine learning based PET-AS methodologies are showing promise for the accurate TVD and based on our results, ATLAAS appears to be an excellent candidate for use in future trials.

The results of this study are limited by a relatively small cohort (n = 20) of patients, although all recruited as part of the same clinical trial in OPSCC. Patients also underwent only one cycle of chemotherapy. Given the trial design, it was not possible to evaluate the accuracy and robustness of the PET-AS methods included in this study against the remaining biological tissue after multiple cycles of chemotherapy or after fractions of radiotherapy. However, this study demonstrates the feasibility of PET-AS TVD after one cycle of ICT. Therefore, ATLAAS could be useful as a basis for treatment adaptation (e.g. dose escalation) during chemotherapy. A further study to investigate the feasibility of PET-AS TVD in adaptive radiotherapy is planned.

V. Conclusion

The ATLAAS segmentation methodology provided a more robust approach to the delineation of the GTV_{PET} after induction chemotherapy when compared to any of the individual PET-AS methods included in this study. ATLAAS is adaptable to differing situations and has been successfully trained on PET data acquired after chemotherapy and therefore could be useful as a basis for treatment adaptation (e.g. dose escalation) during chemotherapy.

VI. Acknowledgments

There are no conflicts of interest. This work was partially funded by EPSRC-DTP grant ref EP/M507842/1 and Velindre NHS Trust grant Number 2016/1. Professor Barrington acknowledges support from the National Institute of Health Research (NIHR) [RP-2-16-07-001]. King's College London and UCL Comprehensive Cancer Imaging Centre is funded by the CRUK and EPSRC in association with the MRC and Department of Health (England). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

VII. References

[1] Vigneswaran N, Williams MD. Epidemiologic trends in head and neck cancer and aids in diagnosis. Oral Maxillofac Surg Clin North Am 2014;26:123–41. doi:10.1016/j.coms.2014.01.001.

[2] Oral cancer statistics | Cancer Research UK n.d. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oral-cancer#heading-One (accessed April 6, 2017).

[3] Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, et al. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. N Engl J Med 2010;363:24–35. doi:10.1056/NEJMoa0912217.

[4] Bar-Shalom R, Yefremov N, Guralnik L, Gaitini D, Frenkel A, Kuten A, et al. Clinical performance of PET/CT in evaluation of cancer: additional value for diagnostic imaging and patient management. J Nucl Med 2003;44:1200–9.

[5] Daisne J-F, Duprez T, Weynand B, Lonneux M, Hamoir M, Reychler H, et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology 2004;233:93–100. doi:10.1148/radiol.2331030660.

[6] Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: A survey of image segmentation techniques. Eur J Nucl Med Mol Imaging 2010;37:2165–87. doi:10.1007/s00259-010-1423-3.

[7] Ypsilantis P-P, Siddique M, Sohn H-M, Davies A, Cook GJR, Goh V, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. PLoS One 2015;10:e0137036. doi:10.1371/journal.pone.0137036.

[8] Niyazi M, Landrock S, Elsner A, Manapov F, Hacker M, Belka C, et al. Automated biological target volume delineation for radiotherapy treatment planning using FDG-PET/CT. Radiat Oncol 2013;8:1. doi:10.1186/1748-717X-8-180.

[9] Bradley J, Thorstad WL, Mutic S, Miller TR, Dehdashti F, Siegel BA, et al. Impact of FDG-PET on radiation therapy volume delineation in non-small-cell lung cancer. Int J Radiat Oncol Biol Phys

2004;59:78–86. doi:10.1016/j.ijrobp.2003.10.044.

[10] Ciernik IF, Dizendorf E, Baumert BG, Reiner B, Burger C, Davis JB, et al. Radiation treatment planning with an integrated positron emission and computer tomography (PET/CT): A feasibility study. Int J Radiat Oncol Biol Phys 2003;57:853–63. doi:10.1016/S0360-3016(03)00346-8.

[11] Lardinois D, Weder W, Hany TF, Kamel EM, Korom S, Seifert B, et al. Staging of non-small-cell lung cancer with integrated positron-emission tomography and computed tomography. N Engl J Med 2003;348:2500–7. doi:10.1056/NEJMoa022136.

[12] Parkinson C, Chan J, Syndikus I, Marshall C, Staffurth J, Spezi E. EP-1333: Impact of 18 F-Choline PET scan acquisition time on delineation of GTV in Prostate cancer. Radiother Oncol 2017;123:S714–5. doi:10.1016/S0167-8140(17)31768-1.

[13] Lee N, Xia P, Fischbein NJ, Akazawa P, Akazawa C, Quivey JM. Intensity-modulated radiation therapy for head-and-neck cancer: The UCSF experience focusing on target volume delineation. Int J Radiat Oncol Biol Phys 2003;57:49–60. doi:10.1016/S0360-3016(03)00405-X.

[14] Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imaging Radiat Oncol 2016;60:393–406. doi:10.1111/1754-9485.12462.

[15] Jiang J, Wu H, Huang M, Wu Y, Wang Q, Zhao J, et al. Variability of Gross Tumor Volume in Nasopharyngeal Carcinoma Using 11C-Choline and 18F-FDG PET/CT. PLoS One 2015;10:e0131801. doi:10.1371/journal.pone.0131801.

[16] Greco C, Rosenzweig K, Cascini GL, Tamburrini O. Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC). Lung Cancer 2007;57:125–34. doi:10.1016/j.lungcan.2007.03.020.

[17] Corvò R. Evidence-based radiation oncology in head and neck squamous cell carcinoma. Radiother Oncol 2007;85:156–70. doi:10.1016/j.radonc.2007.04.002.

[18] Chen AM, Chin R, Beron P, Yoshizaki T, Mikaeilian AG, Cao M. Inadequate target volume delineation and local–regional recurrence after intensity-modulated radiotherapy for human papillomavirus-positive oropharynx cancer: Local–regional recurrence after IMRT for HPV-positive oropharynx cancer. Radiother Oncol 2017;123:412–8. doi:10.1016/j.radonc.2017.04.015.

[19] Eisbruch A, Marsh LH, Dawson LA, Bradford CR, Teknos TN, Chepeha DB, et al. Recurrences near base of skull after IMRT for head-and-neck cancer: Implications for target delineation in high neck and for parotid gland sparing. Int J Radiat Oncol Biol Phys 2004;59:28–42. doi:10.1016/j.ijrobp.2003.10.032.

[20] Dawson LA, Anzai Y, Marsh L, Martel MK, Paulino A, Ship JA, et al. Patterns of local-regional

recurrence following parotid-sparing conformal and segmental intensity-modulated radiotherapy for head and neck cancer. Int J Radiat Oncol Biol Phys 2000;46:1117–26. doi:10.1016/S0360-3016(99)00550-7.

[21] Schaefer A, Vermandel M, Baillet C, Dewalle-Vignion AS, Modzelewski R, Vera P, et al. Impact of consensus contours from multiple PET segmentation methods on the accuracy of functional volume delineation. Eur J Nucl Med Mol Imaging 2016;43:911–24. doi:10.1007/s00259-015-3239-7.

[22] Hatt M, Lee JA, Schmidtlein CR, El Naqa I, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. Med Phys 2017;44:e1–42. doi:10.1002/mp.12124.

[23] Drever L, Roa W, McEwan A, Robinson D. Iterative threshold segmentation for PET target volume delineation. Med Phys 2007;34:1253–65. doi:10.1118/1.2712043.

[24] Jentzen W, Freudenberg L, Eising EG, Heinze M, Brandau W, Bockisch A. Segmentation of PET volumes by iterative image thresholding. J Nucl Med 2007;48:108–14.

[25] Day E, Betler J, Parda D, Reitz B, Kirichenko A, Mohammadi S, et al. A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients. Med Phys 2009;36:4349–58. doi:10.1118/1.3213099.

[26] Geets X, Lee JA, Bol A, Lonneux M, Grégoire V. A gradient-based method for segmenting FDG-PET images: Methodology and validation. Eur J Nucl Med Mol Imaging 2007;34:1427–38. doi:10.1007/s00259-006-0363-4.

[27] Tylski P, Bonniaud G, Decenciere E, Stawiaski J, Coulot J, Lefkopoulos D, et al. 18F-FDG PET images segmentation using morphological watershed: a phantom study. 2006 IEEE Nucl. Sci. Symp. Conf. Rec., vol. 4, IEEE; 2006, p. 2063–7. doi:10.1109/NSSMIC.2006.354319.

[28] Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. Med Phys 2010;37:1309–24. doi:10.1118/1.3301610.

[29] Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. Eur J Nucl Med Mol Imaging 2011;38:663–72. doi:10.1007/s00259-010-1688-6.

[30] Konert T, Vogel W, MacManus MP, Nestle U, Belderbos J, Grégoire V, et al. PET/CT imaging for target volume delineation in curative intent radiotherapy of non-small cell lung cancer: IAEA consensus report 2014. Radiother Oncol 2015;116:27–34. doi:10.1016/j.radonc.2015.03.014.

[31] Berthon B, Marshall C, Evans M, Spezi E. ATLAAS: An automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. Phys Med Biol 2016;61:4855–69. doi:10.1088/0031-9155/61/13/4855.

[32] Berthon B, Evans M, Marshall C, Palaniappan N, Cole N, Jayaprakasam V, et al. Head and neck target delineation using a novel PET automatic segmentation algorithm. Radiother Oncol 2017;122:242–7. doi:10.1016/j.radonc.2016.12.008.

[33] Welsh L, Panek R, McQuaid D, Dunlop A, Schmidt M, Riddell A, et al. Prospective, longitudinal, multi-modal functional imaging for radical chemo-IMRT treatment of locally advanced head and neck cancer: the INSIGHT study. Radiat Oncol 2015;10:112. doi:10.1186/s13014-015-0415-7.

[34] Crandall JP, Tahari AK, Juergens RA, Brahmer JR, Rudin CM, Esposito G, et al. A comparison of FLT to FDG PET/CT in the early assessment of chemotherapy response in stages IB–IIIA resectable NSCLC. EJNMMI Res 2017;7:8. doi:10.1186/s13550-017-0258-3.

[35] Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys 2003;30:979–85. doi:10.1118/1.1568978.

[36] Berthon B, Häggström I, Apte A, Beattie BJ, Kirov AS, Humm JL, et al. PETSTEP: Generation of synthetic PET lesions for fast evaluation of segmentation methods. Phys Medica 2015;31:969–80. doi:10.1016/j.ejmp.2015.07.139.

[37] Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A Fuzzy Locally Adaptive Bayesian Segmentation Approach for Volume Determination in PET. IEEE Trans Med Imaging 2009;28:881–93. doi:10.1109/TMI.2008.2012036.

[38] Arens AIJ, Troost EGC, Hoeben BAW, Grootjans W, Lee JA, Grégoire V, et al. Semiautomatic methods for segmentation of the proliferative tumour volume on sequential FLT PET/CT images in head and neck carcinomas and their relation to clinical outcome. Eur J Nucl Med Mol Imaging 2014;41:915–24. doi:10.1007/s00259-013-2651-0.

[39] Berthon B, Spezi E, Galavis P, Shepherd T, Apte A, Hatt M, et al. Toward a standard for the evaluation of PET-Auto-Segmentation methods following the recommendations of AAPM task group No. 211: Requirements and implementation. Med Phys 2017;44:4098–111. doi:10.1002/mp.12312.

[40] Fayad H, Camarasu-Pop S, Tauber C, Ouahabi A, Kain M, Tan S, et al. The first MICCAI challenge on PET tumor segmentation. Med Image Anal 2017;44:177–95. doi:10.1016/j.media.2017.12.007.

[41] Cheebsumon P, Yaqub M, Van Velden FHP, Hoekstra OS, Lammertsma AA, Boellaard R. Impact of [18F]FDG PET imaging parameters on automatic tumour delineation: Need for improved tumour delineation methodology. Eur J Nucl Med Mol Imaging 2011;38:2136–44. doi:10.1007/s00259-011-1899-5.

[42] Xu W, Yu S, Ma Y, Liu C, Xin J. Effect of different segmentation algorithms on metabolic tumor volume measured on 18F-FDG PET/CT of cervical primary squamous cell carcinoma. Nucl Med Commun 2017;38:259–65. doi:10.1097/MNM.0000000000000641.

[43] Berthon B, Marshall C, Evans M, Spezi E. Evaluation of advanced automatic PET segmentation methods using nonspherical thin-wall inserts. Med Phys 2014;41:022502. doi:10.1118/1.4863480.

[44] Parkinson C, Foley K, Whybra P, Hills R, Roberts A, Marshall C, et al. Evaluation of prognostic models developed using standardised image features from different PET automated segmentation methods. EJNMMI Res 2018;8:29. doi:10.1186/s13550-018-0379-3.

[45] Berthon B, Marshall C, Edwards A, Evans M, Spezi E. Influence of cold walls on PET image quantification and volume segmentation: A phantom study. Med Phys 2013;40:082505. doi:10.1118/1.4813302.

[46] Parkinson C, Marshall C, Staffurth J, Spezi E. ATLAAS - Investigation into the Incorporation of Morphological Data on Automated Segmentation. Eur J Nucl Med Mol Imaging 2018;45:S72–3.

[47] Kaalep A, Sera T, Oyen W, Krause BJ, Chiti A, Liu Y, et al. EANM/EARL FDG-PET/CT accreditation - summary results from the first 200 accredited imaging systems. Eur J Nucl Med Mol Imaging 2018;45:412–22. doi:10.1007/s00259-017-3853-7.

[48] McGurk RJ, Bowsher J, Lee J a, Das SK. Combining multiple FDG-PET radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods. Med Phys 2013;40:1–9. doi:10.1118/1.4793721.