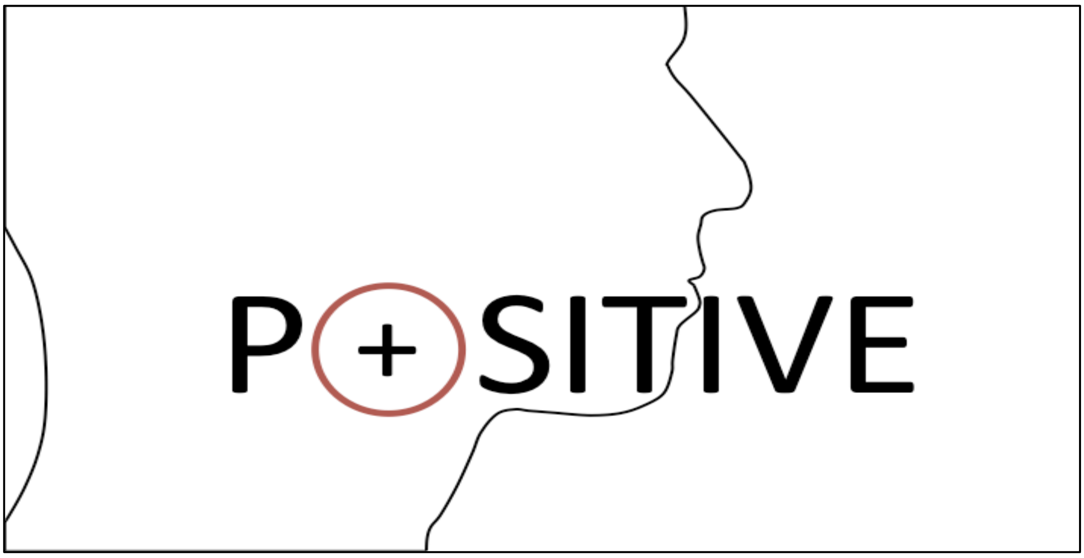


Optimisation Of Positron Emission Tomography Based Target Volume Delineation In Head And Neck Radiotherapy

Beatrice Berthon

Thesis submitted on 8 October 2014 at Cardiff University for the

award of Doctor of Philosophy

Table of Contents

Acknowledgements

The accomplishment of this project was made possible by the financial help of Cancer Research Wales (CRW), and I am very grateful to the CRW staff for their generosity and continuous support during these three years.

I am extremely grateful to my supervisors Dr. Emiliano Spezi and Dr. Chris Marshall for always being there to guide me, and reassure me, and for putting so much faith in my work. It has been a great pleasure and honour working with you and getting to know you.

I would also like to extend my thanks to the clinical team, Dr. Mererid Evans, Dr. Nachi Palaniappan, Dr. Naomi Cole and Dr. Vetri Jayaprakasam, for driving this project toward its clinical application, to Andrew Edwards for his great work on the vacuum-moulding technique and phantom equipment, and to Robin Holmes for his inspiration and help in developing the PET printed SS phantom.

None of this work would have been possible without the help and support of the whole team in PETIC, in particular Lee, Chris and Matthew. A special feeling of gratitude also to Carell, Sara and Guillaume for making PhD days so much more enjoyable. Finally, I am forever grateful to my fiancé Stuart for always being at my side and going through the long task of proofreading this thesis.

I would like to dedicate this thesis to André and Ursula, my loving parents and two of the most wonderful people I have ever known.

Summary

Automatic segmentation of tumours using Positron Emission Tomography (PET) was recommended for radiotherapy treatment (RT) planning of head and neck (H&N) cancer patients, and investigated in the scientific literature without reaching a consensus on the optimal process. This project aimed at evaluating the performance of PET-based automatic segmentation (PET-AS) methods and developing an optimal PET-AS process to be used at Velindre Cancer Centre (VCC). For this purpose, ten algorithms were implemented to represent the most promising PET-AS approaches from a systematic review of the literature. The algorithms' performance was evaluated on filled phantom inserts with variable size, geometry, tumour intensity and image noise. The impact of thick insert plastic walls on both image quantification and segmentation was thoroughly assessed. The PET-AS methods were further applied to realistic H&N tumours, modelled using a printed subresolution sandwich phantom developed and calibrated in house. Results showed that different PET-AS performed best for different types of target objects. An Advanced decision Tree-based Learning Algorithm for Automatic Segmentation (ATLAAS) was therefore developed and validated for the selection of the optimal PET-AS approach according to the target object characteristics. Finally, a protocol was designed for the use of PET-AS within RT planning at VCC. The protocol was used retrospectively on a group of 10 oropharyngeal cancer patients, and the results highlighted the additional information brought by PET beyond anatomical imaging. In a prospective study on 10 additional patients, PET-AS replaced manual PET/CT delineation, and accounted for up to 33% of the modifications of manually drawn CT/MRI contours to derive the final planning contour. This study demonstrated the usefulness and reliability of the PET-AS method in RT planning, and led to modifying the clinical workflow for H&N patients at VCC. This work has the potential to be extended to other tumour sites and institutions.

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed …………………………… (candidate) Date …………………………

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD

Signed …………………………… (candidate) Date …………………………

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.
Other sources are acknowledged by explicit references. The views expressed are my own.

Signed …………………………… (candidate) Date …………………………

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed …………………………… (candidate) Date …………………………

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.

Signed …………………………… (candidate) Date …………………………

List of figures

List of tables

List of equations

Chapter I. Introduction

I. A. Management of Head and Neck cancer

Head and Neck (H&N) cancer is one of the most common cancers in the developed countries, representing 5% of cancers worldwide, with more than 6500 diagnoses in England in 2012 [1]. H&N cancer is most commonly linked to excessive and prolonged consumption of tobacco or alcohol, but can also be caused by viruses such as the Human Papillomavirus, especially in young people [2]. H&N cancer refers to tumours originating in the anatomical regions of the oral and nasal cavity, larynx, pharynx and sinuses. It occurs most often in the oropharynx, although laryngeal and pharyngeal tumours account for more deaths [2]. The majority of H&N cancers (90%) are squamous cell carcinomas (SCCs), corresponding to the growth of the mucosal membrane called epithelium [1]. These cancers often rapidly progress, and are likely to spread to the lymph nodes of the neck, as well as other regions of the body. Tumours are classified following the recently updated TNM (Tumour, Nodes, Metastases) classification system [3], leading to different cancer management scenarios. Although efficient treatment can be provided if the disease is detected early, at present, the chances of survival remain low for advanced disease, with 50-60% survival rate for 5 years [4]. In particular, recurrence rates reach 15-50% for H&N SCC patients [5]. This is due to a number of limiting factors such as the intrinsic tumour radioresistance, the lack of accuracy of current anatomical imaging to determine the extent of disease, and the proximity of surrounding Organs At Risk (OARs) making the dose delivery challenging.

Most H&N cancer patients receive a combination of treatment, including chemotherapy, radiotherapy and surgery. In the 1990's, surgery followed by radiotherapy was considered as the standard of care for advanced stage tumours [6], and is still used for curative purposes of early stage malignancies. However, for advanced disease, the complex anatomy of the H&N makes it difficult to remove all tumour cells without damaging surrounding organs. H&N surgery is, therefore, often associated with life changing cosmetic damage for the patient, and an increased risk of deteriorating vital or essential body functions, such as speech, swallowing and breathing.

In an effort to reduce the toxicity associated with H&N cancer treatments, a small number of authors have published comparative studies showing that treatment with chemo-radiotherapy (cisplatin and fluorouracil) instead of postoperative radiotherapy, allowed organ preservation whilst maintaining local control of the disease and similar survival rates [7], [8].

A meta-analysis of randomised trials conducted between 1965 and 2000 showed a 4.5% survival benefit of the delivery of chemotherapy during the radiotherapy treatment (RT), called concomitant radio-chemotherapy (CRT), as opposed to radiotherapy alone [9], making it the new standard of care for advanced H&N SCC (stage III and IV).

A different type of chemotherapy delivery, induction chemotherapy (IC), is used for reducing the tumour bulk and reducing the microscopic disease extension before CRT or RT alone, to improve its efficacy. Although the evidence for improved overall survival when adding IC to CRT is still sparse, some randomised studies have shown improved survival rates when using additional administration of docetaxel to cisplatin/5-fluorouracil (5-FU)) CRT [10], [11]. IC is currently used in a number of centres in addition to CRT with high survival rates obtained [12].

Alternatively, biological therapies can be used to target the tumour cells and stop the progression of the disease through a number of different possible mechanisms. Cetuximab is an antibody used for the down-regulation of the epidermal growth factor receptor (EGFR), which is overexpressed in cancerous cells, leading to accelerated cell repopulation. There is already some evidence that treatment with Cetuximab for patients undergoing radical radiotherapy increases progression-free survival [13], which has made it the new standard of care alternative to platinum-based chemotherapy.

A small number of studies have focused on the identification of failure patterns within current standard treatments. Pigott *et al.* observed 97% failure within the centre of the tumour in H&N patients treated with radical radiotherapy [14]. The high probability of recurrence within the high dose treatment volume was confirmed by other more recent studies [12]. These results suggest that alternative treatment delivery protocols have the potential to increase locoregional control by targeting tumour areas with high intrinsic radioresistance. In particular, this makes the case for the definition of an additional sub-volume within the tumour for which the dose could be increased or escalated during treatment.

Tumour hypoxia is also a factor that can affect the tumour radiosensitivity. It was suggested that hypoxic regions require a boost of 120-150% to reach normal control rates. Adapting the treatment to hypoxic regions was shown to increase the overall therapeutic benefit [11]. However, there is still a lack of current evidence of improved outcomes when incorporating hypoxia information into clinical management.

I. B. Challenges in radiotherapy delivery and planning

The first curative radiation therapy treatment was applied to H&N patients as early as 1899, after the discovery of X-rays by Wilhelm Röntgen in 1896. It involves irradiating the tumour with ionizing radiation beams, causing damage to the DNA of the targeted cells, leading to the death of radiosensitive tumour cells, which cannot repair themselves.

In the H&N region, the total radiation doses required to eradicate the disease (up to 70 Gy in 35 fractions over 7 weeks) lead to significant acute and late toxicities, because of the proximity of OARs around the tumour. These can include mucositis, xerostomia, dysphagia, radiation dermatitis, pain and fatigue, and at a later stage osteoradionecrosis, skin fibrosis, all resulting in significant morbidity and reduced quality of life during RT [16]. This shows the importance of accurate treatment delivery and planning for H&N patients.

Major advances in the radiotherapy delivery over the last 2 decades have greatly improved the treatment accuracy. Previous techniques involved irradiating the neck via two parallel beams, leading to a homogeneous dose delivered to the whole neck area. The development of Intensity Modulated Radiotherapy Treatment (IMRT) using a larger number of beams (typically 5-7), coupled with a simultaneous integrated boost technique, allowed shaping of the dose delivered to the treatment area, allowing sparing of the surrounding normal tissues. The implementation of IMRT has seen a reduction in toxicities such as xerostomia, caused by the irradiation of the parotid glands during treatment, for the same overall survival rates [17].

However, accurate targeting of the tumour requires careful planning of the beams' shape and position. The time-consuming RT planning is done by dedicated software, which calculates the optimal beam arrangement, based on contour information determined by the planning clinician for the tumour and OARs.

These contours are commonly drawn using information from anatomical imaging, such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), which are sometimes combined to provide the planning contours. The definition of the target dose volume involves different contours defined in guidelines of the Report 50 of the International Commission on Radiation Units and Measurements [18]:

- Gross Target Volume (GTV), corresponding to the tumour burden,
- Clinical Treatment Volume (CTV), which is an extension of the GTV with a margin to account for possible microscopic disease extension,

- Planning Target Volume (PTV), which adds a security margin to the CTV to account for errors in the patient positioning or dose delivery.

With current technologies allowing high precision dose delivery, GTV delineation inaccuracy was recently identified as the major source of error in RT delivery [19]. As a result, there is growing interest in the use of additional information, complementary to commonly used anatomical imaging, provided by other imaging techniques. In particular, a number of clinical oncology groups have investigated the use of Positron Emission Tomography (PET) to improve the accuracy of GTV delineation for RT planning.

I. C. Role of functional imaging in H&N cancer care

The current standard of care for imaging H&N cancer patients involves the acquisition and interpretation of CT and MRI scans. Both modalities provide volumetric and anatomical information. CT and MRI information is used for diagnosis, staging and treatment response assessment and widely used for GTV delineation. The main advantages of CT are its high resolution (1 mm or less for both slice thickness and transverse voxel width), absence of geometrical distortions, and the fact that it provides an estimate of the electron density, which is used in the dose calculation process. However, soft tissue contrast in CT is poor, which leads to large delineation variability across observers [20], [21]. Furthermore, CT imaging is prone to artefacts caused by metallic implants, which can significantly alter the quality of scans in regions such as the oral cavity. MRI data, with different sequences (T1-weighted, T2-weighted, or diffusion weighted MRI) show higher soft tissue contrast and allow better identification of the bone extension. This can potentially lead to smaller inter-observer variability and better accuracy in the GTV delineation when used in combination with CT information [22], [23]. However, no improvement in GTV delineation accuracy was shown when using MRI instead of CT in sites such as pharyngolaryngeal tumours [24], [25]. The heterogeneity often observed within tumours has led a number of authors to recommend the use of additional information for identifying regions of different metabolic activity and biology, which might benefit from different types of treatment. Functional imaging techniques exist, which provide biologic information of the tissues, and can therefore be used to provide additional information to commonly used anatomical CT or MRI imaging. Such information can be used to define one or more biological tumour volumes (BTVs) and potentially increase outcomes.

Positron Emission Tomography (PET) is a functional imaging technique used for quantifying the accumulation of a given radiotracer within the body. It was developed in the 1950's and applied for the first time in 1953 by Sweet and Bronwell [26]. The technique relies on the simultaneous detection of 511 keV gamma ray pairs emitted during the annihilation of a positron with an electron encountered in the surrounding tissue. The positron is itself emitted by the radioactive substance, or radionuclide, injected into the blood flow of the subject to image. The term tomography refers to fact that the image is obtained by acquiring consecutive horizontal slices of the imaged subject, which, for PET imaging systems, progresses horizontally through the vertical detector ring containing the detectors (the gantry).

PET most commonly uses the radiopharmaceutical 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG), a molecule of glucose on which Fluorine-18 (^{18}F), a radioactive isotope of F, has been substituted to a ring hydroxyl group, as shown on Figure 1.

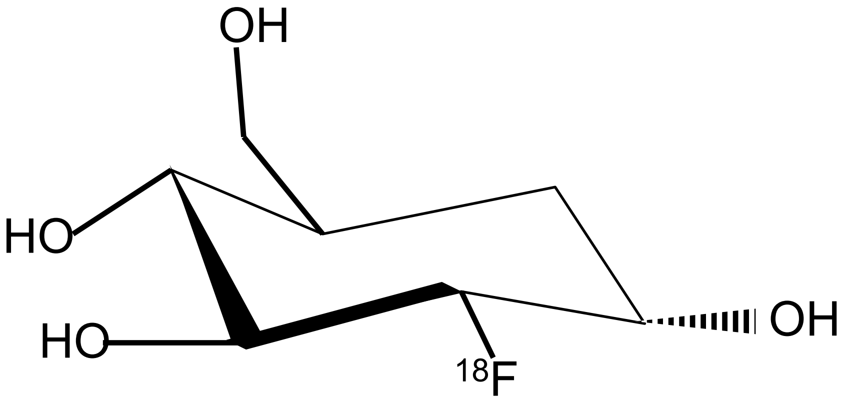


Figure 1. Schematic of a 2-deoxy-2-[^{18}F]fluoro-D-glucose

The half-life of 110 min of ^{18}F allows enough time for the tracer to be taken up by the metabolism before imaging. It also ensures a minimum risk to the patient, as the radioactivity of FDG will decrease to a negligible level within a few hours. As an example, the dose of 400 MBq injected to a patient would have decreased to 9.1 kBq after 10 hours. It also makes it possible to produce the radiopharmaceutical at facilities that may be a few hours away from the institution where they will be used. In addition, FDG proves ideal as a tracer, as the high energy of the photons emitted by the annihilation with an electron of the positron issued from ^{18}F decay allows a good penetration in the biological tissues [27].

The main advantage of FDG as a radiopharmaceutical lies with its biological properties. First, FDG is an analogue of the glucose molecule, which allows it to penetrate cell membranes with the sodium and glucose transport systems [28]. The absence of one of its hydroxyl groups, compared to the glucose molecule, ensures it is not further metabolised. A phosphorylation mechanism occurring when the molecule enters the cell causes it to remain trapped until it decays. Non-decaying molecules are then rejected by the body via the urinary system. FDG therefore allows identifying regions with high glucose uptake such as the brain, heart, and rapidly proliferating cell clusters. Findings from Otto Warbur *et al.* in 1927 showed that tumour tissue was supplied with 70 mg of glucose for 100 mL of blood, compared to 2-16 mg on average for normal tissue [29]. This causes FDG to also highlight metabolically active tumour cells within healthy tissue, making this tracer highly sensitive to the majority of cancers.

Other radiopharmaceuticals have been developed in the past decade as markers of different biological pathways, with a promise for use in RT planning. Proliferating cells show high phospholipidic activity, which is an essential component of the cell membranes. Tracers based on choline, a precursor in the biosynthesis of phospholipids, were developed to image this mechanism, for example 11C-choline or ^{18}F -choline. The reduced renal excretion of choline makes it a marker of choice for prostate cancer. In the H&N, however, the advantage of these components on ^{18}F-FDG is still unclear due to a current lack of data [30]. ^{18}F-Fluorothymidinole (^{18}F-FLT) was validated in a number of studies as another surrogate for tumour cell proliferation, and therefore a marker of cancer [31]. Recent research has focused on its use for assessment of response to therapy and delineation for dose escalation purposes [32]. Finally, tracers based on amino acids allow imaging the protein metabolism, which is increased in cancer cells. Radiopharmaceuticals such as [Methyl-^{11}C]-methionine (^{11}C-MET), 3-[^{18}F]-Fluoro-alpha-methyltyrosine (^{18}F-FMT) and O-(2-[^{18}F]-Fluorethyl)–tyrosine (^{18}F-FET) have recently been developed and evaluated for this purpose, but show no clear superiority identified on ^{18}F -FDG at present [33]**.**

An important area of investigation is the development and validation of radiopharmaceuticals correlating to tumour hypoxia. Hypoxia is an important radioresistance factor, and has been shown to correlate with reduced patient outcome after RT [34]. In addition, specific treatments targeting hypoxic areas are being developed, including the application of dose escalation to those regions. Hypoxia imaging can be done with nitroimidazole components for selective binding to hypoxic cells, which currently include ^{18}F-Fluoromisonidazole (^{18}F-MISO) [35], ^{18}F-Fluoroazomycin arabinose (^{18}F-FAZA) [36] and 3-^{18}F-fluoro-2-(4-((2-nitro-1Himidazol-1-yl) methyl)-1H-1,2,3,-triazol-1-yl)-propan-1-ol (^{18}F-HX4) [37] as the most promising. The lipophilic radioactive metal compound ^{62}Cu-methylthiosemicarbazone (Cu-ATSM) also showed good retention in hypoxic cells [38], but more work is currently required to better understand the underlying mechanism.

FDG-PET has recently become one of the tools required in the highest standard of clinical imaging for oncology. Its use is currently recommended by the British Association of Othorinolaryngology [39] for the diagnosis, staging, and detection of recurrence in H&N. Reports of clinical experience showed that FDG-PET was particularly beneficial for:

- Identifying the unknown primary tumour [40]
- Imaging distant metastases [41], [42]
- Excluding abnormalities (e.g. atelectasis: collapsed lung) from RT plan [43]
- Avoiding geographic miss of the gross tumour [44]
- Determining earlier tumour response to therapy compared to CT [28] [45] , especially with the use of texture features characterising the tumour heterogeneity [45], [46].

As stated by MacManus *et al.* in 2009, PET imaging is therefore likely to play an increasingly valuable role in RT planning for a number of cancers [43]. There is already some evidence of good outcomes or low loco-regional recurrence of patients treated with PET-guided IMRT and using PET information in defining the GTVs. [47]–[49]. Recent studies have highlighted a number of potential benefits of using PET in GTV delineation, compared to the use of CT or MRI data.

Firstly, manually segmented contours were shown to be less dependent on the operator performing the task when using FDG-PET compared to anatomical imaging ([50], [51], [52]). However, work by Riegel *et al.* revealed significant inter-observer variations in delineating 16 H&N patients, in the absence of a well-defined delineation protocol [49]. Experts from the International Atomic Energy Agency (IAEA) recommend the use of “a rigorous visual contouring protocol using predefined window and colour settings and with input from the nuclear medicine physician”, insisting that it “can give highly reproducible results” [43]. Automatic or semi-automatic segmentation has the potential to eliminate operator variability. In addition, such methods can reduce the time consuming task of delineating GTVs down to a few seconds without requiring a high level of expertise from the operator, although expert judgment will always remain critical for the validation of the contours. Alternatively, the availability of FDG-PET to the clinicians as a starting point for manual delineation also has the potential of speeding up the planning process, as well as reducing inter-observer variability, as suggested by Davis *et al.* [54].

Secondly, there is clear evidence to date that FDG-PET-based delineation provides different information beyond CT and MRI data. A study by Daisne *et al*., used CT and MRI contours for comparison, and showed that volumes delineated on PET were the closest to the volumes of the surgical specimen for nine pharyngolaryngeal SCC patients, but were not systematically encompassed in the anatomical contours [24]. Most comparative studies have shown a reduction in the GTV when including PET in the delineation process [42], [55]. This is in line with recommendations of the IEAE experts, stating that the planning volumes “should be kept as small as possible to minimise damage to other tissue” [43]. In a study by Nishioka *et al.* on 21 H&N cancer patients [47], the use of fusion between FDG-PET and MRI/CT allowed sparing of the parotids for 71% of the patients. Barker *et al.* showed a reduction of the irradiated volume and significant reduction of the dose to parotid glands when using FDG-PET/CT fusion compared to CT only in H&N patients [56]. Good clinical outcomes were obtained in studies using fused FDG-PET/CT for image guided IMRT [48], [57].

In addition to the delineation of GTVs, a growing number of research groups are focusing on using FDG-PET for defining a dose boosting or dose escalation volume ([58],[59]). This is also supported by findings such as the ones by Wang *et al.* on 89

H&N patients, which showed that the use of FDG-PET-based contours could help avoiding recurrences [5]. In a study of 13 H&N patients, Thorwarth *et al.* suggest that dose painting, which specifies local dose levels according to the underlying PET intensity, is more effective than delivering a uniform boost to the FDG positive area [60]. However, other tracers might be preferable to FDG in this case, to better consider tumour heterogeneity (the authors based the dose-painting on using ^{18}F-MISO).

I. D. Challenges for PET imaging in H&N cancer care

The mechanism of uptake of FDG makes it a tracer specific to highly metabolic areas. Work by Otto Warburg showed that tumour tissue is supplied with 4 to 30 times more glucose per 100 mL of blood, compared to normal tissue [29]. In a tabulated review of findings in more than 14000 clinical PET studies, Gambhir *et al.* found an average sensitivity of FDG-PET of 84% in clinical evaluation, and a specificity of 88% [61]. Work by Laubenbacher *et al.* found significantly higher sensitivity and specificity for PET compared to MRI on 22 H&N patients in the identification of lymph node (90%/96% compared to 78%/71%) and in the involved neck side (89%/100% compared to 72%/56%) [62].

However, FDG-PET imaging is also subject to a number of artefacts and should be interpreted with caution. First, FDG uptake can occur in healthy tissue in some cases, potentially causing false positives in the interpretation of the FDG-PET data. The experts of the IAEA 2006-2007 commission issued a report in 2009 warning that the uptake of FDG in tumours is affected by a range of factors including tumour blood flow [63], activity of glucose transporters [64], activity of hexokinase, and glucose consumption [65]. As a consequence, FDG uptake can occur in thymic hyperplasia, fat necrosis, and smooth, skeletal or cardiac muscle [43]. In the H&N, the interpretation of FDG-PET images of is made difficult by the presence of FDG uptake by salivary glands and salivary excretion, uptake by some muscles close to the oral cavity and by the vocal cords, uptake by lymphatic structures containing macrophages, as well as by the proximity of structures in the upper aero-digestive tract [66]. On the other hand, FDG uptake is not always visible on PET images for lesions smaller than 5-10 mm with high background uptake [28]. This is particularly problematic in the H&N, where lymph nodes are numerous and lesions smaller than 4 mL are often observed [67]. FDG-PET can also fail to identify superficial mucosal extensions as part of the tumour [24].

Moreover, the technique of PET is limited both by its theoretical modelling and its technical application. Uncertainties in the localisation of the annihilation event are in part linked to the positron mean free path in the tissue, which leads the particle to travel up to 2.4 mm in low density tissues (0.6 mm on average) before encountering an electron. In addition, the positron's residual energy causes the angle between the coinciding gamma rays emitted to be typically about 0.2° different from the expected 180° value, with extreme differences reaching up to 6°. For a gantry of 1 m diameter, this amounts to a typical error of 2 mm in the localisation of the annihilation event. As a result of these uncertainties, the PET data are resampled into large voxels, which are assigned a value representing the mean intensity at the corresponding location. This resampling effect, often called the "tissue fraction effect", causes the information coming from different tissues to be translated into a single value in the resulting image.

Further limitations are due to the characteristics of the detector, consisting of several blocks containing lutetium orthosilicate (LSO) or gadolinium orthosilicate (GSO) scintillation crystals. These convert detected gamma rays into visible light, which photomultiplier tubes (PMTs) turn into an electric signal giving the position and energy of the scintillation event. Detectors have a fixed width, and are connected to a limited number of PMTs. The accuracy of the localisation of a scintillation event in the crystal, and therefore the spatial resolution of the system, is limited by the fixed width of the crystals and the limited number of PMTs. The crystals are not capable of detecting two photons within a certain period of time, t, which also limits the temporal resolution of the scanner. A recent advance in the technique, called Time-Of-Flight (TOF) correction, allows localizing the annihilation event on a straight line drawn between two detectors by measuring the time delay between two coincidence detections. Current state-of-the-art PET scanner systems using TOF correction have a spatial resolution of 4-7 mm, and the use of TOF correction in the reconstruction was shown to improve lesion detection [68]. The finite spatial resolution of PET systems, especially when TOF correction is not applied, leads to a phenomenon of image blurring in 3D. The combination of both tissue fraction effect and 3D blur is referred to as the Partial Volume Effect (PVE) [69] in this thesis. As a consequence of this, objects with dimensions that are small compared to the Full Width at Half Maximum (FWHM) of the imaging system's point spread function will have their activity underestimated on the reconstructed PET image. PVE is one of the factors greatly hampering the detection and accurate delineation of tumours on clinical PET images. In particular, the PVE causes the boundaries of an object to appear blurred on the resulting PET image, making the detection of the object edges difficult, especially for methods based on the identification of gradient crests. In addition, the different positron range and photon scatter properties between media of different density can cause signal from the neighbouring regions to be produced into one another, generating a "spill-out" phenomenon observed in particular at the boundary between high radiotracer uptake regions and air.

I. E. PET-based delineation

Manual delineation by radiology experts currently remains the standard for GTV delineation on FDG-PET in RT planning. However, the limitation in the resolution of PET images described previously and the complexity of biological tumour uptake make manual PET delineation a time consuming and highly operator-dependent process, requiring the availability of specific expert knowledge for FDG-PET. This explains the growing interest in automatic or semi-automatic segmentation tools to assist or perform the GTV delineation, with the additional potential advantages of making the process more reliable by eliminating errors due to human judgement, and making it standardised across different centres. Some studies have already shown the important reduction in inter-observer variability when using automatic PET delineation compared to manual delineation by experienced observers [70].

However, several authors have commented on the lack of consensus on a standardised and accurate segmentation method [71]–[73]. A variety of segmentation techniques have been published or recommended for clinical practice, but there is no recommendation or consensus for a single protocol to use, in particular for H&N cancer. It is crucial to select the most accurate method, as studies comparing different segmentation tools have shown resulting volume differences of up to 200% [74].

Single thresholding methods include in the volume delineated all voxels with intensity higher than a single threshold value. The threshold value specified can be a single intensity value (absolute), or a percentage of the maximum voxel intensity in the image. The latter option allows for a less patient-specific process, especially when the image intensity is expressed in Standardised Uptake Values (SUVs) defined as:

$$SUV = \frac{Measured\ activity\ (Bq/mL)}{Injected\ Activity (Bq)}\ .Patient\ weight\ (g) \qquad \text{Eq. 1}$$

(It is assumed that 1 mL is equivalent to 1 g, as the human body is made mostly of water molecules). The injected activity corresponds to the activity injected corrected for radioactive decay between injection and image acquisition, while the measured activity is the activity read on the PET image.

Although simple to use and intuitive, thresholding methods have been shown to lack in accuracy and robustness, in particular for inhomogeneous and irregular lesions [74], [75]. Work by Geets *et al.* found that the optimal threshold for matching the contours of the macroscopic specimen in pharygolaryngeal cancer patients ranged between 36% and 73% of the maximum intensity, which shows that no single thresholding method can accurately delineate the GTV, even within a single site [76]. In addition, such methods appear very sensitive to the image reconstruction method, tumour size, Tumour-to-Background Ratio (TBR) and system response [77]. Data from several groups also highlighted the high dependency of the delineated volume on the threshold SUV value chosen [74], [77]. Single thresholding is therefore often used with additional features making it more stable and reliable. These can include:

- Region growing techniques, which avoid obtaining disconnected contours by growing a region of connected voxels step-by-step. This is done by including in the growing region the voxels neighbouring the region which have an intensity value above the threshold.
- Defining the relative threshold according to the "peak" SUV, defined as the mean value inside a 1 cm^3 sphere around the maximum SUV voxel [78], which can minimise bias due to noise.
- Adaptive thresholds, which are calculated relative to the difference between the maximum (or peak) intensity value and the background mean value. However, such methods rely heavily on the definition of the background area, which varies largely across publications ([67], [79], [80]).

The full dependency on a single parameter, the threshold value, makes simple thresholding methods practically just as operator-dependent as manual delineation. Several authors have provided methods for the calculation of the optimal threshold value on the basis of the tumour size, intensity or background intensity. Some authors used a linear combination of these parameters ([81]–[83]), while others provided calibration curves obtained with phantom data [67], [77], [84], [85]. The main limitation of such approaches is the need for *a priori* knowledge of the object volume and activity. The use of calibration curves showed good object volume recovery, but required accurate equipment-specific calibration, and was only applied and tested on spherical objects. A multicentre study showed that the optimal threshold to apply for the segmentation of spherical phantom inserts varied across centres with different imaging protocols and reconstruction settings, even as the centres were using the same scanner and reconstruction technique [86].

Following the reports on the lack of reliability and robustness of single thresholding techniques, a number of different segmentation approaches have been investigated in the recent literature. Advanced image segmentation approaches, some of which have been investigated for use on PET images, can be classified into a number of categories:

- Automatic threshold–based approaches that iteratively find the optimal threshold value to recover the object, with no user input or a priori information required [80], [87].
- Region-growing schemes that operate by implementing a step-by-step process to grow a region starting with a single voxel (the seed), by incorporating neighbouring voxels on the basis of their intensity value. Different seed selections, voxel inclusion criteria and stopping criteria for the growing region can be used [88], [89].
- Clustering approaches classify voxels iteratively into a number of groups (clusters) of homogeneous intensity values. The number of clusters identified is specified by the user or by the code itself. Cluster membership can be binary (yes or no, i.e. 0 or 1) or can be expressed as a probability ("fuzzy" clustering) [72], [90], [91]. Other clustering work was based on fitting the cluster intensity distributions to Gaussian distributions [92], [93].
- A growing number of studies have investigated the use of parameters such as Haralick texture features [94], which describe the regional and local distribution of intensities across a region, as a basis for clustering applied to PET images [95], [96]. This was investigated using both PET and CT, as an improvement on the use of PET or CT alone [95], [97].

- Edge detection methods that are based on the identification of rapid changes in intensity, corresponding to the intensity “crests” in the gradient of the original PET image. Several approaches exist for the detection of crests, including gradient-based thresholding [98], region-growing based on gradient threshold [99], and algorithms such as the Watershed Transform [100], [101].
- Active contours methods that are based on successive deformations of a contour to reach an equilibrium, which can be defined by a set of criteria involving the voxel intensities inside and outside the contour, as well as the shape and length of the contour [102], [103].
- Artificial neural networks (ANN) rely on the iterative classification of voxels according to a complex set of relationships between them, involving their location as well as intensity values [104].
- Other machine learning techniques, such as support vector machine have been investigated for modalities such as MRI or CT [105]. These approaches could potentially also be applied to PET, for example in combination with other segmentation approaches [106]. However, published data about this subject is currently very limited, probably due to the higher level of complexity required by such techniques.
- Finally, a number of different techniques can be combined within more complex segmentation frameworks. This has been investigated in a small number of studies using tools such as the simultaneous truth and performance level estimation (STAPLE) algorithm [107], majority voting [108] or probabilistic methods [109].

A small number of methods have been published and validated on test images and patient data. Geets *et al.* have developed a gradient-based method, which showed a good correlation with the ground truth for volumes from seven patients with T3-T4 laryngeal SCC [98]. Day *et al.* developed a 3D region-growing method, which performed better than fixed thresholding schemes on 18 rectal and anal cancer patients [89]. Hatt *et al.* developed and validated a Fuzzy Locally Adaptive Bayesian method (FLAB), based on a fuzzy clustering scheme incorporating an expectation maximisation step [91]. Evaluation of the FLAB method with spherical fillable phantom data and more complex simulated data showed high accuracy of the method compared to thresholding and other clustering methods, especially for small objects. The methods published rely on very different segmentation approaches, and have been validated by their authors on different types data provided by their own centres, which makes it difficult to compare the results obtained. In addition to validation on a large and useful range of data, very few methods have been evaluated in terms of repeatability and robustness.

The past decade has seen a strong effort in the scientific community to investigate alternative methods and a number of automatic or semi-automatic segmentation algorithms have been published. However, the IAEA experts panel notes that the availability of numerous automated segmentation methods and the absence of any reliable inter-comparisons makes it difficult to recommend a single technique, but insist that single-parameter methods are too simplistic for the variety of clinical scenarios encountered and are not recommended [43].

I. F. Thesis aims

The high potential of advanced PET-Automatic segmentation (PET-AS) methods in RT planning for H&N cancer patients is hampered by the current lack of inter-comparison and exhaustive validation of such methods. However, there is enough evidence in the recent literature to suggest that FDG-PET should play a key role in the planning of curative radiotherapy for H&N cancer patients, and that work is needed to identify the optimal protocol for the inclusion of FDG-PET delineation into the RTP process. The project described in this thesis aimed at addressing these issues, in the form of a pilot study titled POSITIVE: Optimisation Of Positron Emission Tomography Based Target Volume Delineation In Head And Neck Radiotherapy. It was funded by Cancer Research Wales and carried out as a collaboration between two different institutions:

- The Wales research and diagnostic PET imaging centre (PETIC), which opened in 2010, offers some of the most advanced imaging equipment in the UK, with a high-resolution scanner providing high quality images for research and clinical purposes. PETIC is operated by Cardiff University in partnership with Cardiff and Vale University Health Board, and is located at the University Hospital of Wales in Cardiff.
- Velindre Cancer Centre (VCC), located in Cardiff, is one of the largest specialist centres for non-surgical cancer treatment in the UK, with over 5000 new patient referrals every year. It boasts high-end equipment, with linear accelerators enabling IMRT and image guided RT (IGRT) procedures, and strong links with the Wales Cancer Trials Unit and the Wales Cancer Bank for conducting world class research through oncology trials.

This thesis therefore aims at addressing the following points:

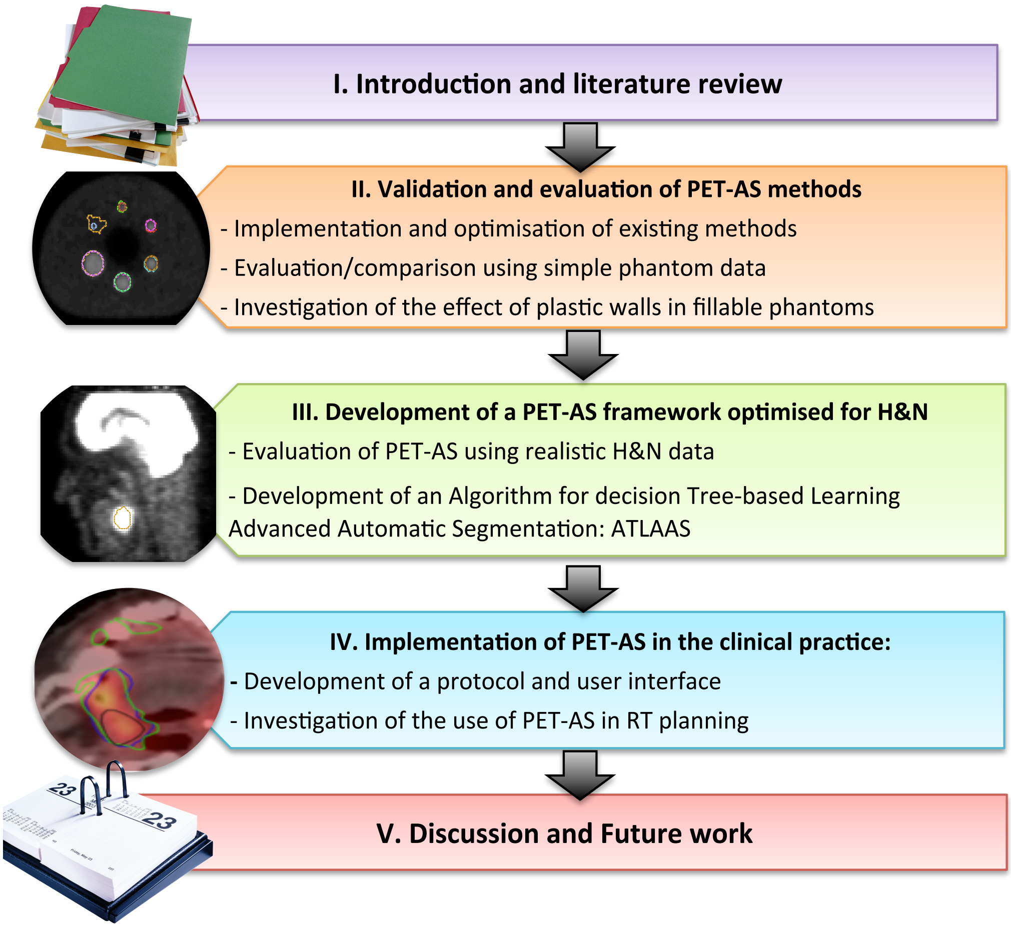
- Provide a solid and exhaustive comparison of advanced PET-AS methods
- Investigate the effect of a range of image parameters on such PET-AS methods
- Provide a limited set of segmentation tools or a single PET-AS tool validated and optimised for use in H&N RT planning
- Develop a protocol for the use of PET-AS for potential use in routine clinical practice for H&N RT planning

For this purpose, it was hypothesised that:

- Advanced PET-AS algorithms provide more accurate delineation than simple thresholding schemes

- The presence of inactive plastic walls in fillable phantoms has a non-negligible impact on the image quantification and segmentation.
- Advanced PET-AS can be used within a clinical protocol
- Optimised advanced PET-AS for H&N reduces the time needed to generate RT plans and reduces observer variability.

The outline of the thesis is described on Figure 2.



I. Introduction and literature review

II. Validation and evaluation of PET-AS methods

- Implementation and optimisation of existing methods
- Evaluation/comparison using simple phantom data
- Investigation of the effect of plastic walls in fillable phantoms

III. Development of a PET-AS framework optimised for H&N

- Evaluation of PET-AS using realistic H&N data
- Development of an Algorithm for decision Tree-based Learning Advanced Automatic Segmentation: ATLAAS

IV. Implementation of PET-AS in the clinical practice:

- Development of a protocol and user interface
- Investigation of the use of PET-AS in RT planning

V. Discussion and Future work

Figure 2. Outline of the thesis and description of chapter contents

Chapter II. Validation and evaluation of PET-AS methods

II. A. Development and validation of segmentation algorithms

This chapter describes the experiments and analysis carried out in order to achieve one of the aims of the POSITIVE project: the investigation of the impact of a number of image parameters on the segmentation accuracy of current published methods. These include object-related aspects (object geometry, size, phantom type), and PET image-related aspects (TBR and image noise). This work was done using fillable phantoms allowing simple and controllable generation of well-defined target objects. This section describes the algorithms implemented and the tools used to evaluate their accuracy.

II. A. 1. Methods and materials used

II. A. 1. a. Scanner

The scanner available for the project was a GE (General Electric Healthcare, Milwaukee, USA) Discovery 690 PET/CT, dedicated to clinical and research work. All experiments were carried out with the acquisition and reconstruction parameters described Table 1. The reconstruction algorithm used was Vue Point FX, which is based on a Maximum Likelihood Ordered Subset Estimation Maximisation (ML OSEM) method with TOF correction. The scanner is shown on Figure 3.

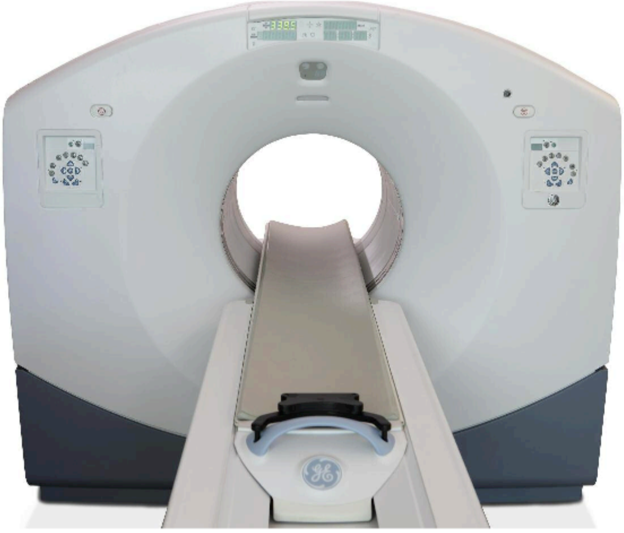


Figure 3. GE Discovery 690 PET/CT scanner used throughout this project.

Parameter to set	Value chosen
Matrix size CT (voxels)	512 x 512 x 47
Matrix size PET (voxels)	256 x 256 x 47
Voxel size CT	1.37 mm x 1.37 mm x 3.27 mm
Voxel size PET	2.73 mm x 2.73 mm x 3.27 mm
Field of View dimensions	700 mm x 153 mm
Duration of bed position	3 min
Reconstruction algorithm	Vue Point FX TOF-corrected
Algorithm settings	3D ML OSEM 24 subsets 2 iterations cut-off
Post-processing filter cut-off	6.4 mm
CT-based attenuation correction	yes

Table 1. Scanner settings used for the acquisition of the phantoms scans.

II. A. 1. b. Phantoms

The NEMA (National Electrical Manufacturer's Association) IEC (International Electrotechnical Commission) body phantom (manufacturer: The Phantom Laboratory, Salem, USA), used for quality assessment of the scanner images, was available for this project. It consists of a 9700 cm^3 fillable sealable plastic tank containing six spherical fillable inserts of inner diameters 10, 13, 17, 22, 28 and 37 mm, corresponding to volumes of 0.5, 1.2, 2.6, 5.6, 11.5, 26.5 mL respectively. The phantom is shown on Figure 4. It includes a non-fillable central insert of low density representing lung tissue, which was used in all scans of this phantom for this project.

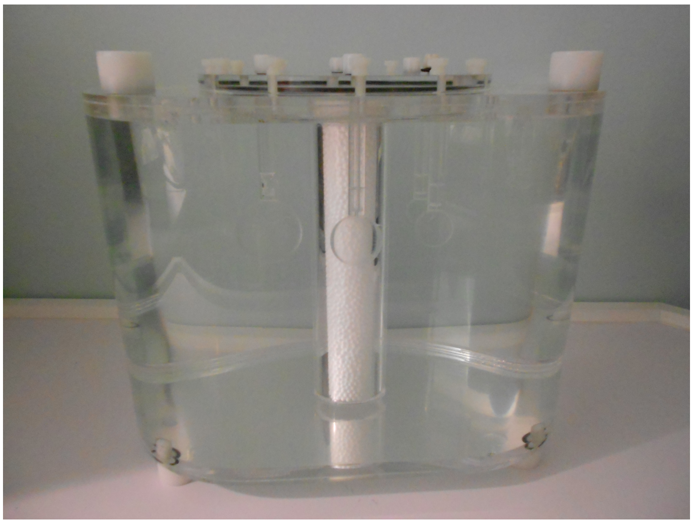


Figure 4. Picture of the filled NEMA IEC body phantom with central lung insert.

II. A. 1. c. Hardware and software

A 2.7 GHz quad-core Intel Core i5 computer was dedicated to the project. The algorithms were developed in the Matlab programming language with a Matlab 2010b licence (The Mathworks, Natick, USA), including the Image Processing Toolbox. The visualisation and processing of CT and PET images was done with the open source software CERR (a Computational Environment for Radiotherapy Research) [110]. CERR was developed at the university of St Louis (Michigan, USA) and is currently maintained at the Memorial Sloane Kettering Cancer Centre (MSKCC) in New York (USA). The statistical analysis software SPSS 20 (IBM, Chicago, USA) was used throughout this project.

II. A. 1. d. Segmentation algorithms

Following a review of the recent literature, a number of PET automatic segmentation (PET-AS) approaches were selected as the most promising for PET delineation. These were chosen based on the segmentation categories described by Bankman *et al.* [111], and implemented in house into a common framework as fully automatic methods.

In addition to the PET-AS methods implemented in this study, three basic thresholding algorithms using thresholds of 42% and 50% of the maximum SUV value in the tumour (FT_{42} and FT_{50}) and a threshold of 2.5 SUV ($SUV_{2.5}$) were sometimes used for comparison. The threshold values selected correspond to commonly used delineation methods (cf. I. E).

II. A. 1. d. i. Thresholding methods

AT: Automatic iterative thresholding

The AT method implemented iteratively modifies the contour by applying to the image successive thresholds. These are calculated for every iteration i from an estimation of the background mean intensity as follows:

$$T^{i+1} = 0.45 * \left({SUV_{max}}^{i} - {B^{i}}_{mean}\right) + {B^{i}}_{mean} \qquad \text{Eq. 2}$$

with SUV_{max} the maximum SUV value inside the lesion, B_{mean} the mean background (non-lesion) intensity, and T the absolute threshold intensity to apply to the image. This method was based on the method developed by Drever *et al.* [80], but was implemented with different initialisation and stopping criteria. The algorithm is initialised with a mean background value calculated on voxels with intensity lower than 50% of the maximum intensity, and a value of 0.4 for the relative threshold applied in Eq 2 (instead of 0.45 for subsequent iterations). Equation 2 is applied at each step until the region delineated changes by one voxel or less. The method is illustrated on Figure 5.

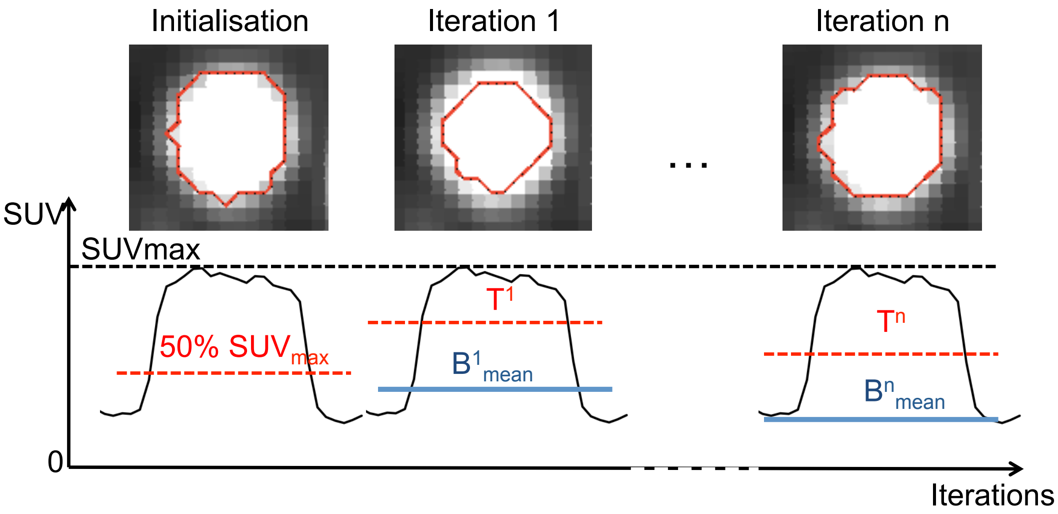


Figure 5. Illustration of different iterations in PET-AS method AT.

II. A. 1. d. ii. Gradient-based methods

Gradient-based segmentation algorithms are based on the image intensity gradient map, which is calculated in Matlab using the two-dimensional (2D) or three-dimensional (3D) Sobel operator as described in [112]. The Sobel operator calculates an approximation of the image gradient using a discrete differentiation of the image intensity function.

GC: Gradient-based contouring

This segmentation algorithm utilizes the method used by the Pinnacle^3 (Phillips Healthcare, Guildford, UK) software as briefly described by Ford *et al.* [77]. It uses the gradient image obtained slice-by-slice from the original image by applying the Sobel filter in the transverse plane. The algorithm searches voxel-by-voxel for the highest gradient neighbour in a clockwise manner, starting from the highest gradient value voxel in the image. This process is illustrated on Figure 6.

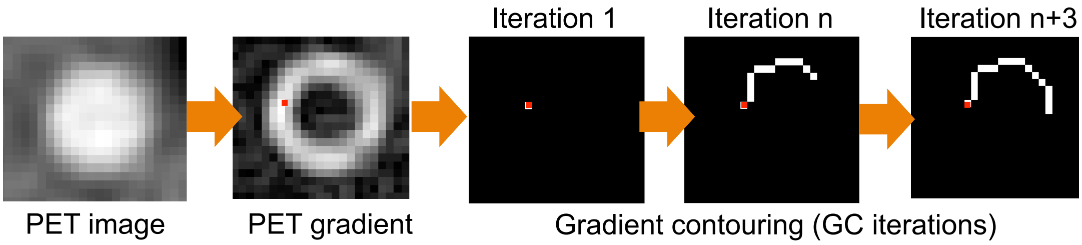


Figure 6. Illustration of PET-AS method GC, extracting the gradient from the original image, and following the highest gradient crest clockwise from the seed (red).

WT: Watershed Transform-based segmentation

This method was based on the Watershed Transform algorithm, described in several studies [98], [100], [113], which finds the "crests" of the gradient image by simulating a water level rising from the local minima in the image gradient. The process is carried out until only one closed contour remains. The algorithm was fully written in-house, only using the Sobel operator available in Matlab to derive the gradient image. It is illustrated on Figure 7.

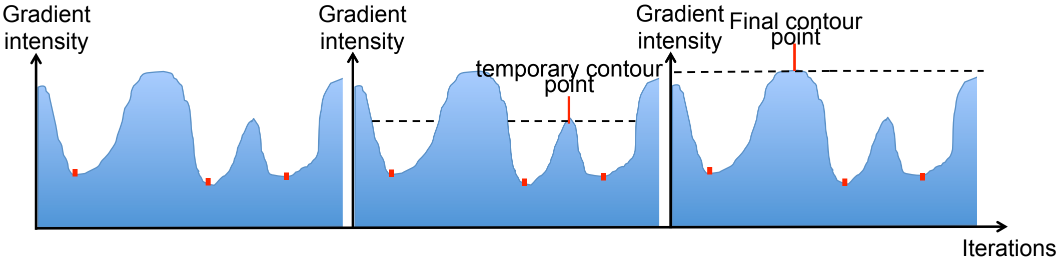


Figure 7. Illustration of WT method using seeds (in red).

II. A. 1. d. iii. Region-growing methods

RG: Region growing

This algorithm selects one voxel as a seed and grows a region step-by-step by including some of the voxels at the border of the growing region on the basis of their intensity value. The method, based on the work of Day *et al.* [89], was developed and optimised in-house, by automatically choosing the highest intensity voxel inside the hottest region of the image as a seed, and stopping the algorithm when the number of voxels added represents less than 5% of the total number of voxels in the growing region. This value was chosen as a good trade-off between computation time and accuracy, for a number of voxels ranging between 9 and 5000 corresponding to typical lesions observed at Velindre Cancer Centre. Voxels are added to the growing region if their intensity is within the Confidence Interval (CI) of the mean intensity in the growing region. CI was chosen as follows:

$$CI = \min\left(2\sigma, (T_{mean} - B_{mean} - \sigma_B)\right) \qquad \text{Eq. 3}$$

with T_{mean} the mean intensity value inside the lesion, B_{mean} the mean background intensity, σ and σ_B the standard deviation (SD) of the intensities in the lesion and background respectively. This criterion was chosen to take into account cases where the intensity distributions for lesion and background are well separated, as well as cases where they overlap. This is illustrated on Figure 8.

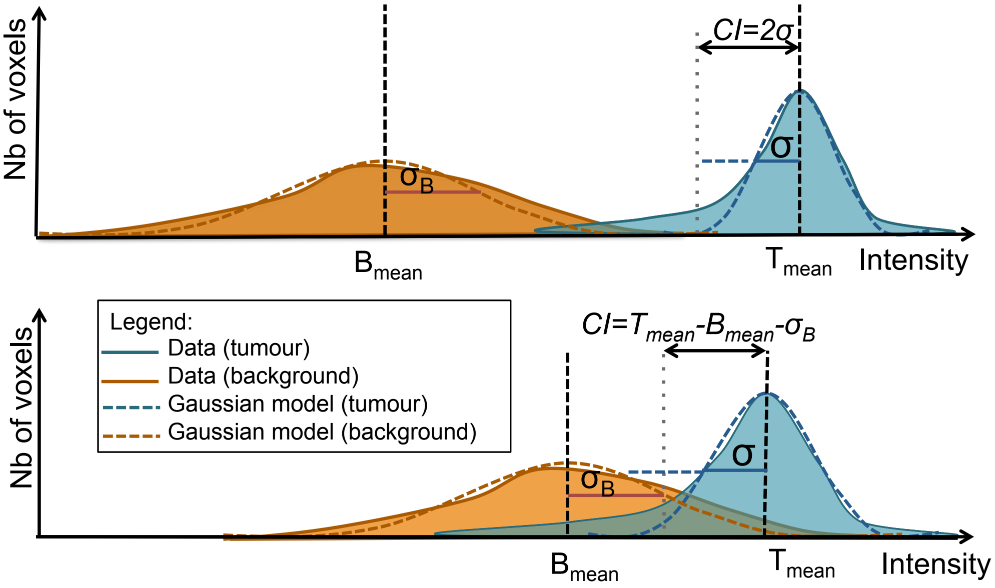


Figure 8. Illustration of the confidence interval CI for method RG a) when lesion and background intensity distributions are well separated and b) when they overlap.

II. A. 1. d. iv. Clustering methods

Clustering methods have acquired great popularity in the last few years. These methods are based on the iterative classification of the voxels into a defined number of categories called clusters, and voxels are classified according to the updated parameters so as to produce homogeneous regions. This is done iteratively, calculating parameters describing the clusters (e.g. mean intensity value or SD) at each iteration, and updating the cluster memberships of each voxel to the different clusters. The methods described in this section used an updated mean intensity value M for each cluster k at iteration i, calculated as:

$$M_k^i = \frac{\sum_j u_k^{i-1}(x_j) * I(x_j)}{\sum_j u_k^{i-1}(x_j)} \qquad \text{Eq.4}$$

where $I(x_j)$ is the intensity value of voxel x_j, and $u_k^{i-1}(x_j)$ the cluster membership of voxel x_j for cluster k at the previous iteration. In the final step, all but the lowest intensity clusters are considered to form the tumour, and the remaining cluster, the background. All clustering methods were implemented so as to be able to detect a given number K of clusters, which was done in parts of this thesis. The clustering process is illustrated on Figure 9 for K=5.

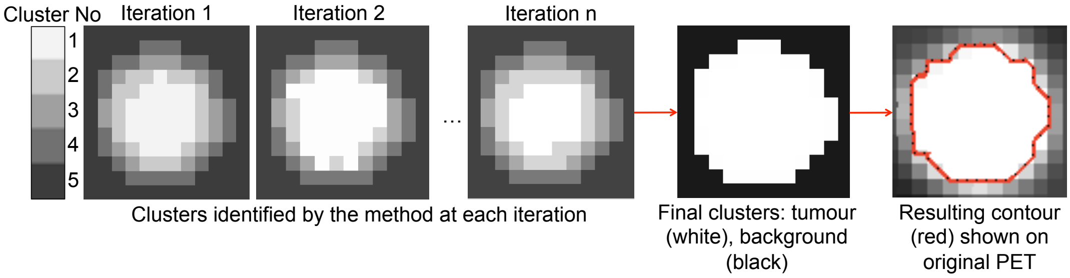

Cluster No
1
2
3
4
5
Iteration 1
Iteration 2
...
Iteration n
Clusters identified by the method at each iteration
Final clusters: tumour (white), background (black)
Resulting contour (red) shown on original PET

Figure 9. Description of the segmentation process using a clustering method, in the case of K=5 levels.

The following clustering algorithms were implemented:

KM: K-means clustering

This algorithm assigns each voxel of the initial image to the cluster with mean intensity value closest to its own value. This corresponds to

$$u_k^{i-1}(x_j) = \begin{cases} 1 \; if \; \|I(x_j) - M_k^i\| = \min_l \left(\|I(x_j) - M_l^i\|\right) \\ 0 \; otherwise \end{cases} \qquad \text{Eq.5}$$

where "|| ||" represents the absolute difference. The method was based on the method described by Zaidi *et al.* [41] with a customised initialisation considering a partition of the image intensity range into the number K of levels chosen by the user.

FCM: Fuzzy C-means clustering

This algorithm was developed to account for the uncertainty arising at tumour boundaries in particular, by using a fuzzy classification instead of a binary one. It was based on the work described by Belhassen *et al.* [72]. In this case, each voxel is assigned a membership value for each cluster, ranging between 0 and 1. The membership value of a voxel x at iteration i is calculated as a probability to belong to the cluster k considered, according to the difference between the voxel intensity and the cluster mean intensity:

$$u_k^i(x) = \frac{\left\| I(x) - M_k^i \right\|}{\sum_j \left\| I(x) - M_j^i \right\|} \qquad \text{Eq.6}$$

GCM: Gaussian Fuzzy C-means clustering

This algorithm is based on the FCM algorithm, with the difference that each cluster is assumed to have a Gaussian intensity distribution, of which mean and SD are calculated at each step. The cluster membership for each voxel is the probability of the voxel intensity value being generated by the cluster Gaussian distribution:

$$u_k^i(x) = \exp\left(- \frac{\left\| I(x) - M_k^i \right\|^2}{2\left(\sigma_k^i\right)^2} \right) \qquad \text{Eq.7}$$

where $\left(\sigma_k^i\right)^2$is the variance of intensities in cluster k at iteration i. The method was implemented based on the modifications of FCM suggested by Hatt *et al.* [91].

The clustering algorithms were first implemented for the detection of 2 clusters, and the names KM, FCM and GCM refer to the versions of the algorithms corresponding to K=2. These were named KM2, FCM2 and GCM2 in later studies to avoid confusion with other versions involving higher numbers of clusters.

II. A. 1. d. v. Deformable models

AC: Active contours

The active contours method implemented was based on previously published work by Sussman *et al.* [114]. It uses a level set approach, defining the contour on each slice using the signed distance function ϕ. This function returns for each point of the image its signed Euclidian distance to the closest contour point, which takes negative values inside the contour, positive values outside, and equals 0 for the contour points. The contour is evolved at each iteration I by updating ϕ, in order to satisfy the following equation for a voxel k on the contour:

$$\phi_k^{i+1} = \phi_k^i - F^i\left|\nabla\phi_k^i\right| \qquad \text{Eq. 8}$$

where F represents the force deforming the contour, ∇ the gradient operator. The process is illustrated on Figure 10.

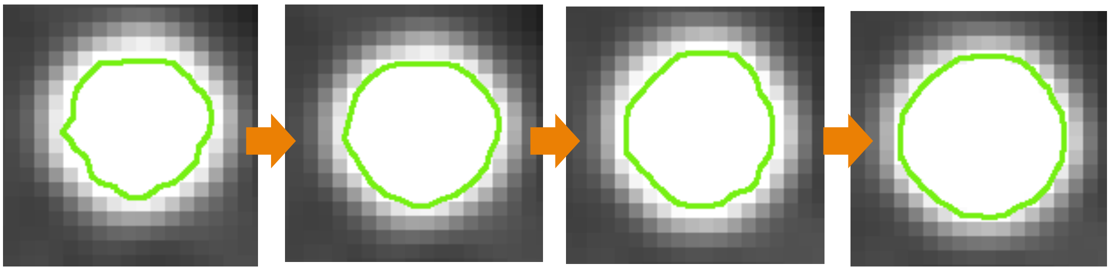


Figure 10. Illustration of different steps in the 2D active contours process used in AC, with successive contours (green) shown on the original image.

The force F was chosen so as to minimize the differences between the values of the level set curve and both the values outside and inside the curve, while limiting the length of the curve with a curvature term α as described in the literature [115], [116]:

$$F^i = \left(I({\phi_k}^i) - {Mean^i}_{int}\right)^2 - \left(I({\phi_k}^i) - {Mean^i}_{ext}\right)^2 + \alpha C^i \qquad \text{Eq. 9}$$

for iteration i, with I the original image intensity, ${Mean^i}_{int}$ and ${Mean^i}_{ext}$ the mean intensity values inside and outside the curve respectively, and C^i the contour curvature. The parameter α was set to 0.1, which provided the best results after some preliminary tests.

The gradient at voxel of x and y coordinates (i,j) was calculated using discrete formulation as follows (for a voxel size of s in both x and y directions):

$$\nabla\phi = \begin{bmatrix} \frac{(\phi_i - \phi_{(i-1)}) - (\phi_{(i+1)} - \phi_i)}{2s} \\ \frac{(\phi_j - \phi_{(j-1)}) - (\phi_{(j+1)} - \phi_j)}{2s} \end{bmatrix} \qquad \text{Eq. 10}$$

II. A. 1. e. Description of the metrics used

Throughout this work, the accuracy of the PET-AS methods was evaluated by quantifying the agreement between the 3D test contour obtained and a reference 3D contour, which were both extracted from binary masks in Matlab. This section describes the different metrics used to quantify the accuracy of the segmentations throughout the thesis. The following metrics were selected following the literature review, as a set of the most commonly used metrics in the field providing complimentary information, and were all implemented in house:

- Relative Volumetric Error (RVE) was used to evaluate the delineation accuracy in terms of volume. It was calculated as:

  $$RVE = \frac{(Y-X)}{Y} \qquad \text{Eq. 11}$$

  with X the volume obtained using the developed PET-AS methods, and Y the volume corresponding to the reference contour. RVE can take any positive (for a volume produced smaller than the true volume) or negative value (for a volume produced larger than the true volume).

- Dice Similarity Coefficient (DSC) was calculated to quantify the similarity between the structure delineated and the ground truth, providing values between 0 and 1. A DSC above 0.7 was used as an indicator of good overlap as suggested in the literature [117]:

  $$DSC = \frac{2 * |X \cap Y|}{|X| + |Y|} \qquad \text{Eq. 12}$$

The following additional metrics were used for each algorithm to further evaluate the performance of each algorithm individually:

- Sensitivity (S) gave the rate of "tumour" voxels detected by the algorithm:

$$S = \frac{TP}{FN+TP} \qquad \text{Eq. 13}$$

with FN the number of false negative voxels, TP the true positives. S ranges from 0 to 1.

- Positive Predictive Value (PPV) was used to determine the proportion of the delineated volume accurately classified as tumour:

$$PPV = \frac{TP}{FP+TP} \qquad \text{Eq. 14}$$

with FP the number of false positives. PPV can range from 0 to 1.

- Modified Hausdorff Distance:

$$HD = \max\left(\frac{1}{N_A}\sum_{a\in A} d(a,B), \frac{1}{N_B}\sum_{b\in B} d(b,A)\right) \qquad \text{Eq. 15}$$

with A and N_A, B and N_B the set of points and number of points within the test contour and true contour respectively, and $d(a,B)$, the minimal Euclidian distance between point a and the points in B. HD returns a positive value in cm. Distance metrics are used to quantify the distance between the contour points of two different outlines. This particular metric was chosen following the work of Dubuisson *et al.* [118], which shows the superiority of this method in quantifying the similarity between two contours, as opposed to other definitions of the distance metrics suggested by authors such as Huttenlocher *et al.* [119].

II. A. 2. Algorithms implementation and optimisation

II. A. 2. a. Algorithms 2D vs 3D implementation

II. A. 2. a. i. Purpose

PET-AS algorithms as presented in the literature can be implemented as a slice-by-slice process, or be applied to a full 3D image. Although some authors have suggested that full 3D segmentation could be more accurate than 2D methods, some published work has reported lower segmentation accuracies obtained for 3D implementations of some 2D algorithms [101]. However, such studies are scarce, and include little evidence or discussion of the type of implementation most suited to the segmentation of PET images. This section describes work carried out in order to check for any advantage of a 3D implementation over a slice-by-slice version of the algorithms implemented for this project.

The superiority of 3D implementation was hypothesised for the following reasons:

- Use of more spatial information (larger number of neighbours) e.g. for clustering and region growing methods
- More degrees of freedom for curve expansion or drawing (AC and GC)
- Taking into account the global intensity distribution in the case of methods or initialisations based on the maximum tumour intensity value.

II. A. 2. a. ii. Experimental protocol

Test images were generated using the NEMA phantom described in II. A. 1. b. Three different TBRs (3, 5 and 8) were achieved by filling the background and spherical inserts with two FDG solutions of different concentrations. For this series of experiments, 75 MBq were measured with a radionuclide calibrator, and injected into the phantom background filled with water. This value was chosen to lead to a background concentration of 5 kBq/mL at the time of the scan, which is representative of clinical soft tissue values. The precision error associated with the calibrator is within ±2% of the measured value, which leads to an uncertainty of ±4% on the actual TBR value obtained. The plastic phantom was then shaken to homogenize its contents, and the volume was completed with water. Next, the desired amount of FDG (calculated so as to achieve the target TBR) was drawn, and was diluted in a 1 L vial available in the laboratory. Finally the spheres were filled with the solution prepared using a shielded syringe, and the phantom was sealed. During the phantom preparation, the times of injection for both spheres and background were recorded, so as to account for tracer decay.

The phantom was positioned on the scanner bed with the sphere rods in the superior-inferior direction, and aligned with the lasers so as to keep the midline of the spheres at the same level in the transverse plane.

The algorithms AT, RG, KM, GC, WT and AC were implemented both on a slice-by-slice basis and in 3D, and used to segment the same test images. KM was chosen to represent the group of clustering algorithms. The accuracy of the resulting twelve algorithms was assessed by comparing the contours generated to the reference ground truth contour, using the metrics described in II. A. 1. e. This allowed looking for the effect of the implementation version (2D or 3D) on the delineation performance, and the variation of this effect with object size and contrast. The ground truth was extracted for each scan and each sphere by automatically generating a spherical contour of the same diameter as the sphere using a Matlab function written in-house. This contour was then positioned on the high resolution CT so as to match the inside of the sphere delimited by the visible plastic walls. It was then copied onto the registered PET image using a function available in CERR.

The Mann-Whitney U-test in SPSS was used to test for statistically significant differences in median between metric values obtained for 2D and 3D versions of each algorithm, with a statistically significant p-value of $p=0.05$.

II. A. 2. a. iii. Results

Figure 11 shows higher average DSC values for the 3D implemented version of AT, RG, KM, WT. PET-AS methods AC and GC reached higher DSC when implemented in 2D. In the case of WT, the difference of the average DSC of 2D and 3D method was smaller than the SD of results obtained across spheres.

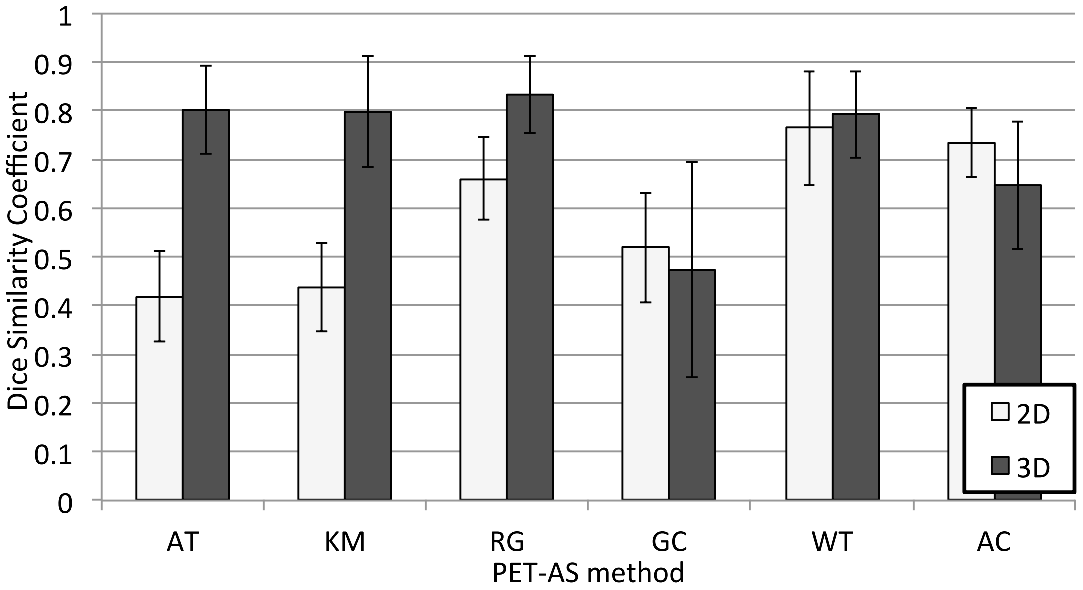


Figure 11. Comparison of DSC obtained on average on all NEMA spheres for the 6 PET-AS methods implemented in 2D and 3D. Error bars correspond to one SD across sphere sizes

The results of the Mann Whitney U-test, used to determine if the difference in median values between 3D and 2D implementation is statistically significant, are shown in Table 2. Results are given for values taken by RVE, DSC and HD. Differences in median are reported for the value reached by the 2D dataset subtracted from the value for the 3D dataset. Negative differences therefore indicate better accuracy for the 3D implementation for RVE and HD (higher RVE and HD correspond to lower accuracy), whereas a positive difference indicates a higher accuracy for the 3D version in the case of DSC, for which the values increase with the segmentation accuracy. Differences between 2D and 3D implementation were statistically significant for AT, RG and KM, and for GC except for HD.

	AT	RG	KM	GC	WT	AC
				RVE		
Difference in median (3D-2D)	-1.55	-0.33	-1.46	0.94	0.08	-0.06
U	282	259	289	281	187	196.5
sigma	$5.8\text{x}10^{-5}$*	$1.6\text{x}10^{-3}$*	$1.6\text{x}10^{-5}$*	$6.8\text{x}10^{-5}$*	0.44	0.28
				DSC		
Difference in median (3D-2D)	0.42	0.18	0.39	-0.36	-0.03	-0.01
U	282	269.5	287	273	168	179.5
sigma	$5.8\text{x}10^{-5}$*	$3.8\text{x}10^{-4}$*	$2.4\text{x}10^{-5}$*	$2.4\text{x}10^{-4}$*	0.94	0.58
				HD		
Difference in median (3D-2D)	-0.77	-0.28	-0.75	-0.03	0.00	0.04
U	312	318	314	195	166	189.5
sigma	$<10^{-6}$*	$<10^{-6}$*	$<10^{-6}$*	0.31	0.91	0.39

*Statistically significant

Table 2. Results of the Mann Whitney U-test comparing accuracy metrics obtained for 2D and 3D algorithms.

Figure 12 shows the change in accuracy achieved by using a 3D implementation compared to a 2D implementation for all algorithms at TBR=3 for the different sphere sizes. Results were similar but of lower magnitude for other TBRs. AT, RG and KM showed a large improvement of the 3D implementation on the 2D scenario, except for the smallest sphere where the 2D version of AT and KM performed better. GC showed systematically higher accuracy for a 2D implementation. WT and AC were little affected (less than 10% change in DSC) by the implementation type for spheres larger than 5 mL, but WT was largely (73% change) improved by the 3D implementation for the smallest sphere, while AC reached 93% lower DSC.

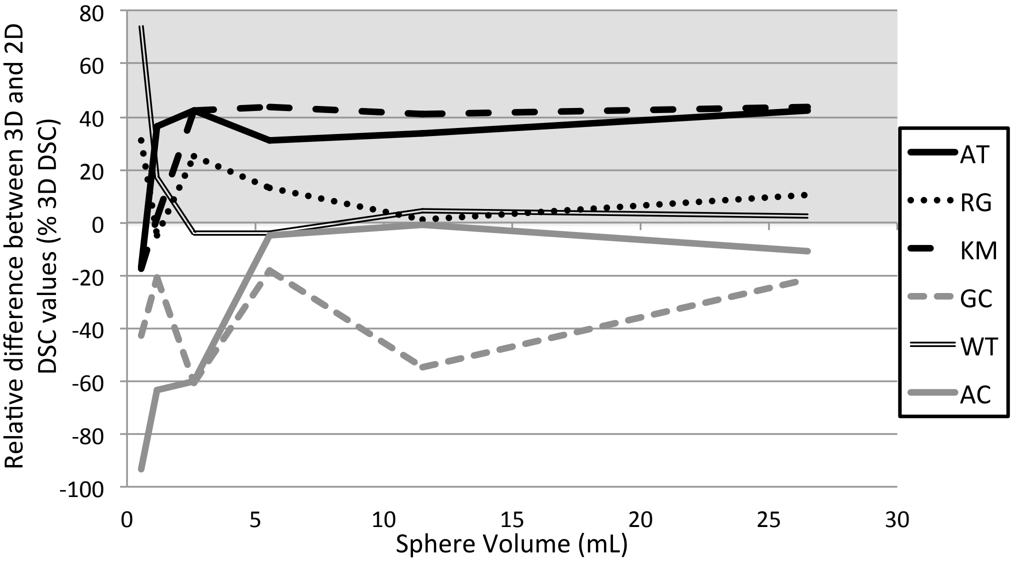


Figure 12. Change in DSC of the 3D implementation for the different algorithms and sphere volumes at TBR=3. The light grey area corresponds to a higher DSC for 3D implementation.

II. A. 2. a. iv. Discussion and conclusions

The effect of 2D or 3D implementation was different for the different algorithms tested. AT, RG and KM, which only use the original image intensities, showed significantly higher accuracy when implemented in 3D. This can be due to the following:

- RG and AT rely on the determination of the maximum intensity voxel in the image in order to define an intensity-based inclusion criteria, or a threshold to apply. This means that slices with very different contrasts (e.g. the middle slice compared to the last slice of a hot sphere) will generate different inclusion criteria or threshold values, leading for example to very large contours for a low contrast slice (low threshold value), but very small contours for a high contrast slice (high threshold value). In addition, the PET-AS methods are then unable to detect slices on which the structure does not appear, and will thus generate irrelevant contours. This may also explain why the improvement seen for these methods was lower for the smallest sphere, which only consists of 2 to 4 slices.

- RG and KM rely on the definition of 3D neighbouring regions and the calculation of their mean intensity. Such regions will have different and unrelated mean intensities across slices in the 2D version. For this reason, the segmentation will again generate contours that do not match the ones on neighbouring slices.

GC showed significantly lower performance when implemented in 3D (cf. Figure 12). This is likely to be due to the initial 2D-design of this method, and the way it was implemented in 3D. In 2D, the algorithm calculates for each slice the image gradient, chooses the highest gradient point and follows a clockwise path of highest gradient points, neighbour-by-neighbour. In 3D, the gradient and highest point, are calculated for the whole volume, and the implementation needs to be modified to adapt to a 3D image. In the present version, a 3D region corresponding to the object surface is grown, rather than a 2D-contour, with the region growing function. Another approach, consisting in finding all neighbouring highest gradient voxels until a closed 3D-contour is reached did not yield any satisfying results. Due to the added complexity, there are many possible approaches to adapting GC to a 3D version, which makes a direct comparison between 2D- and 3D-implementations difficult. However, the two approaches used in this study seem to suggest the superiority of the 2D-method.

The difference between 2D- and 3D-implementation was less consistent for the remaining algorithms (cf. Figure 11 and Figure 12), which also make use of the gradient of the image intensities. WT showed higher DSC and lower RVE when implemented in 3D, especially for the lowest sphere, whereas AC only showed a minor difference (< 10% for larger sphere volumes). In turn, AC showed much higher RVE when implemented in 3D. However, none of these observations were statistically significant across the whole dataset. This suggests that the 3D-method for those algorithms could be used in some cases preferably to others. As an example, the WT method could be applied in 3D preferably in the case of small structures (< 8 mL) with low uptake

(TBR<6), as suggested by the results shown on Figure 12. This is in line with the results obtained by Drever *et al.* investigating the accuracy of an axial and tri-axial implementation of the WT algorithm for three spherical inserts at TBR values ranging within 2-15 [120].

The AC method was transposed into 3D in a very straightforward way, only modifying the computation of the force and curvature parameters. However, it could be that this method needs tuning for the 3D case, in order to increase the weight of some parameters, which could have been averaged or smoothed by the 3D effect. In addition, the gradient calculation in AC assumes a cubic voxel size, which is not the case for the images used within this project. An adaptation of the gradient formulation may have led to better results.

This study helped selecting the best implementation method for all algorithms, which is 3D for AT, RG and clustering methods and 2D for GC. In the case of WT and AC, the expected improvement seems limited to a small portion of the image parameters range and it was therefore decided to use the 2D version in this study. However, this work has shown a potential improvement of these methods, and these results could be used in a different study focusing on these two algorithms.

II. A. 2. b. Implementation of a pre-processing step

II. A. 2. b. i. Purpose

One of the challenges in the accurate delineation of GTVs on PET images is due to the PVE inherent to PET imaging. The PVE includes a phenomenon of image blurring in 3D, which causes neighbouring regions of different intensities to "spill-out" into each other, and a resampling effect causing signal emanating from within small or neighbouring regions to be translated into a single combined intensity value. As a consequence, accurate edge detection of tumours on PET images is difficult, particularly for methods based on the identification of gradient crests. Gradient-based segmentation methods are also known to be particularly sensitive to image noise [120], because of the gradient calculation process, which is based on the intensity difference between consecutive voxels. A number of image processing techniques can be used to recover the image quality via de-blurring, which consists in applying a de-convolution filter to the PET image to reverse the effect of the system Point Spread Function (PSF) [72], [121]. Alternatively, some methods already include a form of uncertainty at the lesion edges accounting for this effect, in the form of a fuzzy cluster membership [91]. In addition, smoothing can be applied to reduce the effect of statistical noise in the image. Pre-processing combining the two approaches was applied by Geets *et al.* [98] within their segmentation algorithm. It has the potential to increase the accuracy of segmentation algorithms, and particularly for gradient-based methods. In this study, the effect of such a pre-processing (PP) step on some of the PET-AS methods implemented was investigated.

II. A. 2. b. ii. Experimental protocol

PP was applied in two different steps, reproducing the method used in [98]:

1. First a bilateral de-noising (DN) filter was applied to the image to prepare for the de-blurring step, which is known to increase the noise in the image [122].
2. A de-convolution filter was then applied iteratively to the image in the form of the Van Cittert de-blurring (DB) algorithm.

The bilateral filter used was applied voxel-wise, assigning to each voxel a new value corresponding to a weighted mean of the neighbouring values:

$$I(x) = \frac{1}{C}\sum_{y\in N(x)} I_o(y)w(x,y) \qquad \text{Eq. 16}$$

with I_o the original image, I the filtered image, $N(x)$ the voxels in the neighbourhood of voxel x, $w(x,y)$ a weighting factor assigned to the pair of voxels (x,y) and C the sum of all weights of (x,y) couples in the neighbourhood of x.

The weighting factor for each couple of neighbouring voxels x and y is a convolution of two kernels:

$$w(x,y) = \exp(-\frac{\|x-y\|^2}{2{\sigma_S}^2})\exp(-\frac{(I_o(x)-I_o(y))^2}{2{\sigma_R}^2 I_o(x)}) \qquad \text{Eq. 17}$$

The first kernel, or the space kernel, assigns higher weights to neighbours y closer to x. The second kernel, or radiometric kernel, takes into account the intensity variations in the neighbourhood of x. It results in lower weights for high intensity gradients, which means that gradient crests or edges are preserved from smoothing.

Kernel SDs σ_S and σ_R were chosen so as to approximate the PSF of the scanner, according to measurements made by GE in 2010 at the Wales Research & Diagnostic PET Imaging Centre, using hematocrit capillary tubes filled with ^{18}F at 0 mm and 100 mm distance from the centre of the Field Of View (FOV). The NEMA spheres considered in this study were on average 50 mm away from the centre of the FOV, which led to approximating the PSF by a 6 mm FWHM Gaussian distribution. It is important to note that the presence of $I_o(x)$ in the denominator of the second kernel term is used to account for the fact that the variations in noise (image intensity) follow a Poisson distribution, i.e. the variance σ^2 of the noise at each point x in the image $I_o$can be approximated by the intensity value $I_o(x)$. This adaptation of the standard bilateral filter was used in [98].

Next, the de-blurring step was written as a filter applied iteratively to the image:

$$I^{(n+1)} = I^n + F * (I_o - F * I^n) \qquad \text{Eq.18}$$

with I_o the original image, I^n the deblurred image at step n, $*$ the convolution sign and F a kernel representing the point spread function of the scanner. F was chosen as a 3D Gaussian kernel using the smooth3 Matlab function with a SD of $\sigma = \frac{3}{\sqrt{2\ln(2)}}$ to approximate a 6 mm FWHM Gaussian, and a filter kernel size of 3x3x3 voxels to encompass the whole width of the Gaussian kernel (width of 6σ).

The number of iterations to achieve convergence was chosen as the number of steps after which the relative intensity change per voxel between two consecutive steps was smaller than 5%. The best combination of DN iterations and DB iterations was identified from a selection of cases as the best trade-off between:

- Good mean intensity recovery (<5% true value)
- Good homogeneity in the lesion (SD lower than 10% of the mean value)
- Low computation time.

The PET-AS methods AT, RG, GC, KM, WT and AC were applied and evaluated with test images corresponding to images of the NEMA phantom described previously (cf. II. A. 1. b). Methods FCM or GCM were not considered in this study because their design already includes a form of uncertainty accounting for PVE at the lesion edges (fuzzy membership)[91].

II. A. 2. b. iii. Results

Figure 13 shows the relative intensity change per voxel between the image obtained at each step and the previous image, for up to 100 iterations in the DB filter. This analysis was done on sphere S37 at TBR=3, because of a higher level of noise expected at this TBR. The number of iterations necessary for acceptable convergence was 31 in that case. This finding is consistent with the number of iterations used by Geets *et al.* A total number of 35 iterations were used in the study, to account for further small fluctuations of the change between DB steps.

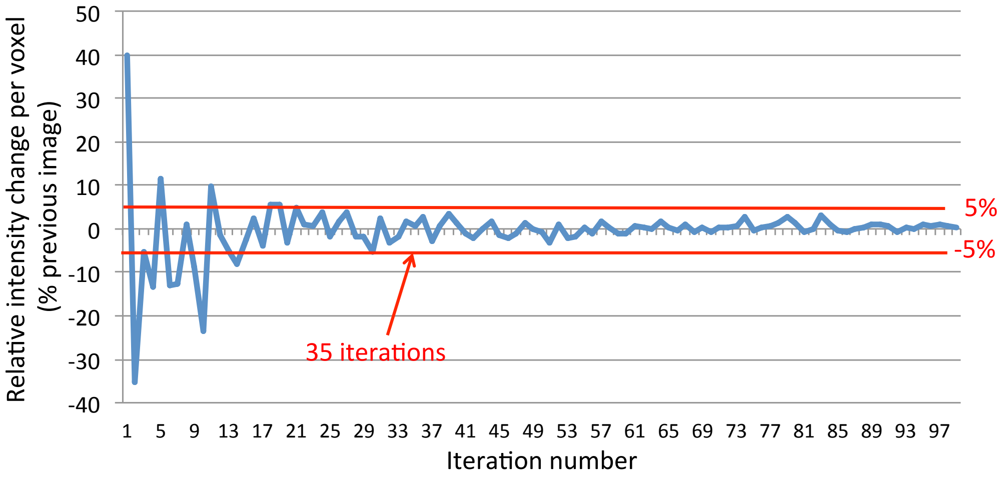


Figure 13. Visualisation of DB filter convergence, and chosen value of 35 iterations (red) required for acceptable convergence (<5% relative intensity change/voxel) at TBR=3.

The effect of the de-blurring and de-noising filters is illustrated on Figure 14. Figure 15 shows the mean sphere intensity recovery obtained for some of the combinations tested on a 3570 voxels image. The difference between the filled-in mean sphere intensity and the measured mean sphere intensity decreased with an increasing number of DB steps. It reached values lower than 5% of the true intensity for three DB steps, with a computation time of 46 seconds for 3570 voxels on a dual core. Further DB steps provided even better intensity recovery, but also higher intensity heterogeneity in the lesion. The SD of voxels within the lesion increased with the number of DB steps, reaching 10% of the mean intensity for 4 iterations. A number of 3 iterations was used for the purpose of this work. The addition of DN iterations improved the voxel homogeneity within the lesion, particularly when applied before the de-blurring. The best result was obtained with two DN followed by three DB steps.

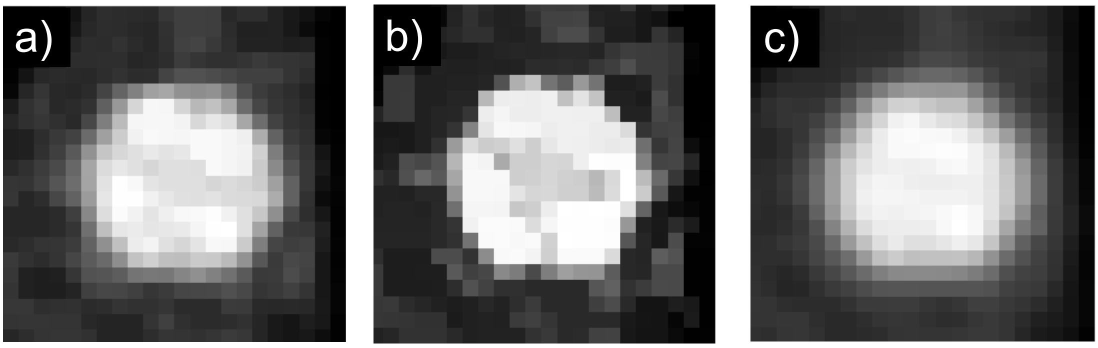


Figure 14. Example of a) original PET image, b) de-blurred PET image using the one DB filter step with 35 iterations. c) de-noised PET image using one iteration of the DN filter.

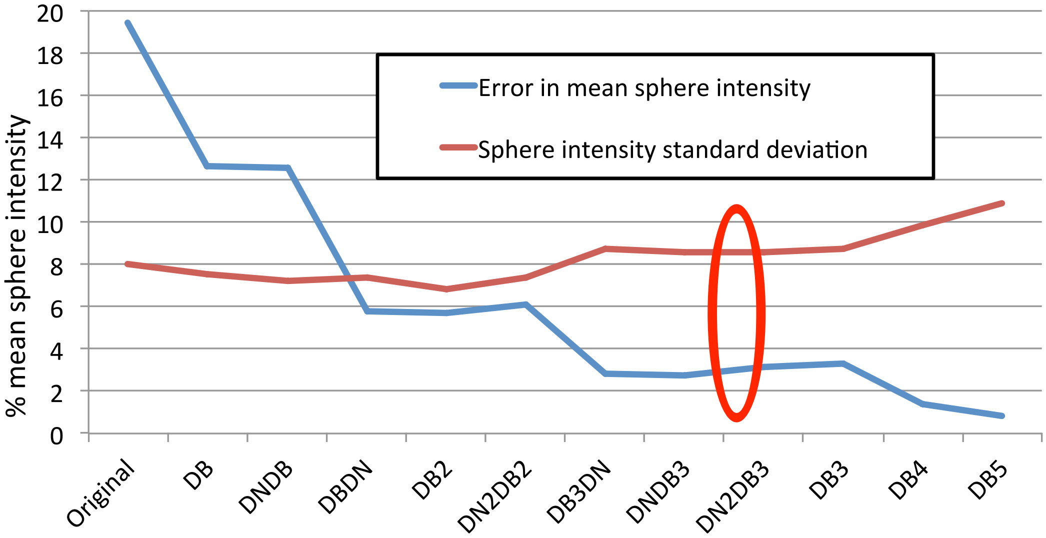


Figure 15. Identification of the optimal DN and DB combination (red) to achieve the lowest error in sphere mean intensity and lowest SD.

Table 3 gives the results of the Mann Whitney U-test for differences in median between RVE and DSC values for each method with and without PP. The p-value is given for the two-tailed test. All methods showed overall higher DSC when PP was not applied, and only AC reached lower absolute median RVE value with the addition of PP. Results obtained with AC were non-significant for both metrics. GC reached significantly lower accuracy when PP was used in terms of both DSC and RVE, with large differences in median. RVE values were significantly different between versions with and without PP for AT, RG and WT while DSC values were significantly different for KM.

	AT	RG	KM	GC	WT	AC
			RVE			
p-value	6.6x10^{-4}*	6.6x10^{-4}*	0.17	0.0048*	0.019*	0.63
U	266	266	206	256	236	1775
Difference in medians	-0.072	-0.20	-0.041	-0.69	-0.26	0.024
			DSC			
p-value	0.091	0.0062*	0.023*	2.7x10^{-7}*	0.13	0.91
U	216	247	234	307	210	165
Difference in medians	0.075	0.090	0.057	0.702	0.076	0.035

*statistically significant

Table 3. Results of the two-tailed Mann Whitney U-test for RVE and DSC testing for differences in median for all six algorithms applied with and without PP. A negative difference in medians corresponds to higher absolute RVE and DSC when PP is applied.

Further analysis was carried out for each algorithm, in order to identify sphere size and TBR ranges for which the de-blurring could improve the performance. RG showed systematically lower DSC and higher absolute RVE when PP was used. AT and KM reached lower accuracy for all cases except for the smallest sphere at TBR=3. In this particular case, PP reduced the RVE from -18% to -12% of the true volume for AT, from -15% to -9.1% for KM, and increased the DSC from 0.098 to 0.15 for AT and from 0.12 to 0.17 for KM. GC reached DSC values more than 20% lower when PP was applied, with a slight improvement in DSC from 0.11 to 0.18 for the smallest sphere at TBR=8. Differences between the segmentation with and without PP were lower than 5% of the value without PP for WT.

Figure 16 shows the improvement in segmentation accuracy of AC obtained by using PP, in terms of both absolute RVE and DSC. Although the difference in accuracy across spheres and TBRs was not statistically significant with the Mann Whitney U-test, accuracy obtained at TBR=3 for the two smallest sphere was up to 40% and 35% higher when PP was used in terms of RVE and DSC respectively. Results at TBR=5 showed no improvement, while values for TBR=8 indicated a reduction of the RVE for the smallest spheres, which was however, translated into a lower DSC.

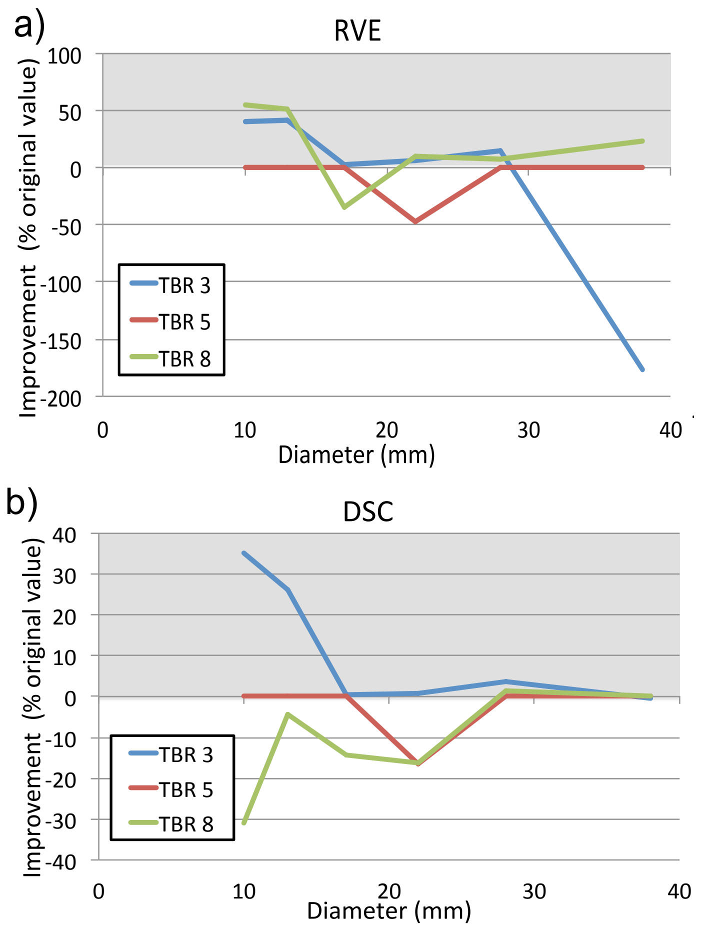


Figure 16. Improvement obtained in terms of a) absolute RVE and b) DSC when using PP prior to AC segmentation, for all TBRs and sphere diameters. The grey plot areas correspond to higher accuracy when PP is used.

II. A. 2. b. iv. Discussion and conclusions

The use of image processing prior to the segmentation was acknowledged and addressed in a number of studies, especially for gradient-based methods. Different PP approaches have been suggested in the literature [69], [123], including iterative combination of segmentation and pre-processing [124]. In this work, one approach containing both de-noising and de-blurring steps was selected, as it previously showed promising improvement of gradient-based techniques [98]. The different filter combinations tested were successfully implemented in the code for each delineation method, and did not add any significant computation time. The delineation performance results obtained varied across PET-AS methods and filter combinations used but were useful to determine which algorithms could be improved by the addition of a PP step, and for which TBR and object size. However, this study was limited to spherical fillable inserts, as it was aimed at gaining a better understanding rather than fully investigating the use of PP for PET-AS methods.

The best performing de-convolution method was slightly different from the one used by Geets *et al.* In this work, the optimal number of de-blurring iterations was 35, and two de-noising steps were applied followed by three de-blurring steps. This can be explained by the difference in the images considered, particularly in terms of noise, as the study by Geets *et al.* used patient data, a Siemens Exact HR camera, without TOF correction, and different acquisition times.

Results showed that the use of PP did not improve the overall segmentation accuracy on a large range of TBRs and sphere sizes, although the segmentations provided reached largely different accuracy values compared to situations with no PP (cf. Table 3). For AC, which is based on the values of the gradient image for the definition of the force deforming the adaptable contour, PP appeared to improve the delineation at the smallest TBR and for the two smallest spheres (cf. Figure 16). Even for AT and KM, which are not based on the gradient image but make use of the global intensity value in the target and background, the use of PP proved beneficial for the smallest sphere at the lowest TBR. The PP did not improve the delineation of non-gradient based algorithm RG, nor of WT and GC except in one case. However, the accuracy of these PET-AS methods was not negatively affected by the use of PP. This work has shown the potential benefits of using a PP step before the segmentation in some specific cases, which correspond mainly to small target objects and low TBR values. This is in line with the fact that the PVE is strongest in those cases. However, these situations might not be representative of typical clinical data. As an example, primary tumours often present as large lesions with a high TBR. The use of PP could be useful as an optional tool in specific cases where lymph nodes detected on a different modality, but with a low PET uptake need to be delineated on PET.

II. A. 2. c. Algorithms initialisation

II. A. 2. c. i. Purpose

The algorithms developed in this project are designed to be fully automatic. However, the presence of the brain and other high intensity regions in some H&N scans leads to the necessity of excluding these areas in the segmentation process. In their investigation of a framework for segmentation evaluation [125], Udupa *et al.* insist on the difference between lesion recognition, which consists in broadly identifying the area to segment, and lesion delineation, corresponding to the detection of the lesion's edges. The recognition process is necessary at some point in the segmentation process. Some methods suggested in the literature leave the recognition task to the clinician, after the delineation algorithm generated a number of contours as a result of segmenting the whole clinical image [72], [96], [104]. However, the lesion recognition is most often applied before any delineation takes place, by applying the delineation to a specific volume of interest (VOI) corresponding to a portion of the whole scan including the target lesion. In this work, all algorithms implemented are therefore applied to a VOI defined by the operator as a rectangular box. For the PET-AS methods implemented, the operator input was reduced to a minimum, corresponding to the initialisation VOI selection. The work described in this section aimed at evaluating the degree to which this specific step of the process makes the PET-AS methods operator dependent. For this purpose, the operator variability caused by different choices for the initialisation box was evaluated for the different PET-AS.

II. A. 2. c. ii. Methods

For this work, PET images acquired with the NEMA IEC body phantom were used. The phantom was filled with a higher concentration of FDG in the six plastic spherical inserts to achieve a TBR of 8. The known diameters of the spheres and the high resolution CT image were used to generate ground truth contours for the different inserts. All segmentation approaches described above were applied to all spheres for the following different initialisation VOIs:

- Bounding box of the true contour with a 2.5 mm, 5 mm, 10 mm and 17.5 mm margin. These cases are named B025, B050, B100 and B175 respectively.
- Bounding box of the true contour with a 10 mm margin shifted by 10 mm in left-right, antero-posterior and superior-inferior directions. These cases are named BshiftLR, BshiftAP, and BshiftSI respectively.

For each insert the difference in volume was calculated as well as the DSC between contours obtained for case B100 and with the other initialisation VOIs.

II. A. 2. c. iii. Results

Figure 17 shows results of the segmentation by AT of the largest sphere in the NEMA phantom (S37), and the outline of the corresponding initialisation VOI for B025, B050, B100 and B175, and for BshiftLR, BshiftAP and BshiftSI respectively. Table 4 gives for each method tested the average across all inserts of the DSC values and difference in volume (as a percentage of the true insert volume) between the contours generated with initialisation VOI B100 and other initialisation VOIs. The fixed thresholding methods tested showed very little variation with the initialisation VOIs, with less than 1.1% difference of the total volume for a maximum difference in VOI dimensions (in width, height and depth) of 7.5 mm. Similarly, FCM proved very robust to the initialisation, with less than 2% difference on average across all spheres, in all VOI cases. All PET-AS except GCM returned contours with conformity to the contour for B100 higher than 0.95 DSC, and less than 15% difference in volume on average, except for B025 in the case of RG, KM and WT. For all PET-AS, a shift of 10 mm in the box position did not affect the final result more than 10% of its volume on average on the different inserts, with conformity to the B100 contour higher than 0.95 DSC. The largest differences were observed for GCM, which reached a difference in volume of 7.7

mL between the contours obtained from B025 and B100 for the largest sphere, corresponding to 29% of the true volume.

Method	FT42	FT50	AT	RG	KM	FCM	GCM	WT	AC
					DSC				
B025	0.99	1.00	0.90	0.77	0.88	0.99	0.59	0.89	0.92
B050	1.00	1.00	0.96	0.95	0.94	0.99	0.76	0.96	0.93
B175	1.00	1.00	0.99	0.98	0.97	0.97	0.83	0.99	0.95
BshiftLR	1.00	1.00	1.00	0.99	0.98	1.00	0.97	0.98	0.99
BshiftAP	1.00	1.00	1.00	0.99	1.00	1.00	0.96	0.97	0.98
BshiftSI	1.00	1.00	0.99	0.99	0.98	1.00	0.95	1.00	1.00
				Volume difference (% true volume sphere)					
B025	2.02	0.39	13.50	41.95	24.47	1.31	54.70	19.99	10.11
B050	0.00	0.00	5.88	9.58	13.45	0.93	35.87	8.16	5.22
B175	1.04	0.60	2.16	5.23	7.09	1.79	40.15	1.88	5.76
BshiftLR	0.00	0.00	0.16	1.83	3.79	0.06	6.09	3.46	0.85
BshiftAP	0.86	0.00	0.27	1.66	0.85	0.19	8.15	6.51	1.62
BshiftSI	0.00	0.00	0.84	2.85	3.95	0.00	8.98	0.00	0.00

Table 4. Effect of the use of a different initialisation VOI on the contour conformity (DSC) and volume for the different PET-AS methods tested. Values are given as an average for the six spherical inserts of the NEMA phantom.

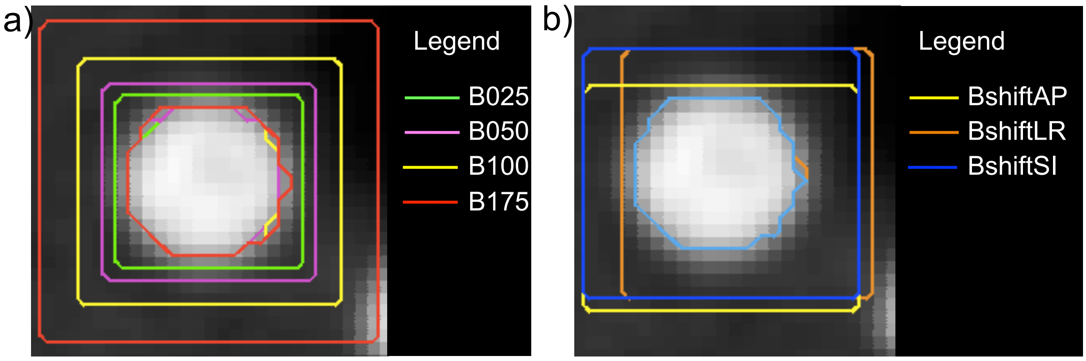


Figure 17. Initialisation box and associated segmentation result (same colour) for sphere S37 by AT for a) B025, B050, B100 and B175 and b) BshiftAP, BshiftLR and BshiftSI.

II. A. 2. c. iv. Discussion and Conclusions

This study was carried out to gain a better understanding of the relative sensitivity of the different PET-AS methods to the initial VOI used in the segmentation process. A more exhaustive study would require investigating this effect for more complex and realistic target objects, but this was not the aim of this work.

The PET-AS methods used showed a good robustness to the initialisation VOI, which was varied between extreme cases in these experiments. All PET-AS methods were particularly robust to a shift in the localisation of the initialisation VOI, but depended more on the VOI size. Initialisation VOI B025 is very tight around the target object, and represents an extreme case of initialisation, unlikely to be selected by an operator. Initialisation VOI B175 represents the other extreme of a very large initialisation, which even includes a couple of high intensity voxels belonging to the neighbouring sphere S5. This can explain some of the large differences observed for these initialisations. For B050, only GCM reached differences larger than 15% of the sphere volume on average. This is expected for clustering methods, which consider the voxels in the background at each step of the segmentation. Gradient-based and threshold based PET-AS, on the contrary, focus more on the object edges and are therefore less influenced by the background voxels. The small effect of the initialisation on AT, compared to FT42 and FT50 is due to the fact that the mean background intensity is calculated at each step to update the threshold to apply. The fuzzy voxel membership calculated in FCM allows the algorithm to be less dependent on the intensity distribution in the background and tumour clusters, as "edge" voxels will have a lower weight in the calculation of the cluster mean intensity at each step. GCM, however, is particularly sensitive to the inclusion or exclusion of background voxels, as these will modify the parameters of the Gaussian distribution fitted to each cluster at each step.

Although a certain influence of the initialisation VOI was observed for some of the PET-AS methods, it remains negligible in most cases. In particular, the variations observed are very small compared to variability values reported in the literature between manual segmentation by two different operators, which can reach 90% of the lesion volume delineated with both CT and PET data [53], [126]. In addition, the variability of the initialisation VOI definition is likely to be small, especially in the H&N because of the presence of other high intensity uptake regions, which operators won't include in the initialisation.

Eventually, non-rectangular initialisation VOIs could also be defined, to account for nearby structures that should not be considered in the segmentation process. This approach was considered at a later stage (cf. IV. B).

II. A. 3. Investigation of the influence of several image parameters on the segmentation results

II. A. 3. a. i. Purpose

The accuracy of PET-AS algorithm depends on the characteristics of the target object. The segmentation of large objects with high contrast to the background often leads to higher accuracy scores than more difficult situations of smaller and less intense objects. Although a number of published PET-AS methods have been validated on phantom data, this is often done for a limited set of imaging conditions. As an example, the data used for the validation of the FLAB method was only derived for two different contrast ratios and three different image noise values [91]. Studies often focus on the variation of the segmentation accuracy with a single image parameter, such as TBR or object size [127]. The investigation of the segmentation accuracy for varying image parameters is crucial for understanding the performance of PET-AS methods and for proper inter-comparison with other algorithms. This series of experiments aimed at evaluating the impact of some key image parameters on the performance of the PET-AS methods developed for this project, using extensive ranges of values for each parameter.

II. A. 3. a. ii. Experimental protocol

Experimental data were acquired for the NEMA IEC body phantom with six fillable spheres of different diameters described in II. A. 1. b) for 8 different TBRs of 1.5, 2, 2.9, 4.3, 4.9, 5.5, 7.4 and 9.3.

After acquisition, a range of numerical noise realisations was simulated for the case of TBR=4.9. The realistic noise values added were determined from 10 clinical

H&N scans, by calculating the coefficient of variation (COV) in homogeneous Regions of Interest (ROIs) such as fat and muscle tissue. The COV is defined as:

$$COV = \frac{\sigma}{\mu} \qquad \text{Eq. 19}$$

where μ is the mean intensity value and σ the SD of the measured intensities. It is expressed in this work in % of the mean intensity value. A template image was made from the PET image obtained at TBR=4.9 by assigning to both background and spheres regions their mean intensity value in the original scan. Gaussian noise was added as a random number generated from a normal distribution with parameters μ and σ corresponding to the noise level to apply. The image obtained was then smoothed with a Gaussian kernel to reproduce the effect of the scanner PSF. The kernel size was chosen to be 3x3x3 voxels to model the scanner PSF of 6 mm. The method is illustrated on Figure 18.

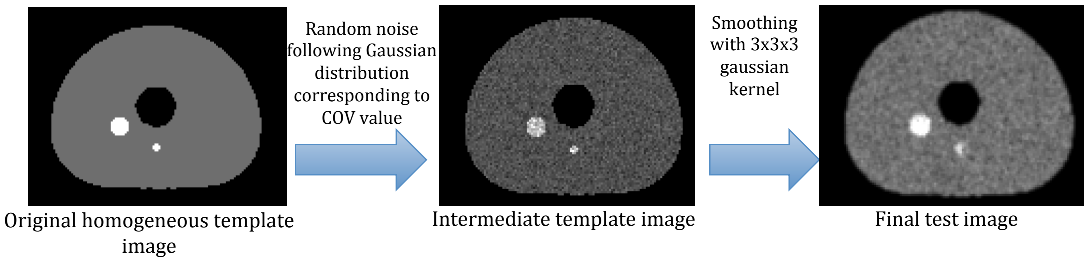

Random noise following Gaussian distribution corresponding to COV value
Smoothing with 3x3x3 gaussian kernel
Original homogeneous template image
Intermediate template image
Final test image

Figure 18. Method for the addition of numerical noise to the original template image extracted from the case of TBR=4.9.

PET-AS methods AT, RG, KM, FCM, GCM, WT and AC were applied to the test images obtained. In addition, the fixed thresholding methods FT42, FT50 and SUV2.5 were applied to the same images, to represent standard PET image segmentation methods. RVE and DSC were calculated for each case, and the values were compared across segmentation algorithms.

For the images containing additional numerical noise, the variation of the methods' accuracy with the COV was assessed by fitting the curves representing the average DSC on the six spheres for each COV values with a linear regression. The resulting slope and R^2 coefficient were calculated and compared.

The Friedman test for k related samples available in SPSS was used to check for statistically significant differences among the methods, while the Sign test was used to compare each method with FT42 and FT50 in terms of RVE and DSC. Non-parametric tests were chosen because the values obtained did not follow a normal distribution for most methods.

II. A. 3. a. iii. Results

A total of 48 test images were obtained for the six sphere sizes and 8 TBRs. In addition, 10 synthetic images of the NEMA phantom with added numerical noise were obtained, each one containing six spheres. The results obtained for the resulting 108 test images generated were first checked visually for non-usable contours. These corresponded to the following cases:

- the contours were empty,
- the contours did not include any voxel of the ground truth,
- the contours included the whole input image.

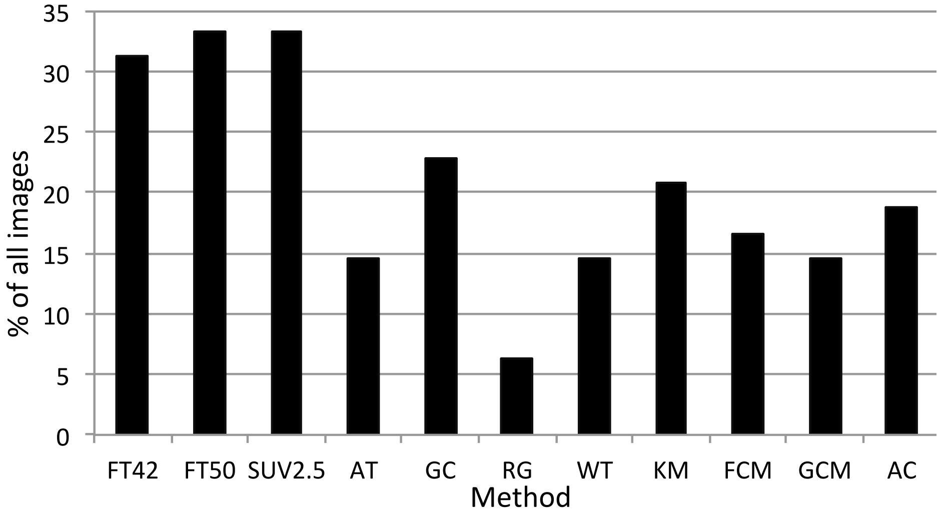


Figure 19. Percentage of images with non-usable contours returned for each algorithm.

The proportion of non-usable contours obtained by each method is displayed on Figure 19. In the following analysis, these contours, for which the metrics could not be calculated or would not be relevant, were discarded.

The average values across all images for RVE and DSC, S and PPV and HD for each PET-AS method are presented on Figure 20, Figure 21 and Figure 22 respectively. Higher absolute RVEs were obtained for FT42 and FT50 compared to the PET-AS methods, as well as lower average DSC (except for GCM). On Figure 20, large error bars, which do not cross the line corresponding to 0% RVE, are observed for FT42, FT50 and KM. Average RVE lower than 50% were obtained by GC, RG, WT, FCM and GCM. The highest average (SD) DSC of 0.76 (0.10) was obtained for RG, with an average (SD) negative RVE of 37% (99%). GC reached the lowest average (SD) RVE of 5.5% (22%), with an average (SD) DSC of 0.66 (0.14). PET-AS methods also showed higher average PPV (above 0.66 for all methods) and lower S (below 0.86 for all) compared to the thresholding methods tested. RG reached the lowest average (SD) HD value of 0.20 cm (0.10 cm). Only AC yielded higher average HD (0.41 cm) compared to the thresholding methods, with a SD of 0.25 cm.

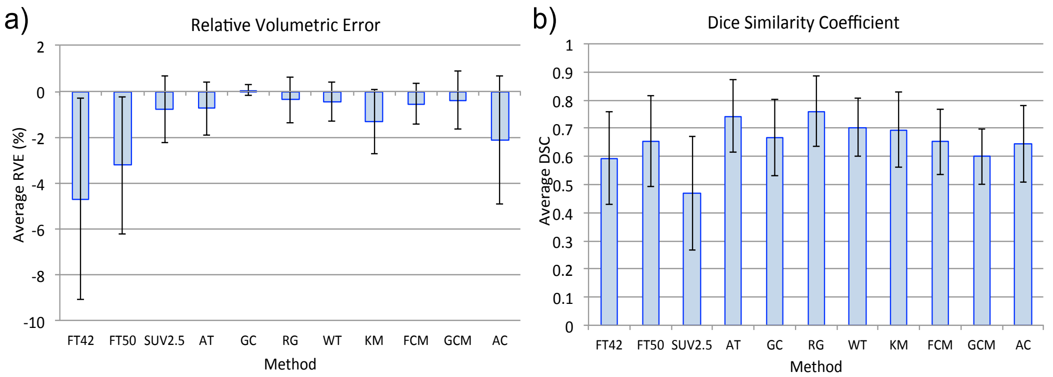


Figure 20. Average values for a) RVE and b) DSC across all images for the different PET-AS methods. Error bars correspond to one SD of the values.

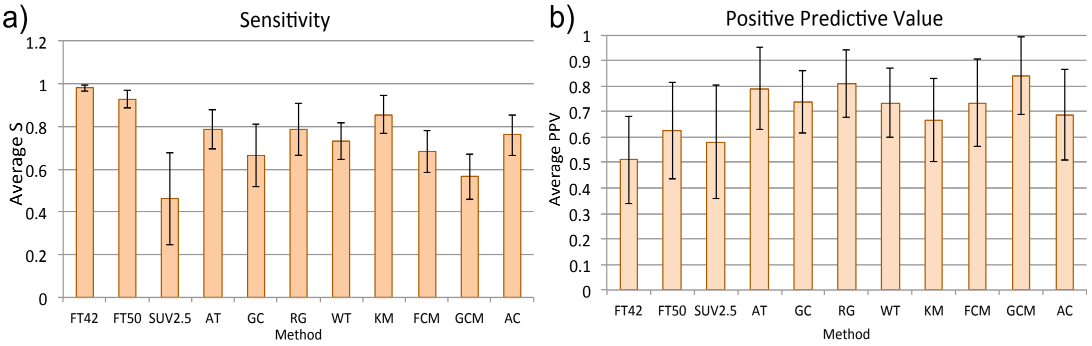


Figure 21. Average values for a) S and b) PPV across all images for the different PET-AS methods. Error bars correspond to one SD of the values.

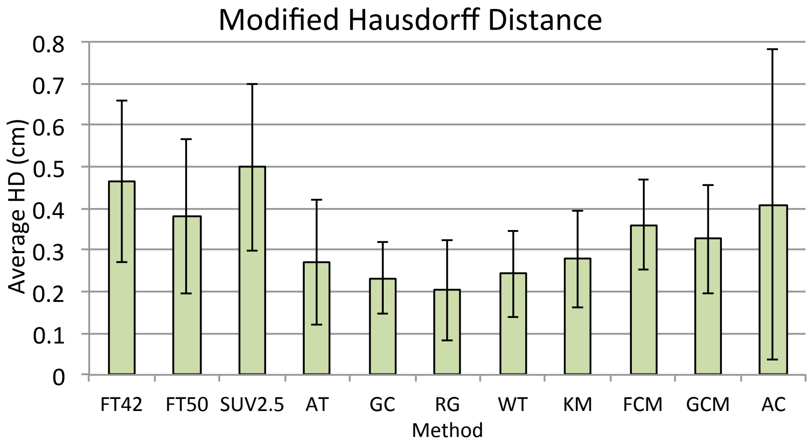


Figure 22. Average values for HD across all images for the different PET-AS methods. Error bars correspond to one SD of the values.

The Friedman test for k related samples, used to check for statistically significant differences among the methods, returned a p-value smaller than 10^{-4} for both DSC and RVE values. Table 5 summarises the results of the pairwise comparisons of each of the PET-AS methods with FT42 and FT50 using the pairwise Sign test. RVE values were significantly higher for all PET-AS compared to FT42, and for AT, GC and RG compared to FT50. All PET-AS reached higher median DSC values compared to FT42, which was significant for AT, GC, RG, KM, GCM and WT. RG also reached significantly higher median DSC than FT50.

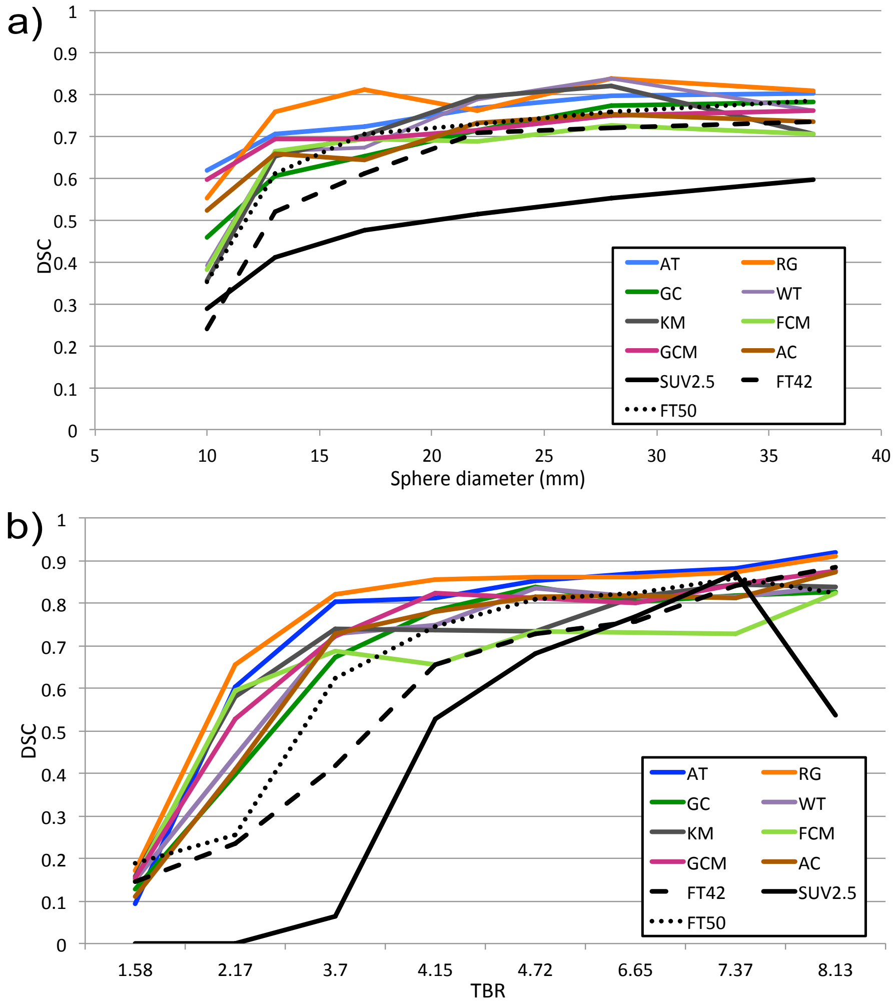


Figure 23. Influence of a) sphere size and b) TBR on DSC values for each PET-AS method. Results are averaged on TBR and sphere sizes respectively.

Figure 23 illustrates the influence of sphere size and TBR on the DSC obtained for each PET-AS method. The influence of noise on the delineation performance, using the 60 test images generated for six spheres with 10 levels of numerical noise, is shown on Figure 24. Due to the definition chosen, a high COV value corresponds to a noisy image. SUV2.5 was discarded from this particular study because the method failed to delineate the target images in more than 50% cases. The delineation accuracy of the different PET-AS methods decreased with increasing COV. The lowest DSC values were reached by FCM, while AT reached the highest values for COV lower than 30%.

	PET-AS Method	**AT**	**GC**	**RG**	**KM**	**FCM**	**GCM**	**WT**	**AC**
		Comparison to FT42							
RVE	**p-value**	$<10^{-4}$ *	$<10^{-4}$ *	$<10^{-4}$ *	$<10^{-4}$ *	$<10^{-4}$ *	$<10^{-4}$ *	$<10^{-4}$ *	$<10^{-4}$ *
	Difference in medians	0.122	0.128	0.193	0.085	-0.039	0.023	0.161	0.019
DSC	**p-value**	0.009 *	0.211	$<10^{-4}$ *	0.002 *	0.312	0.02 *	0.01 *	0.16
	Difference in medians	-0.11	-0.072	-0.12	-0.062	-0.016	-0.066	-0.078	-0.047
		Comparison to FT50							
RVE	**p-value**	$<10^{-4}$ *	0.001*	0.001 *	0.024 *	$<10^{-4}$ *	$<10^{-4}$ *	0.292 *	$<10^{-4}$ *
	Difference in medians	0.0016	0.0082	0.073	-0.035	-0.16	-0.097	0.041	-0.10
DSC	**p-value**	0.196	0.461	$<10^{-4}$ *	0.094	0.186	0.961	0.723	0.581
	Difference in medians	0.013	0.043	-0.00065	0.054	0.098	0.049	0.037	0.069

*Statistically significant (*p<0.05*)

Table 5. Results of the Sign test for paired samples of RVE and DSC values for each method tested against FT42 and FT50, with a positive difference in medians corresponding to higher RVE and lower DSC for the thresholding method.

Table 6 shows the results obtained by a linear regression of the average DSC obtained at each COV for the different methods. The R2 value obtained for FT42 was very low and does therefore not allow drawing any conclusions. For the other methods, the highest slope coefficient (a) value was obtained for KM. Methods GC, KM, and WT all reached coefficient absolute values higher than 0.03. The lowest value for the slope coefficient (a) was obtained for FT42, followed by FCM. Methods FT50, AT and AC all reached absolute values for coefficient (a) lower than 0.02.

	Regression model: DSC= a*(COV)+b									
	FT42	**FT50**	**AT**	**GC**	**RG**	**KM**	**FCM**	**GCM**	**WT**	**AC**
a	-0.0056	-0.019	-0.019	-0.030	-0.023	-0.033	-0.011	-0.028	-0.032	-0.019
R^2	0.092	0.71	0.67	0.81	0.77	0.75	0.49	0.91	0.845	0.79

Table 6. Results of linear regression of the average DSC values obtained at each COV for the different PET-AS methods.

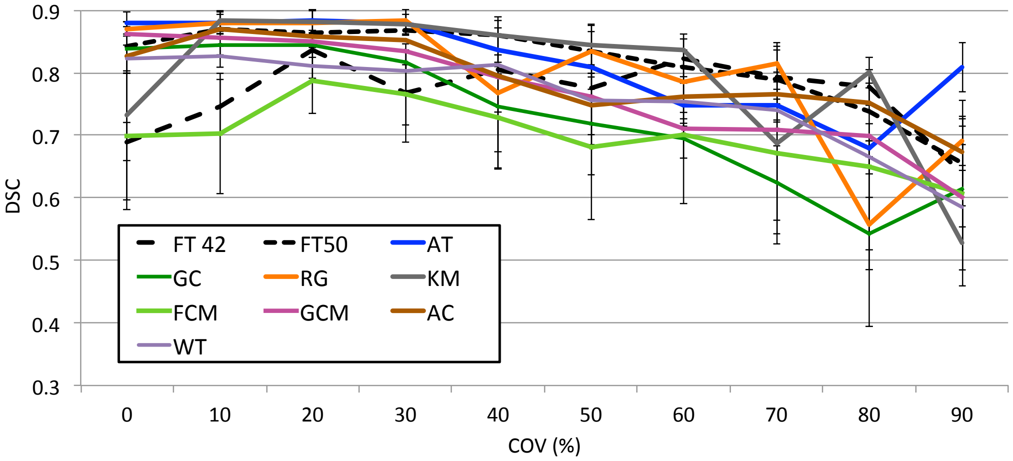


Figure 24. Average DSC on all spheres at different COVs for the different PET-AS methods tested. Error bars given represent one SD of the values.

II. A. 3. a. iv. Discussion and conclusions

This study aimed at evaluating the PET-AS methods implemented with data covering a wide range of image parameters. Eight images, on which some of the algorithms produced non-usable contours for calculating RVE, were discarded when showing the corresponding results. Method SUV2.5 in particular was unable to appropriately delineate any of those images, and returned particularly low accuracy values throughout the study. Method SUV2.5 was suggested in the literature for the delineation of lung tumours, which are most commonly in a cold or very low activity background. However, this project focuses on delineation of H&N tumours, for which the background intensity is much higher than in the lungs. SUV2.5 appears unsuited to this type of data, and was therefore ignored in the rest of this project.

Results on Figure 20 a) and b) show RVEs of less than 100% (except for KM and AC) and good overlap between delineated structure and ground truth (DSC>0.6) for all eight algorithms developed. These results include the delineation of challenging cases with spheres of less than 1 mL volume and TBR lower than 2. These cases were used to investigate the robustness of the PET-AS methods to extremely challenging cases. More detailed analysis on Figure 20 shows that the PET-AS methods reach DSC values higher than 0.7 for TBR higher than 3.7 and spheres of diameter larger than 15 mm. RG was identified as the most promising method, with high average DSC and low RVE, followed by AT showing very good performance as well. RG also performed better than FT42 and FT50 for both DSC and RVE, which was statistically significant as shown in Table 5. The lowest DSC was reached by GCM, while AC had the highest average RVE.

On average over the whole dataset, PET-AS methods appear to perform better than the three standard thresholding methods FT42, FT50 and SUV2.5. All PET-AS methods implemented reached average DSC values higher than the thresholding methods. This was confirmed by lower PPV values shown on Figure 21 and higher HD values on Figure 22. The low RVE values on Figure 20 a) for SUV2.5 are due to the fact that a large number of cases for which the RVE could not be calculated were discarded. The higher accuracy of the advanced PET-AS methods was statistically significant compared to FT42 for all methods, and to FT50 for RG. Results on Figure 20 a) also showed that the RVE values vary within a range of negative values for FT42, FT50 and KM, while the line corresponding to RVE=0 is crossed for all other methods. This shows that systematic errors occur only for these methods while other PET-AS have been tuned to avoid such systematic errors. Information provided by more specific indicators (cf. Figure 21 and Figure 22) suggested how some algorithms could be improved. As an example, KM has the highest S of all methods, but one of the smallest PPV, suggesting that the contours generated by KM were in average larger than the true contour. Similarly, GCM might reach higher DSC by increasing its sensitivity. In the case of KM however, only parameter the clustering stopping criterion can be modified, but this did not help improve the performance of the PET-AS.

Figure 23 a) and b) and Figure 24 illustrate the strong influence of the sphere size, TBR and noise level on the delineation accuracy, showing an overall decrease of the delineation accuracy with noisier images, smaller spheres and lower TBR for all methods. This study was useful to identify the most robust methods as well as the critical range of parameter values for which the algorithms reach different accuracy values. In terms of the influence of TBR, all methods reached low average DSC with very similarly low values at the lowest TBR. Differences occur at TBR higher than 2 where RG reaches a value of 0.66, while GC does not attain 0.40. The curves show large variations for TBR lower than 4, with most methods reaching a saturation of their DSC values at higher TBRs. This was the case for AT and RG at TBR higher than 3.7, for GCM at TBR higher than 4.15 and for WT at TBR higher than 4.72. This type of trend is representative of the PVE described in II. A. 2., which hampers the delineation process, and is particularly strong for small spheres at low TBRs. Thresholding methods still returned DSC values lower than 0.30 at TBR of 2.17. Similar trends are observed for the influence of the sphere size on the delineation accuracy on Figure 23 a). RG reached its threshold accuracy of 0.65 DSC for spheres larger than 13 mm diameter. In this case, the methods can be differentiated at the lowest sphere size, where AT reached a DSC of 0.69 compared to the lowest DSC of 0.35 reached by KM. Methods GCM and AT appear most robust to the sphere size, with a maximal difference in DSC across sphere sizes of 0.14 and 0.18 respectively. KM in comparison, reaches a maximum difference across spheres of 0.47 DSC, making it the method least robust to sphere size on the dataset analysed. It is also interesting to note that WT, AC, and GC, which were implemented as 2D processes, showed poorer performance than the 3D methods at TBRs lower than 4.7. This may be due to lower accuracy on some slices near the object axial boundaries, for which the TBR measured in 2D appears lower.

The robustness of the methods to the addition of Gaussian noise in the image was quantified using a linear regression, as the curves obtained on Figure 24 were more suited to this type of analysis. The results of the regression given in Table 6 highlighted FCM, AC and AT as the methods most robust to noise, whereas WT, KM and GC varied most with the addition of noise in the image. In the case of GC and WT, this is likely to be due to the use of a single seed voxel in the process, which is applied to both background and tumour for WT. In addition to this, GC follows voxel-wise process, making it highly dependent on single voxel values, and therefore on noise. KM showed large fluctuations at COVs larger than 60%, but remained very stable with high DSC values for lower noise levels. The remaining methods make more use of mean region values, which may explain their higher robustness to noise.

The results of this study provided a validation of the PET-AS methods implemented, and showed their superiority to commonly used fixed thresholding methods FT42, FT50 and SUV2.5. In addition, these results have highlighted the most promising methods, such as AT and RG, as well as the weakest ones, such as FCM, GC and WT. Promising results were also obtained in specific cases, such as for GCM and AC in terms of robustness to image parameters. The test images generated provided a good starting point for the validation and comparison of the PET-AS algorithms, but have little clinical relevance due to the fixed geometry of the inserts and their thick plastic walls. Work presented in section II. B was aimed at addressing these issues.

II. B. Evaluation of PET-AS methods using thin plastic wall inserts

II. B. 1. Investigation of the effect of cold plastic walls on image quantification and segmentation

II. B. 1. a. Purpose

Fillable plastic inserts are commercially available for use in PET phantom studies. Such inserts are useful in order to simulate hot regions in a “cold” or “colder” background, by filling the inserts and the background phantom with different radioactivity concentrations. Plastic inserts are increasingly used as a validation or calibration tool within PET studies, as they provide a convenient way to perform such simulations because of their known ground truth dimensions. One of the main drawbacks of these inserts is the plastic wall, usually a few millimetres thick, which separates the inner active region from the background active region. Due to the lack of activity in these walls and the low resolution of PET images, the overall activity in the spheres observed on the PET scans is different from the actual activity injected. This is due to the PVE causing both hot region and background region to "spill out" into the cold region, and is especially true for small spheres for which the wall thickness represents a high percentage of the overall sphere volume. As a consequence, the images are quantitatively biased, with a reduced activity recovery [128] compared to a wall-less case, and are not representative of a clinical situation where cold walls do not separate active regions from background tissue. These issues have been discussed in several documents, some of which suggested the use of different types of inserts, or assessed the impact of inactive sphere walls on quantitative image analysis methods [84], [85], [129], [130]. Bazanez-Borgert *et al.* [130] reported up to 21% higher activity recovery when using wall-less spheres compared to plastic inserts. The groups of F. Hofheinz [84] and J. van Dalen [85] derived theoretical models of PET-intensity profiles, which included the presence of cold walls, and applied those to the investigation of the effect of the inactive shells on threshold-based volume recovery. Hofheinz *et al.* concluded on the systematic bias introduced when using phantoms with standard plastic inserts for the calibration of optimal thresholding algorithms at finite background levels. Although these works provided thorough investigations of the cold walls phenomenon, further systematic work is required on this subject. In particular, its influence should be investigated in some more extreme imaging conditions such as low TBRs and small sphere sizes, and the effect of the inactive wall thickness should be quantified. There is also a need for a study of the impact of the cold walls effect on the delineation of volumes with automatic methods other than thresholding. This study aimed to quantitate the influence of cold walls thickness on physical phantom PET images, for an extended range of TBRs and sphere sizes, and at different wall thicknesses. In this work the effect of the thickness of cold plastic walls on a range of fully automatic segmentation algorithms was also evaluated.

This work was peer reviewed and published in [131].

II. B. 1. b. Methods

A custom-made phantom, consisting of a cylindrical plastic body containing six removable inserts, was designed for the purpose of this study. Inserts from the Liqui-Phil body phantom (The Phantom Laboratory, Salem, USA) were used, as well as in-house made plastic inserts with inner diameters replicating the Liqui-Phil set (10 mm, 15 mm, 20 mm, 30 mm, 38 mm and 58 mm). These six spheres were manufactured with a "vacuum – moulding" technique allowing wall thicknesses of about 0.2 mm, compared to wall thicknesses of 1-2 mm measured on the commercial spheres. Plastic wall thicknesses were measured for both sets using a digital calliper (Absolute Digimatic, Mitutotyo UK Ltd, Andover, UK) with a precision of 0.01 mm. Five measurements were made and averaged for each sphere. The inner volume of each insert was measured by weighing the spheres empty and filled with water. The phantom is pictured on Figure 25 a), followed by the 58 mm diameter sphere of each set on Figure 25 b).

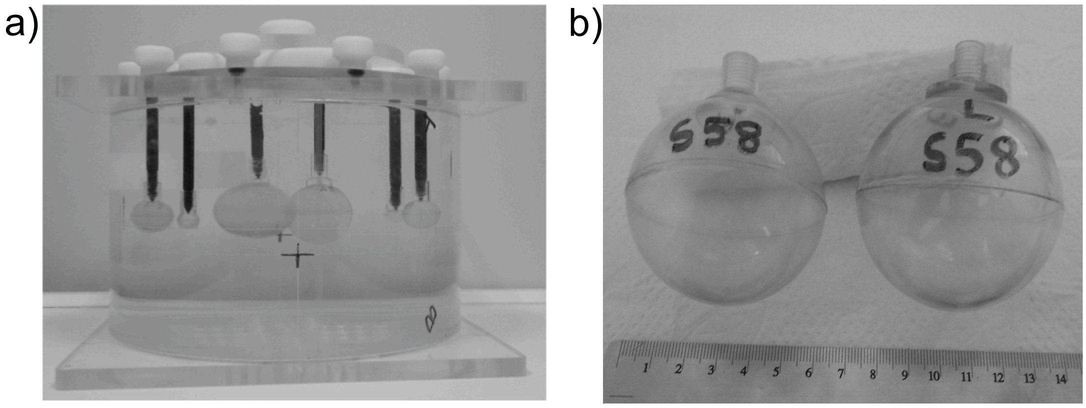


Figure 25. a) Picture of the custom phantom used containing six spheres, and (b) of the 58 mm diameter sphere of each set (vacuum-moulded on the left, commercial on the right).

Six different TBRs, ranging from 2 to 7, were targeted by varying the activity injected into the spheres while keeping the same background activity concentration of 5 kBq/mL. This value is recommended for the use of NEMA phantom devices as an approximation of the typical background uptake observed on clinical data [132]. For each of the six sphere sizes, the corresponding thin- and thick-wall inserts were scanned simultaneously at the six different TBRs targeted. The analysis was thus performed on 72 different test images corresponding to a certain sphere size, wall type (thin or thick) and TBR. All scans were acquired in house with the parameters described in II. A. 1. a.

2D PET intensity profiles were generated with the method used by Hofheinz *et al.* [84], which is based on the convolution of the true profile with a Gaussian distribution simulating the PSF of the scanner. The equation used to derive the radial profile P of a sphere of radius R (i.e. intensity as a function of the distance r to the centre of the sphere) can be expressed as a function of the transformed radial coordinate $z = \frac{2\sqrt{\ln{(2)}}}{FWHM} r$, and the transformed radius $Z = \frac{2\sqrt{\ln{(2)}}}{FWHM} R$. For unit activity and no background activity, this gives:

$$P_Z(z) = \begin{bmatrix} \operatorname{erf}(Z) - \frac{2}{\sqrt{\pi}} Z e^{-Z^2} & if\ z = 0 \\ \frac{1}{2}[\operatorname{erf}(z+Z) - \operatorname{erf}(z-Z)] - \frac{1}{2\sqrt{\pi}} \left[\frac{e^{-(z-Z)^2} - e^{-(z+Z)^2}}{z} \right] & if\ z \neq 0 \end{bmatrix} \qquad \text{Eq. 20}$$

with FWHM the Full Width at Half Maximum of the scanner PSF, and "erf" the error function.

The radial PET-intensity profile A in the case of plastic inserts of wall thickness w is then calculated by superposing the profiles obtained for the background of activity B, the sphere of activity S, and the inactive wall:

$$A(z) = \mathrm{S} \cdot \mathrm{P}_Z(z) - B \cdot \mathrm{P}_{Z+w}(z)\ + B \qquad \text{Eq. 21}$$

These profiles were calculated for each of the spheres, knowing their diameter and wall thickness, and for each TBR, with B=1 and S the ratio between activities injected in the spheres and in the background. According to measurements performed on the scanner by the manufacturer the PSF of the scanner is assumed to be a Gaussian with FWHM of 6.4 mm in the axial direction and 5.30 mm in the trans-axial direction at 100 mm off the centre of the Field Of View. In this work, the isotropic FWHM value of 6 mm was used at the position of the spheres, in order to use the equation provided by Hofheinz *et al.*

From these profiles, expected values of the maximum, mean and peak intensity were extracted. Mean intensity was calculated by taking into account all voxels with a centre located within the sphere modelled, while the peak intensity corresponds to the mean intensity of a 3x3x3 voxels cube around the maximum value. In addition, the Recovery Coefficient (RC: ratio between the true activity and mean activity measured on the PET image) and the apparent diameter of the spheres (AD), defined as the background-subtracted FWHM of the intensity profiles generated, were calculated. The location on the profile of the maximum intensity gradient point was also extracted for each insert type and compared to the wall-less case to obtain a value of the displacement due to the cold walls effect.

In order to assess the influence of the cold walls thickness on PET image intensity, the test images obtained were analysed qualitatively by using intensity profiles, as well as quantitatively with SUV-based indicators and RCs. In this study, PET-image intensity was converted to SUVs by entering the phantom weight and dimensions into the scanner protocol.

PET intensity profiles were derived on the scans obtained across the different spheres, in order to visualize any difference in intensity distribution within spheres of both sets. Lines were drawn on the middle slice for both thin- and thick-wall inserts, and the AD of the spheres was derived as described previously. This was done five consecutive times at different angles, and the average AD values are reported in this document.

The activity inside the spheres was quantified with three different SUV-based indicators, also used in the rest of this thesis:

- The mean activity inside the spheres was calculated and named SUV_{mean}
- The value of the maximum intensity voxel was named SUV_{max}

- SUV_{peak} was introduced as a potentially more robust indicator based on SUV_{max}. It considers an average value of neighbouring voxels in order to reduce the influence of noise on the maximum SUV measured, as suggested by P. Julyan *et al.* [133]. SUV_{peak} was calculated as the mean intensity value in a region of 16 voxels, comprising the maximum SUV in the insert and its 15 nearest neighbours. Given the voxel size of 2.73 mm x 2.73 mm x 3.27 mm, this approximated to a 1 cm^3 sphere around the hottest voxel as suggested in the PERCIST method used for assessment of tumour response to therapy [134].

The RC was determined in each sphere and compared between thick- and thin-wall inserts to quantify the difference between the activity observed on the scan and the actual activity in the spheres. It was calculated from the equation used by E. Prieto *et al.* [128].

$$RC = \frac{Mean\ activity\ measured\ (\frac{Bq}{mL})}{Activity\ injected\ (\frac{Bq}{mL})} \qquad \text{Eq. 22}$$

The effect E of the cold walls thickness on each one of these indicators was calculated as the relative difference between indicator values obtained for thin-wall inserts ($V(thin)$) and thick-wall ($V(thick)$) inserts. As an example, the effect of the cold walls thickness on the maximum SUV value measured in the sphere was determined as:

$$E(SUV_{max}) = \frac{(SUV_{max}(thick) - SUV_{max}(thin))}{SUV_{max}(thick)} \qquad \text{Eq. 23}$$

with E negative when the indicator value obtained is higher for thin wall inserts. When considering the global magnitude of the cold walls thickness on a group of test images, the absolute value of E was averaged to obtain the metric $\overline{\text{E}}$. The variation of the effect of cold walls with different image parameters was investigated by plotting E for each indicator against the TBR used for each sphere.

The software SPSS was used to detect any statistically significant differences in median between all E values obtained for thin- and thick-wall spheres. The Wilcoxon signed-rank test was applied to values obtained for all three SUV indicators, with a significance level set to *p=0.05*.

Segmentation results were first examined visually, in order to identify the test images on which the algorithms failed to segment the sphere. This inspection pointed out images with SUV lower than 2.5 or small SUV ranges, for which $SUV_{2.5}$, FT_{42}, and all other methods using thresholds could not successfully delineate the spheres. The contours obtained in those cases were excluded from the study for both thin- and thick-wall inserts to avoid any bias in the results. This amounted to 10% of the test images discarded, corresponding to the lowest TBRs and smallest spheres.

The delineation accuracy of all segmentations performed was assessed, by comparing the contour obtained to CT-derived ground truth. Ground truth was determined for each insert by drawing a sphere of the same diameter as the insert and positioning it on the associated high resolution CT scan to match the image of the sphere walls. The metrics described in II. A. 1. e) were used to directly compare the performance of the algorithms

The overall relative difference $\overline{\mathrm{E}}$ in accuracy metrics between values obtained for thin- and thick-wall inserts was calculated for each sphere across all TBRs and PET-AS methods used, in order to assess the general effect of the cold walls on the performance of PET-AS algorithms. Indicators S and PPV were also averaged for each sphere on all TBRs and PET-AS methods, to compare values obtained for thin- and thick-wall inserts on the whole set of images.

The Wilcoxon signed-rank test was applied to RVE and DSC values obtained for each of the PET-AS methods, to detect any significant difference in the median of the values obtained for thin- and thick-wall spheres. The significance level was again set to *p=0.05*.

The results obtained for thin- and thick-wall spheres were also compared for the different methods. For this purpose, the difference between metric values obtained for thin- and thick-wall inserts on the whole set of images used was derived for each method. The difference calculated was relative to values for thick-wall inserts, and therefore RVE was positive and DSC, S and PPV negative when the algorithm performed better on thin-wall inserts. Differences smaller than 5% were considered negligible in the interpretation of the data.

II. B. 1. c. Results

The plastic wall thicknesses of the different inserts, measured with the digital calliper, are displayed in Table 7 together with the inner volumes of the spheres, as well as the radii derived from these volumes. Measurements showed a reduction in plastic wall thicknesses for the custom spheres ranging from 83% to 93% compared to commercial inserts. This corresponded to a range of 1.12-1.49 mm difference between the wall thicknesses of both types of inserts. The mean (SD) measured wall thickness was 0.18 mm (0.06 mm) for the custom inserts, and 1.53 mm (0.14 mm) for the Liqui-Phil inserts. The spheres were renamed according to their derived diameter (cf. column 1 in Table 7).

The TBR values achieved in the experiments were 1.4, 2.1, 2.7, 2.8, 4.8 and 6.4. These values correspond to the ratio between spheres and background activities filled in the phantom, and were used as parameter S when deriving the PET-intensity profiles from the theory (with B=1).

	Measured wall thickness (mm)		Measured volume (mL)		Measured diameter (mm)		Relative difference in diameter (%)
Name	Liqui-Phil	Vacuum-moulded	Liqui-Phil	Vacuum-moulded	Liqui-Phil	Vacuum-moulded	
S10	1.35 (0.05)	0.24 (0.07)	0.53	0.48	10.06	9.74	3.2
S15	1.55 (0.09)	0.15 (0.06)	1.70	1.74	14.81	14.93	-0.85
S20	1.72 (0.24)	0.24 (0.07)	4.16	4.13	19.90	19.95	0.26
S30	1.43 (0.18)	0.096 (0.052)	14.47	14.02	30.24	29.92	1.1
S38	1.48 (0.17)	0.14 (0.05)	29.99	28.20	38.55	37.77	2
S58	1.65 (0.11)	0.23 (0.18)	102.20	97.38	58.02	57.09	1.6

Table 7. Measured inside wall thickness (SD of measurement), inner volume and inner diameter of Liqui-Phil and custom spheres. The last column gives the difference in diameter relative to the values for Liqui-Phil spheres.

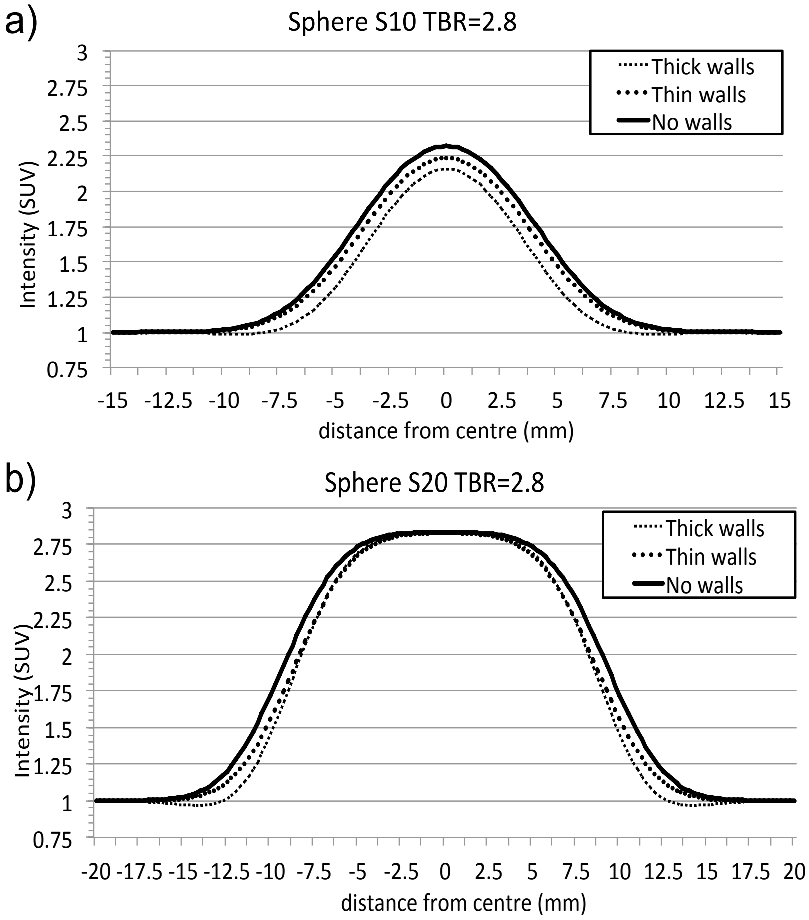


Figure 26. Theoretical radial profiles obtained for the different insert types at TBR = 2.8 for (a) S10 and (b) S20. The true profile, corresponding to the case of no walls, is shown by the continuous line.

Figure 26 a) and b) provide examples of radial profiles generated from Eq. 20, for spheres S10 and S20 respectively, at TBR=2.8. The effect of the PVE is visible on these profiles, where only a fraction of the voxels (none for S10) in the centre of the sphere reached the nominal SUV of 2.8. A difference can also be noted between the profiles generated for all three different types of inserts. The presence of the cold walls causes the intensities inside the sphere to be lower than for the wall-less case. This effect, however, appears smaller for the thin-wall inserts, for which the profile generated was closer to the expected wall-less profile on Figure 26. This observation justifies the investigation of thin-wall inserts, as it predicts a reduced cold wall effect.

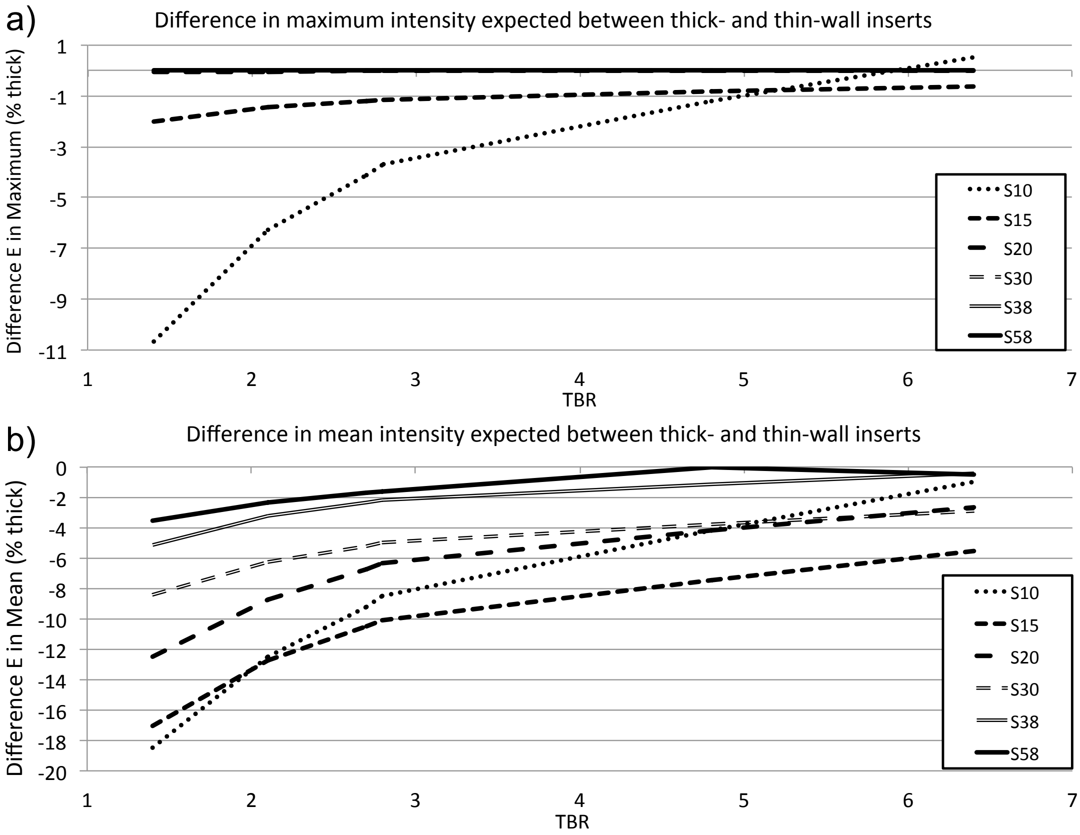


Figure 27. Expected differences between thick- and thin-wall inserts in (a) maximum and (b) mean intensity values. The differences shown are relative to the values for thick-wall inserts

The difference between thin- and thick-wall inserts (relative to the values for thick-wall inserts) in terms of mean and maximum intensity values were extracted from the theoretical radial profiles, and are plotted on Figure 27 against the TBR value for each sphere size. The results indicated a systematic increase (except for S10 at TBR=6.4) in mean and maximum SUV values when using thinner inserts. The smaller maximum SUV value obtained for the thin-wall S10 at TBR=6.4 was due to a slightly smaller diameter than the corresponding thick-wall insert. This effect appeared most important on small spheres and at low TBRs, and decreased with both parameters. The expected magnitude of this effect was larger for mean (up to 18% increase in mean SUV) than for maximum values (up to 11% increase in max SUV). The results predicted a non-negligible effect (>5%) of the difference in cold wall thickness only for the smallest sphere, at TBR lower than 3, for the maximum value. In the case of the mean intensity value, the difference was non-negligible (>5%) for the two smallest spheres at all TBR, and for all spheres except the largest at TBR=1.4. It is therefore reasonable to expect that the differences observed on the experimental results would be larger for the mean intensity values, and will decrease with increasing sphere size and TBR. The same trend should be observed for RC values, which rely on the mean SUV.

The displacement of the maximum intensity gradient point was minor (<0.1 mm) for thin-wall inserts, and reached 1.1 mm for thick-wall inserts.

Figure 28 illustrates the PET-TOF intensity profiles drawn on the images for thin- and thick-wall spheres S58 and S15. The profiles on Figure 28 show some visible difference in the apparent sphere diameters AD between the two sphere sets, especially for smaller spheres, as expected from the theoretical profiles. The SD of the measured AD values did not exceed 4 % of the mean value for any of the spheres. The calculated difference (SD) in measured AD between the two sphere sets showed a mean value of 2.1 mm (2.0 mm) across all the scans. These discrepancies are on average twice as large as the maximum differences in measured wall thicknesses. It can also be noted that SUV_{max} for S15 is 25% higher for the thin-wall insert, as shown in Figure 28 c).

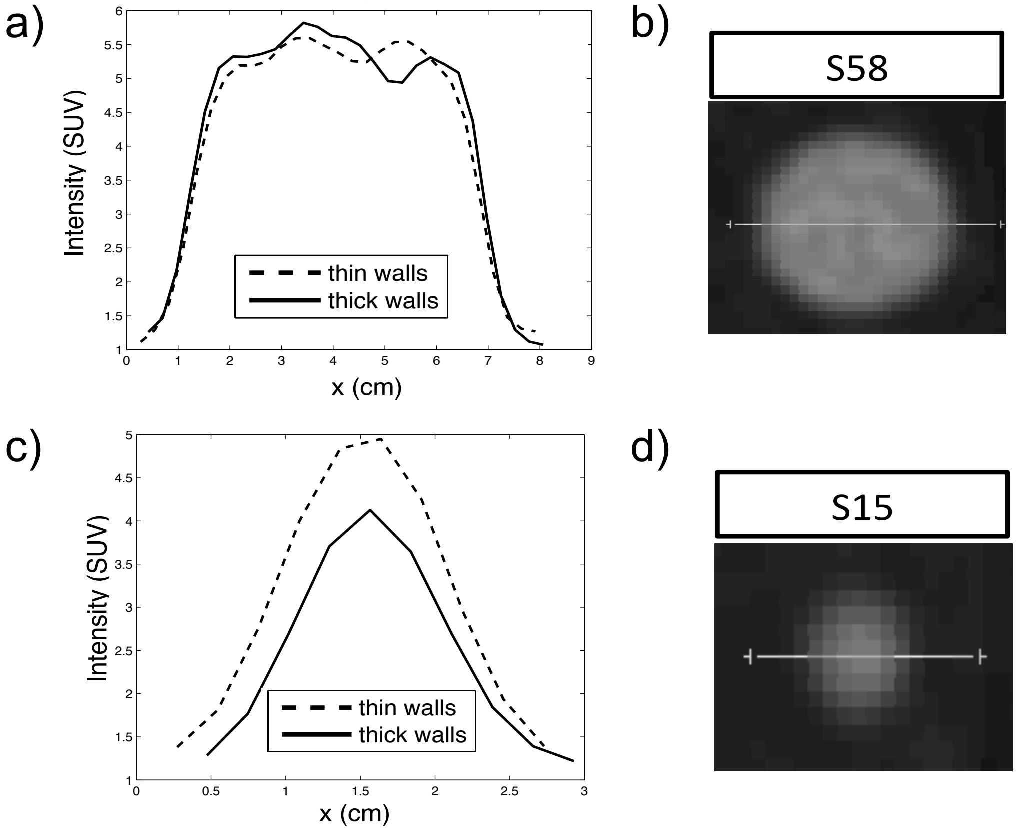


Figure 28. Thin- and thick-wall PET-TOF intensity profiles at TBR = 4.8 for (a) S58, (c) S15. (b) and (d) show the line drawn across the middle slice of S58 and S15, respectively

The variation of the three SUV indicators with TBR for each sphere size is presented in Figure 29. Results show higher SUV_{mean} values for the thin-wall inserts (up to 25%) except in one case. SUV_{max} and SUV_{peak} both show substantial differences between thin- and thick-wall inserts for the smallest sphere at TBR=1.4, thin-wall inserts yielding values up to 26% higher for SUV_{max}, 51% for SUV_{peak} (cf. Figure 29 b) and c)). This observation is consistent with the results of the theoretical data shown on Figure 27, which predicted a non-negligible effect (E>5%) of the difference in wall thicknesses only for the smallest sphere at the lowest TBRs. The metric SUV_{peak}, which was used as a more robust alternative to $SUV_{max,}$ as suggested by Wahl *et al.* [134], confirmed the result observed on the SUV_{max} data, of a large difference between inserts for the smallest sphere at TBR=1.4 only. Figure 29 also highlights the increase in the effect of cold walls with decreasing TBR for all indicators, and shows the substantial (E>5%) effect of the cold walls on all indicators for TBRs smaller than 4.

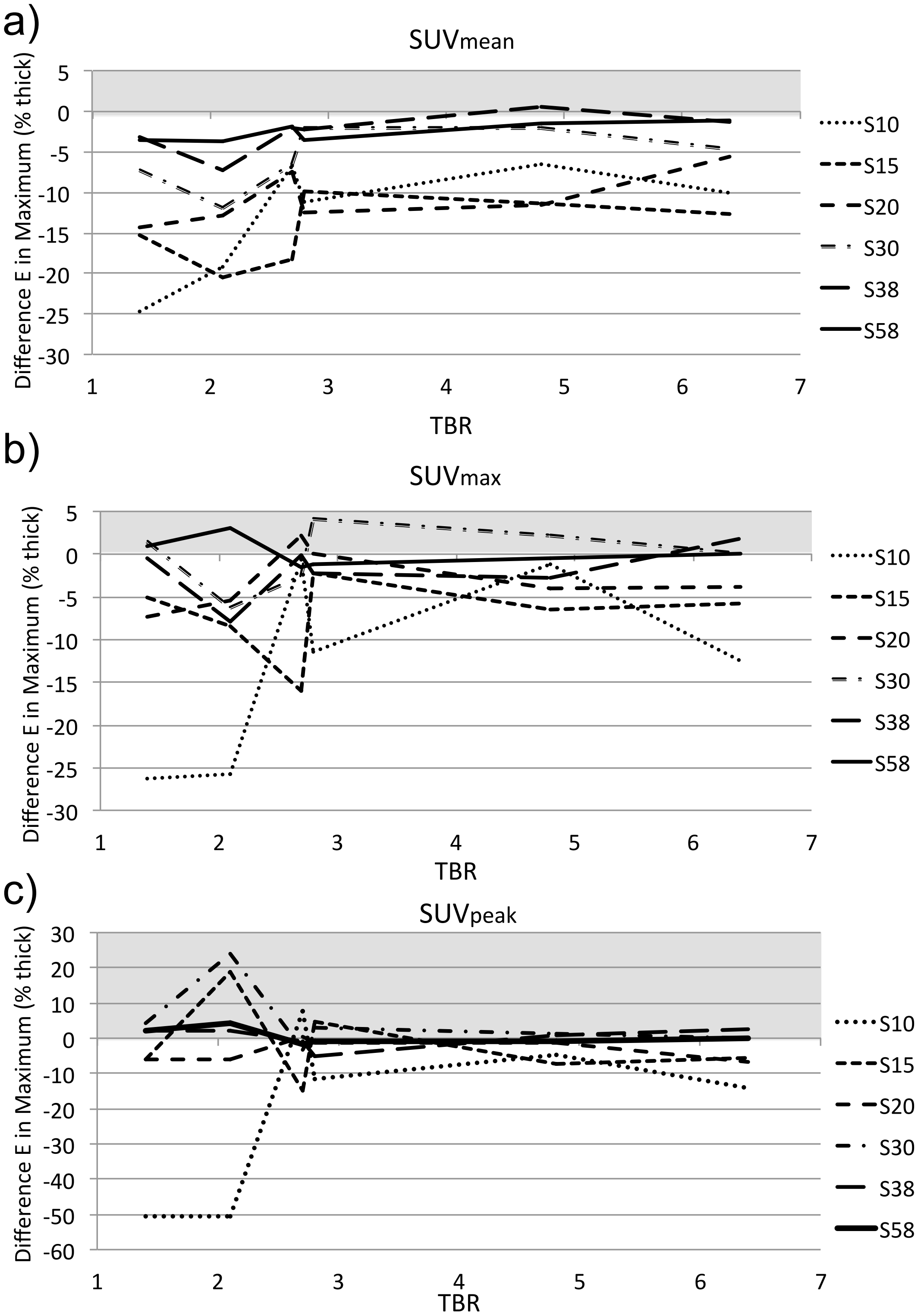


Figure 29. Influence of the cold wall thickness (relative difference E) on the different SUV indicators at different TBRs: (a) SUVmean, (b) SUVmax, and (c) SUVpeak. The grey area highlights positive values of E [i.e., SUV (thick) > SUV (thin), cf. Eq. (23)].

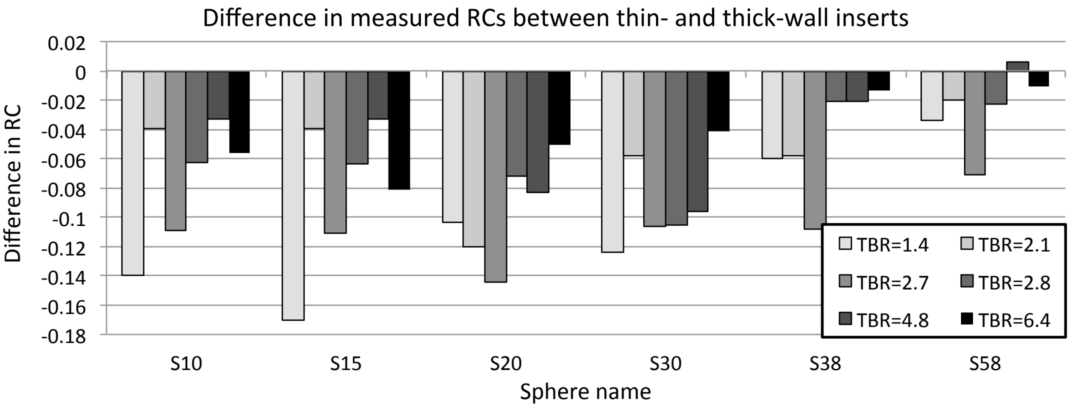


Figure 30. Difference between RCs obtained for thin- and thick-wall inserts for each sphere size and TBR.

Figure 30 depicts the difference in RCs observed between both sphere sets, for each sphere size. The RC appears systematically higher for thin-wall plastic inserts, which is in agreement with the results of the simulated data. Relative differences between thin- and thick-wall inserts reach values of 0.17 (RC of 0.86 and 0.69 respectively), i.e. 17% of the nominal activity. The smallest RC value obtained, corresponding to the smallest sphere was 0.50 for thick-wall inserts and 0.54 for thin-wall inserts.

Table 8 summarizes the results of the Wilcoxon signed-rank test for each indicator used in this study, displaying the difference in median values obtained between thin- and thick-wall inserts and the corresponding p-value returned by SPSS. The results show a statistically significant difference (*$p<0.05$*) in the median SUV_{mean} value of both insert sets of 0.19. The difference in medians for SUV_{max} reaches the smaller value of 0.14, which is also statistically significant (*$p<0.05$*). RC values yielded a statistically significant difference in median values (0.028) between both sphere types. The difference in medians, relatively high compared to the range of RCs observed (0.17-0.68), shows the high influence of the cold wall thickness on the recovery of injected tracer concentration.

Indicator	SUV_{mean}	SUV_{max}	SUV_{peak}	RC
Difference in medians (thin-thick)	0.19	0.14	-0.029	0.028
p-value	<0.01 *	0.003*	0.139	<0.01 *

*Statistically significant (*p<0.05*)

Table 8 . Results of the Wilcoxon signed-rank test for the different SUV indicators.

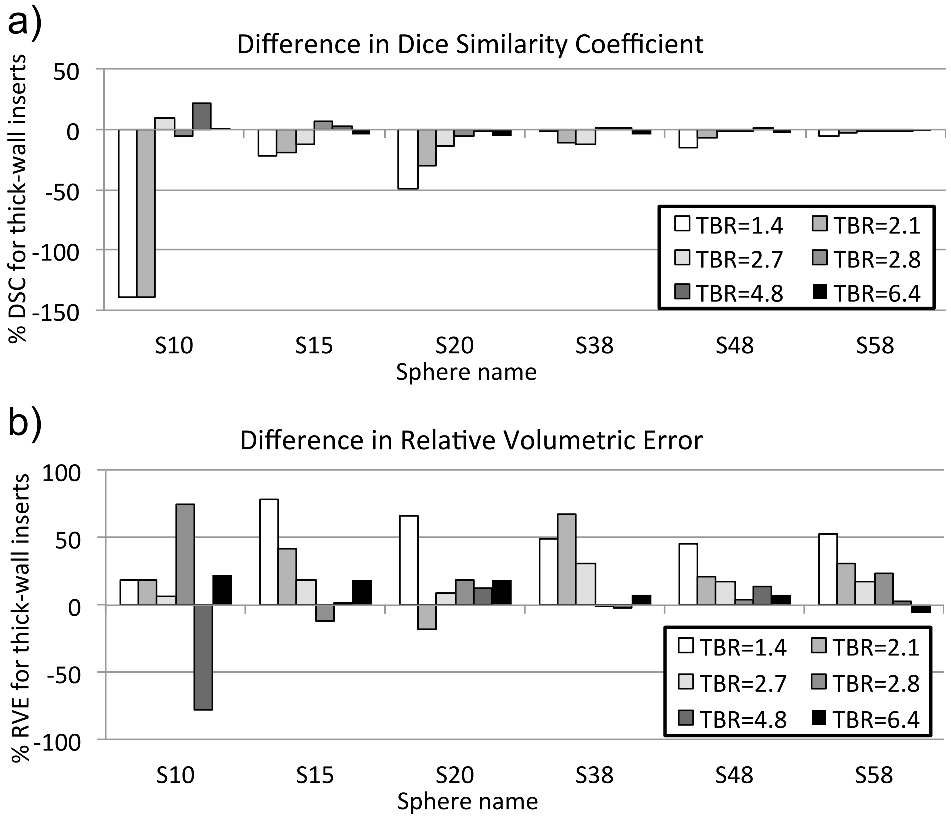


Figure 31. (a) DSC and (b) RVE values obtained averaged on all PET-AS methods for each sphere size and TBR.

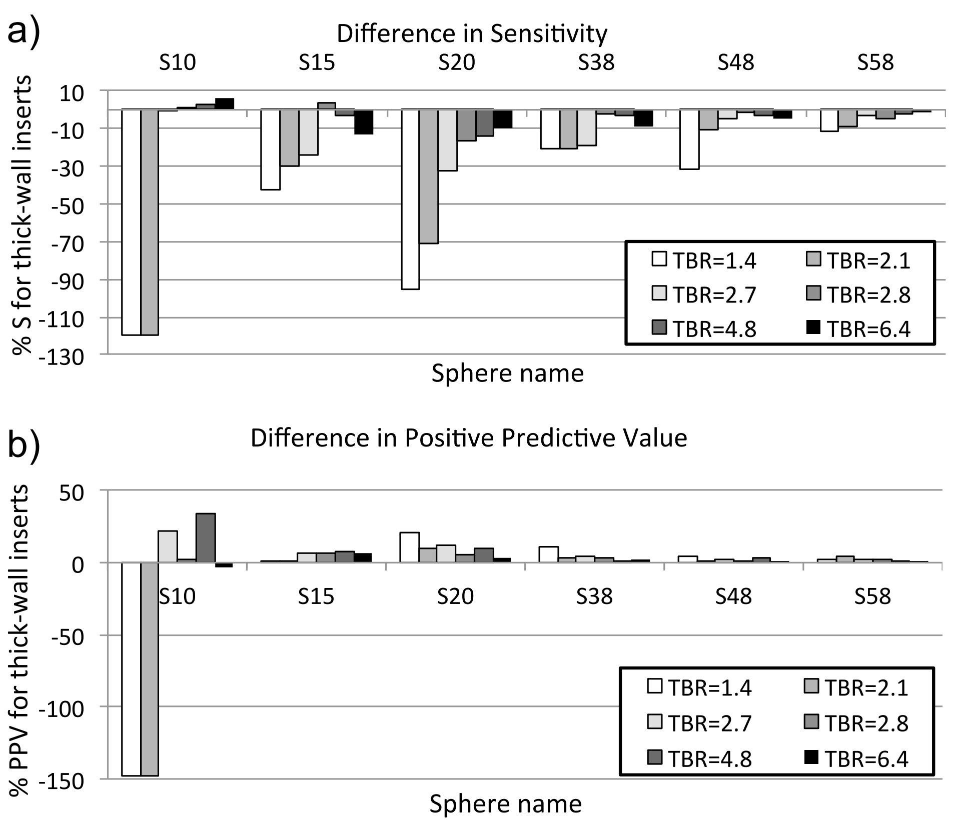


Figure 32. (a) S and (b) PPV values obtained averaged on all PET-AS methods for each sphere size and TBR.

RVE and DSC values obtained, averaged on all PET-AS methods are compared on Figure 31 between both sets of spheres for each sphere size and TBR. The same comparison is presented for S and PPV on Figure 32. The data show that segmentations performed on thin-wall inserts yielded lower volumetric errors (up to 31%) and higher similarity with the true contours (up to 14%) on average on all methods used. Results of the Wilcoxon signed-rank test comparing values for RVE and DSC are given in Table 9. For each PET-AS method, the difference in median values obtained between thin- and thick-wall inserts and associated p-value returned by SPSS are reported. The data show statistically significant differences in delineation accuracy (RVE and DSC) between thin- and thick-wall spheres for all methods except GC, KM and AC.

PET-AS Method	FT$_{42}$	SUV$_{2.5}$	AT	GC	RG	KM	FCM	GCM	WT	AC
Difference in medians	-0.14	-0.15	-0.090	-0.079	-0.061	-0.10	-0.064	-0.56	-0.072	-0.14
p-value (RVE)	0.003*	0.00*	0.00*	0.267	0.00*	0.001*	0.002*	0.00*	0.00*	0.005*
Difference in medians	-0.042	0.019	0.031	0.012	0.012	0.0085	0.051	0.041	0.0043	-0.027
p-value (DSC)	0.002*	0.004*	0.007*	0.372	0.008*	0.110	0.00*	0.001*	0.022*	0.481

*Statistically significant (*p<0.05*)

Table 9. Results of the Wilcoxon signed-rank test applied to RVE and DSC values obtained with each PET-AS method on both sets of spheres. The difference in medians is relative to thin-wall inserts.

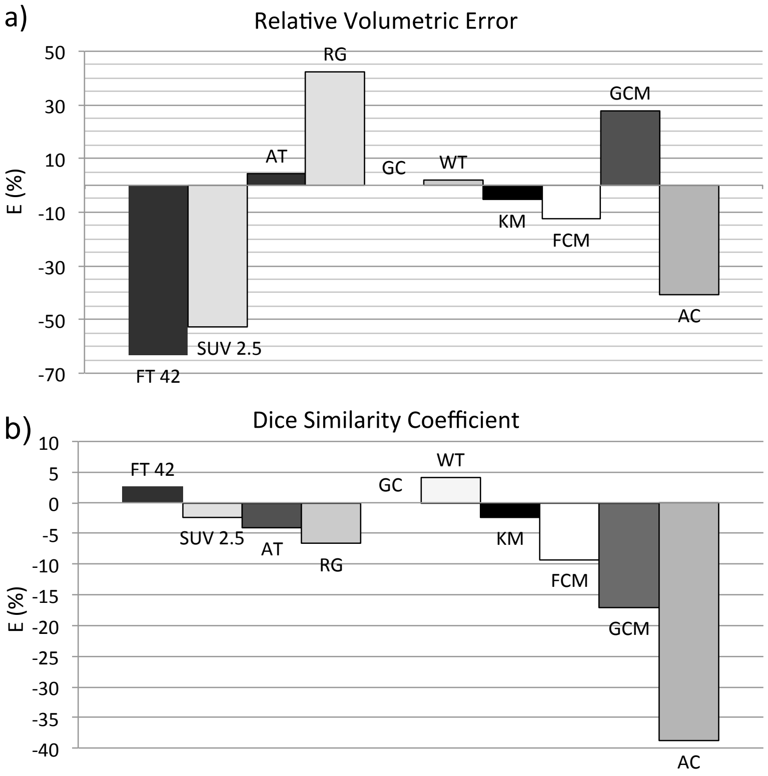


Figure 33. (a) Difference in RVE and (b) difference in DSC between delineations on thin- and thick-wall inserts for the different algorithms.

Figure 33 compares the effect of the cold walls on the different algorithms in terms of relative difference E in RVE and DSC values, averaged on all TBR and spheres. Important differences between algorithms of different types can be observed. WT and AT were the least affected by the wall thickness, with negligible differences for both RVE and DSC. AC appears to be the most overall affected method, with 40% larger RVEs but 39% higher DSC in average for thin-wall inserts.

II. B. 1. d. Discussion and conclusion

In this work, the effect of cold walls thickness was quantitated on a large range of TBRs and sphere sizes, and it was shown that plastic inserts with walls of 0.1-0.2 mm provide improved SUV quantification in PET images. The results of this study demonstrated a statistically significant effect of the cold walls thickness on image activity recovery coefficients as well as on the delineation accuracy of PET-AS algorithms. The effect on the delineation performance was dependent on the algorithm used. The use of wall-less activity regions in a hot background is therefore recommended for the validation and inter-comparison of PET-AS algorithms in clinically relevant conditions.

This study provides a thorough investigation of the effect of the cold plastic walls thickness in spherical phantom inserts, supported by both theoretical and experimental results with thin- and thick plastic wall inserts. For this purpose, six spherical plastic inserts (cf. Figure 25) were successfully manufactured with a reduction of the wall thickness reaching 90%, using the vacuum-moulding technique, and scanned. The equations used to derive 2D-intensity profiles have already been used and validated experimentally by Hofheinz *et al* [84]. However, their study focused on the calibration of the optimal volume recovery threshold, and investigated the case of thick plastic walls at TBRs higher than 3. The results of this study, which extended the TBR and sphere size range to low values, showed a good agreement with the theory (cf. Figure 26 and Figure 28, Figure 27 and Figure 29), although some fluctuations could be observed in the experimental data. These fluctuations could be due to the presence of noise, which was not taken into account in the theoretical model. In this study, higher activity levels were used, in order to reproduce typical clinical background intensity values (50-70 kBq/mL in the phantom background compared to 0-46 kBq/mL for Hofheinz *et al.*). This would be the main source of noise in the present data. On the other hand, the longer emission times used by Hofheinz et al probably compensated the expected lower sensitivity of their PET scanner. The use of a transformation to radial coordinates, as applied by Hofheinz *et al.*, could have helped minimize the impact of image noise in this study. However, the cold walls effect was found to be statistically significant in the data presented despite the presence of observable fluctuations from the theoretical values. In addition, the aim of this study was to assess the magnitude of the cold walls effect directly on quantification and segmentation of typical PET intensity data with a large range of TBRs. This implied using standard routine scanning protocols and settings leading to typically observed levels of noise in the data, as well as commonly available image processing tools, which operate directly on the 3D matrix of image intensity voxels, and might not include such a transformation. The application of a 3D voxel-wise de-noising filter like the one used by Geets *et al.* [98] would be useful in a subsequent study detailing the role of noise on the cold walls effect.

The data showed that higher $\mathrm{SUV_{max}}$ values are measured in thin-wall inserts for the smallest spheres only, while $\mathrm{SUV_{mean}}$ is systematically lower in thick-wall inserts. This can be explained by the fact that the mean SUV includes voxels located close to the sphere walls. $\mathrm{SUV_{peak}}$ did not show an overall statistically significant effect of the cold wall thickness (cf. Figure 27 and Figure 29), suggesting that the significant effect observed for $\mathrm{SUV_{max}}$ should be considered with caution, as it could be affected by noise. The investigation of the cold walls effect at different TBRs and sphere size, which had been previously investigated on a smaller range of values [84], [85], showed that it decreases with increasing TBRs and sphere sizes. The theoretical framework developed by Hofheinz *et al.* [84] predicted that this effect vanishes when the TBRs reaches infinity, i.e. in the case of a cold background. The model also suggests that PET-image quantification should be carried out, when possible, in phantoms with inactive background, especially when performing algorithms calibrations on a large range of sphere sizes and TBRs. However, this situation is not encountered in clinical conditions, therefore this method would not be useful in studies simulating metabolic activity distributions, for example when testing segmentation algorithms destined to be used in radiotherapy treatment planning. This work focused on the experimental validation and inter-comparison of advanced automatic algorithms for a clinically useful range of TBRs and sphere sizes.

Large differences can be observed in the RC between results obtained for thick- and thin wall inserts, as expected from the theory. The use of thin-wall inserts allowed a systematic reduction of the quantification errors in terms of activity recovery in the image (cf. Figure 30). This was confirmed by statistical analysis (higher median value for thin-wall spheres with 0.028 difference, *p<0.05*). Bazanez-Borgert *et al.* [130] quantified the gain in activity recovery obtained when using wall-less inserts compared to fillable plastic ones, in a study validating the wax inserts suggested by Turkington *et al.* [135]. They also found a tendency for high recovery coefficients for the wax spheres. These results are particularly important in quantitative studies, and point out the risk of underestimating the SUV measured within a plastic insert. In this work, thin-wall inserts did not systematically reduce the error in activity recovery to negligible levels (E<5%). This suggests that the effect of plastic walls on image intensity values cannot be completely eliminated by using minimal plastic wall thicknesses. The effect of a 90% reduction in cold wall thickness on the image quantification, as well as on the evaluation and comparison of fully automatic segmentation algorithms was quantified.

In this manuscript, the impact of cold walls on the delineation accuracy was quantified for several fully automatic segmentation algorithms (cf. Table 9 and Figure 31). In particular, Figure 32 shows higher sensitivity values obtained for thin-wall spheres of all sizes, but lower positive predictive values for thin-wall inserts for all spheres except the smallest, due to the overall decreased activity in the spheres. However, differences in S were much larger than differences in PPV (up to 33% for S, 10% for PPV). This indicates that the reduction in plastic wall thickness allows the algorithms to detect all voxels within the lesion, at the cost of including a few background voxels. This may be the case in particular for high intensity lesions, which can appear larger than their actual because of the PVE causing the high intensity region to spill out into the background. Figure 31 b) and Figure 32 show that the values obtained for two very close TBR values of 2.7 and 2.8 reached highly different metric values, in particular for the smallest spheres. The values of 2.7 and 2.8 are within 4% of each other, which corresponds to the uncertainty of the TBR value caused by the calibrator precision error (cf. II. A. 2. a. ii). As a consequence, the difference between the values obtained at these two TBRs cannot be interpreted as being due to the TBR. The discrepancies observed on Figure 31 b) and Figure 32 are more likely to be due to the high sensitivity of some methods to the image noise, leading to large differences in values in particular for RVE, which is calculated relatively to the volume of the spheres. This calls for caution when interpreting individual data points, and suggests that the interpretation of the results should rather focus on the trend shown across TBRs.

Finally, methods based on different mathematical algorithms or initialisations appeared to be influenced differently by the presence of cold walls as shown in Figure 33. For example AT, which is based on the maximum SUV, was less affected by the cold plastic walls than algorithms based on regional mean SUVs, such as the clustering algorithms and RG which perform significantly better on the thin-walls set. The high differences in RVE observed for fixed thresholding methods were due to outliers at the smallest TBRs and sphere sizes. On the other hand gradient-based methods GC and WT showed little variation between thin- and thick-wall spheres. This is explained by the fact that both algorithms rely on the position of highest intensity gradient in the image. The theoretical analysis showed that, for the data presented here, the maximum displacement of the highest intensity gradient only reached a value of 1.1 mm, which is smaller than half the smallest voxel dimension (2.734 mm). This can explain why these methods are less influenced by the wall thickness than SUV_{mean} –based methods. This does not apply to AC, for which the initialisation step also uses a version of the GMM method. These data show the importance of considering the thickness of cold walls when evaluating and comparing the performance of segmentation algorithms using fillable inserts as reference.

Cold background can be used for the calibration of segmentation algorithms as a way of avoiding errors due to the cold walls effect. In this work, however, the algorithms compared did not require any tuning. Therefore, thin wall inserts in a hot background instead, to create more clinically realistic conditions, while accounting for background scatter and heterogeneities. The results of this study are in line with the decision of some groups to use wall-less inserts in their quantitative image analysis studies, as this method did not completely eliminate the cold walls effect. Montgomery *et al.* [90], for example, acknowledged the possible effect of using inactive walls, and used the method described by Turkington *et al.* [135] to generate wax inserts for the validation of their segmentation algorithm. However, such techniques present technical difficulties, in particular for achieving homogeneous activity distribution inside the wax, as pointed out by Bazanez-Borgert et al [130]. Further work was carried out to validate a technique for the production of printed subresolution sandwich phantoms to be used as reference data for PET segmentation testing. This is described in details in section III. A. 1 of this thesis.

II. B. 2. Evaluation of the segmentation of non-spherical volumes

II. B. 2. a. Purpose

The validation of PET-AS methods is often done with test data generated from 3D fillable plastic phantoms, such as the Spherical Lucite Phantom (Canberra-Packard, Zellik, Belgium), the NEMA IEC body, and the IEC quality phantom [77], [91], [98]. The main drawback of this type of method is that images generated do not represent the reality of tumours in the clinical setting, particularly because most the inserts used have a spherical shape and thick plastic walls. Although non-spherical regular [80], [87] and irregular [87], [136] inserts have been used elsewhere, their geometrical characteristics were not systematically used to assess the performance of the PET-AS methods. This series of experiments relied on the hypothesis that carefully designed inserts of different geometry, beyond fillable spheres provide relevant and complimentary information about the performance of PET-AS methods. The aim of this work was to generate a range of test inserts reflecting clinical situations, to compare the performance of the implemented PET-AS approaches, and identify their weaknesses and potential areas of improvement. This work was peer reviewed and published in [137].

II. B. 2. b. Methods

II. B. 2. b. i. Description and quality assessment of test inserts

A total of 16 non-spherical fillable plastic inserts, listed in Table 10, were designed with input from a consultant radiologist to represent geometrical characteristics of interest for clinical tumours, such as high Aspect Ratio (AR: ratio of length to maximum diameter) often encountered in H&N, ellipsoidal shape typical of malignant lymph nodes, and necrotic centres often observed for lung tumours. All inserts were generated in house with a 0.18 mm wall thickness using the vacuum-moulding technique previously described (cf II. B. 1). Ellipsoidal and toroidal (doughnut-shaped) inserts were derived with the same volumes as spheres S15, S20, S28 and S38 of the Raydose phantom (1.8, 4.2, 11.5 and 28.7 mL). The remaining inserts were designed to investigate of the impact of a particular parameter on the delineation accuracy (e.g. sharp corner, high aspect ratio etc.). For this purpose, tubes with rounded ends, drop- and pear-shaped inserts, were manufactured with a constant volume of 28 mL (corresponding to sphere S38), representing typical stage T3 tumours encountered in the H&N at Velindre Cancer Centre. The inserts were derived with three different Aspect Ratios (AR: ratio between the largest diameter and smallest perpendicular diameter) of 2, 2.5 and 3.

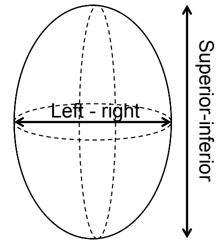
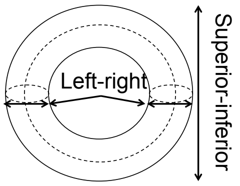
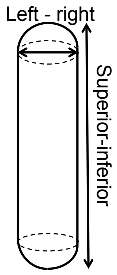
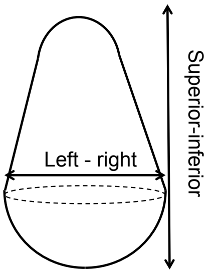
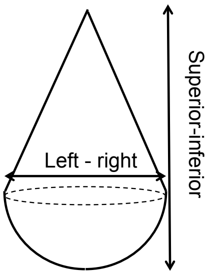
Geometry	Names	Purpose
Ellipsoid  	E15, E20, E28, E38	- Aspect Ratio higher than 1 - modelling structures such as malignant lymph nodes
Torus  	To20, To28, To38	- small cross-section - modelling of necrotic region
Tube  	Tu38a, Tu38b, Tu38c (AR, of 2, 2.5, 3 respectively)	- Aspect Ratio higher than 1 - modelling "long" tumour
Pear  	P38a, P38b, P38c (AR, of 2, 2.5, 3 respectively)	- Aspect Ratio higher than 1 - modelling "long" tumour with asymmetric extension
Drops  	D38a, D38b, D38c (AR, of 2, 2.5, 3 respectively)	- Aspect Ratio higher than 1 - modelling "long" tumour with asymmetric extension

Table 10. Description of the vacuum-moulded plastic inserts used in this study

The actual volume of the non-spherical inserts was measured by subtracting from the weight of the insert filled with water the weight of the empty insert. The

XP205 DeltaRange Analytical balance (Mettler Toledo LLC, Columbus, USA) was used after previous calibration.

II. B. 2. b. ii. Comparison of PET-AS methods

All images were obtained with the settings previously described (cf. II. A. 1. a). For the sake of comparison, spherical and non-spherical inserts of matched volumes were scanned simultaneously in the custom-made phantom described previously (cf. II. B. 1). All scans were acquired with a TBR of 5, corresponding to an intermediate value in the range of TBRs observed at Velindre Cancer Centre.

All PET-AS (cf. II. A. 1. d) and basic segmentation methods were applied to all test images obtained. Non-usable contours were visually identified and rejected from the study as in II. A. 3. a. ii. The segmentation results were compared to the ground truth using the metric DSC as previously described (cf. II. A. 1. e).

In addition the recovered physical dimensions of non-spherical inserts were measured for each accepted contour (except ellipsoids and tori), by determining the maximum diameter of the contour obtained in the transverse direction, and the maximum length of the contour in the superior-inferior direction. The error in recovered dimensions corresponded to the difference between the dimensions of the contours generated and the true inserts dimensions, expressed as a percentage of the true dimensions.

II. B. 2. c. Results

II. B. 2. c. i. Generation of PET test inserts

Some of the manufactured inserts are shown on Figure 34. The measured volumes of the inserts were compared to the values targeted for the manufacturing process. The target volume was achieved (difference <5% target volume) in all cases except for the toroidal objects, which were proved most challenging to manufacture. The error for toroids ranged within 18% and 35% of the target volume for To20 and

To38 respectively. This was due to a difference in the dimensions of the insert, while the targeted toroidal geometry was maintained. As a consequence, the toroidal reference contours used in the study were adapted to the manufactured inserts' measured dimensions.

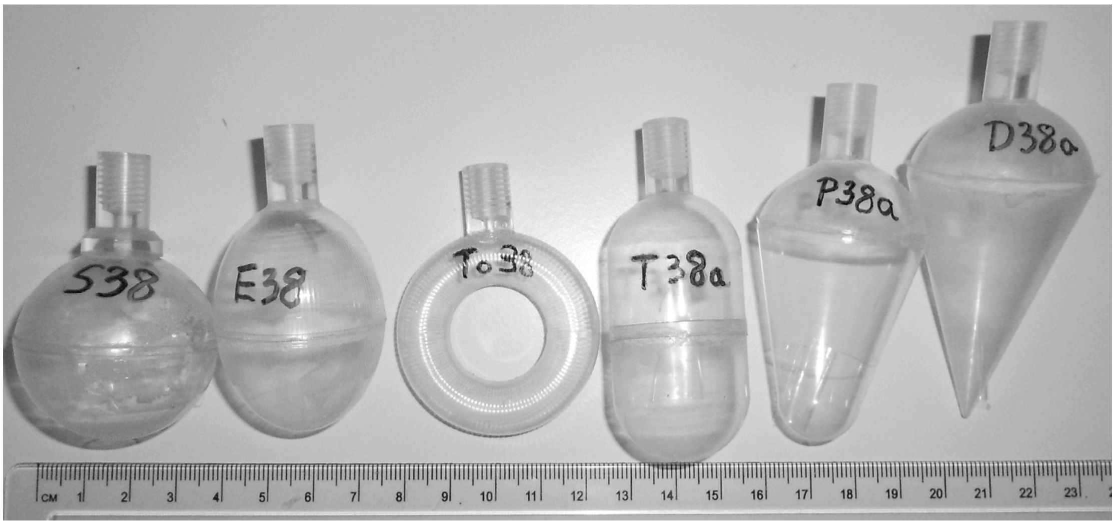

S38
E38
To38
T38a
P38a
D38a
CM 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Figure 34. Vacuum-moulded plastic inserts with volume corresponding to the 38 mm diameter sphere.

In a few cases the inserts produced were slightly different from the geometrical models. This was due to:

- the presence of a plastic rod used to position the inserts,
- for D38c: a slight bend in the vertical axis (less than 1 mm deviation for 60 mm height),
- a truncation by 1-2 mm of the tip of drop-shaped inserts due to the manufacturing process.

In those cases, the contours were manually edited on the high resolution CT to match the analytical model.

II. B. 2. c. ii. Evaluation of PET-AS methods

The DSC values obtained by the different PET-AS methods for each non-spherical insert are reported in Table 11.

	Dice Similarity Coefficient							
Insert	**AT**	**GC**	**RG**	**KM**	**FCM**	**GCM**	**WT**	**AC**
E15	0.83	0.73	0.76	0.74	0.79	0.79	0.76	0.78
E20	0.90	0.89	0.96	0.89	0.82	0.91	0.90	0.85
E28	0.93	0.92	0.94	0.95	0.80	0.91	0.91	0.89
E38	0.95	0.94	0.96	0.96	0.81	0.93	0.93	0.93
To20	0.52	0.31	0.45	0.39	0.38	0.33	0.34	0.41
To28	0.77	0.37	0.67	0.58	0.62	0.76	0.42	0.39
To38	0.83	0.49	0.78	0.74	0.83	0.80	0.42	0.38
Tu38a	0.95	0.94	0.95	0.95	0.76	0.92	0.91	0.92
Tu38b	0.94	0.94	0.95	0.94	0.75	0.91	0.92	0.89
Tu38c	0.93	0.92	0.94	0.95	0.73	0.89	0.92	0.88
P38a	0.93	0.90	0.93	0.92	0.78	0.91	0.85	0.90
P38b	0.91	0.90	0.92	0.92	0.77	0.90	0.86	0.92
P38c	0.95	0.91	0.87	0.93	0.76	0.92	0.86	0.92
D38a	0.93	0.89	0.93	0.94	0.78	0.91	0.88	0.92
D38b	0.93	0.85	0.91	0.93	0.77	0.91	0.87	0.91
D38c	0.94	0.91	0.92	0.94	0.77	0.92	0.87	0.93

Table 11. Accuracy (DSC) obtained by the different PET-AS methods on non-spherical inserts.

Differences between the volume delineated and the true volume were smaller than 2% of the true volume for all ellipsoids, and corresponded to an under-estimation of the volume for all methods. DSC values were higher for the spheres compared to the non-spherical inserts of matched volume in 88% of the cases, and this was true for the smaller insert for all methods. RG produced contours with a high level of conformity to the reference contour (DSC>0.94) for the three largest ellipsoidal inserts, and AT reached DSC values higher than 0.83 at all ellipsoid sizes. FCM did not reach DSC values higher than 0.82 for ellipsoidal inserts. The maximum differences between spherical and ellipsoidal DSC values were obtained at the smaller volume (except for FCM), and reached 0.17, 0.11, 0.15 and 0.14 DSC for RG, KM, FCM and GCM respectively.

In the case of toroids, three scans were acquired with the inserts aligned with an angle of 0^{o}, 45^{o} and 90^{o} of the sagittal plane, to check for any impact on the image quantification and segmentation. No trend linked to the insert position was observed, and the average DSCs across the three instances are therefore reported in Figure 35.

The DSC values obtained were smallest for To20 (DSC ranging from 0.38 to 0.83). WT and AC underestimated the volumes of To28 and To38 (volumes lower than 3 mL and 5 mL respectively), resulting in low DSCs. GC also underestimated the volume of the largest torus.

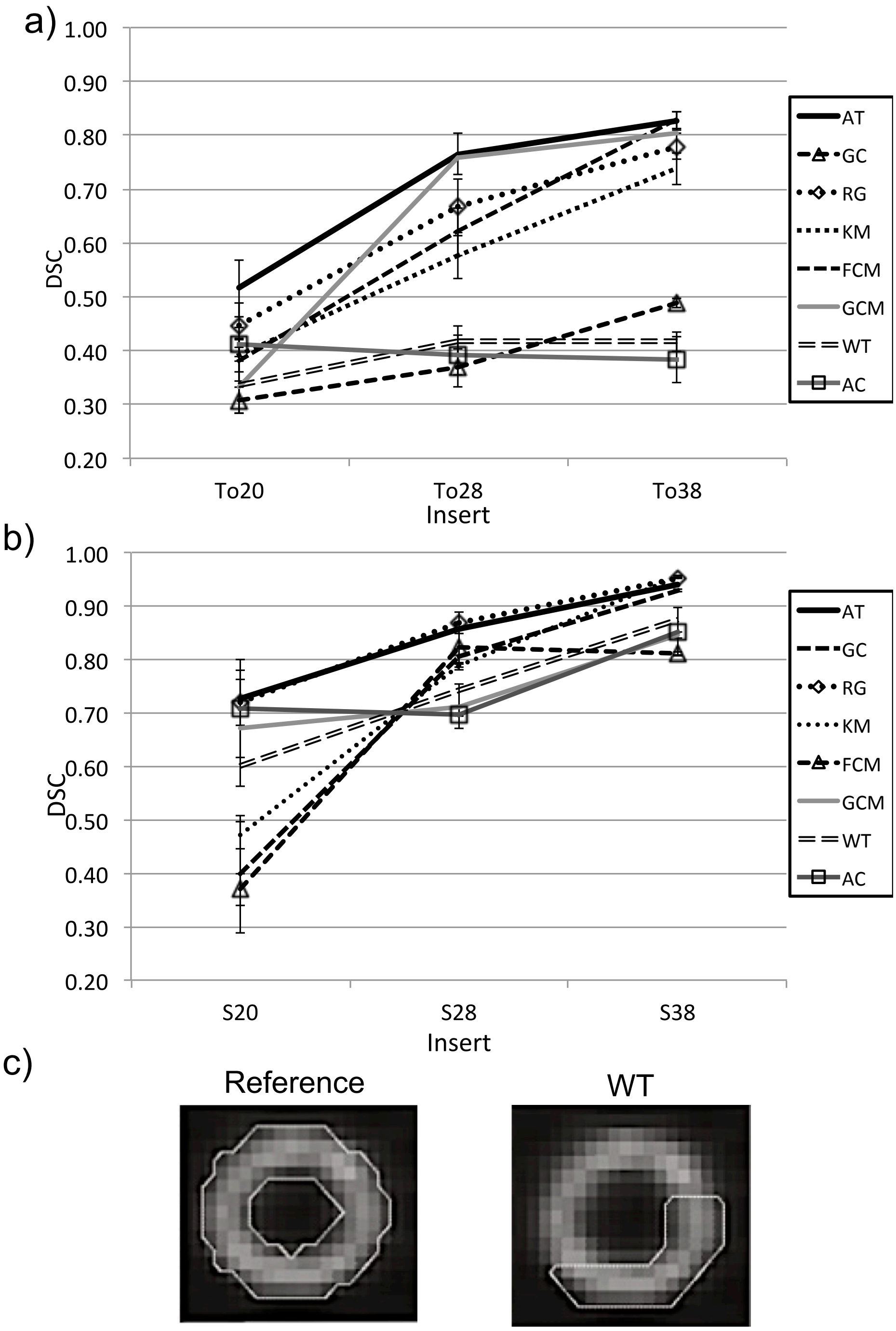


Figure 35. DSC obtained by the PET-AS methods on a) toroidal and b) spherical inserts of matched volume, and c) reference and WT contour obtained for To38 aligned with the coronal axis.

Figure 35 a) and b) show the DSC values obtained for each method on toroidal and spherical inserts respectively. The values were lower for toroids than for spheres for all methods. GC, WT and AC yielded much lower DSC values (up to 0.46 lower) for all toroidal inserts. This is depicted on Figure 35 c), showing that WT only recovered a small part of the To38 torus, which was also observed for GC and AC.

The data in Table 11 show that AT, RG and KM were the best PET-AS methods overall for recovering the volumes of tubes, pear- and drop-shaped inserts, with DSCs higher than 0.91 (except for P38c for RG). AT and KM systematically under- and over-estimated the volumes respectively, whereas RG did not show any systematic trend. High DSC values were also obtained for GCM and AC on all inserts (>0. 89 and >0.88 respectively). FCM reached the lowest DSCs (DSC<0.79) for these inserts.

Figure 36 shows reference and PET-AS contours generated in the coronal plane for D38c, and highlights the difficulties encountered by some PET-AS methods to recover the tip of the drop-shaped insert. Methods such as GC, AC and WT tended to generate a contour wider than the true insert, whereas methods such as RG or AT failed to include the tip of the insert in the delineation. The figure also shows large errors or the GC algorithm on the transverse plane near the insert tip. The measured values of error in the height recovered by the PET-AS methods are presented on Figure 37 show that the methods tended to under-estimate the object length, especially in the case of drop-shaped inserts (12-36% error). All methods achieved errors smaller than 22% of the true length, except for RG (>30%). AT, KM, GCM, AC and WT all achieved errors in the length smaller than 14% of the true volume for all geometries. RG yielded high positive errors in length for all geometries, which indicates a poor shape recovery of such thin inserts.

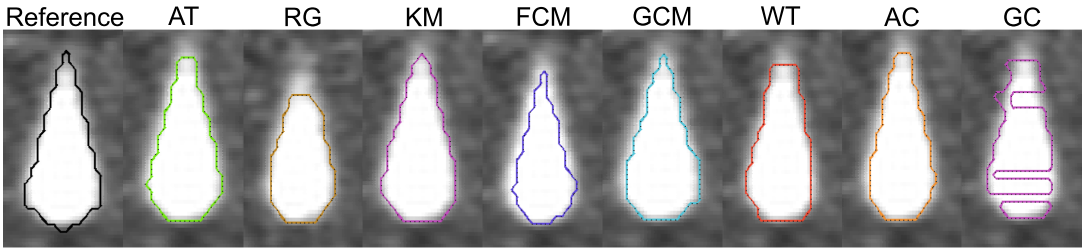


Figure 36. Reference and PET-AS contours generated for D38c. Contours are shown on the coronal plane.

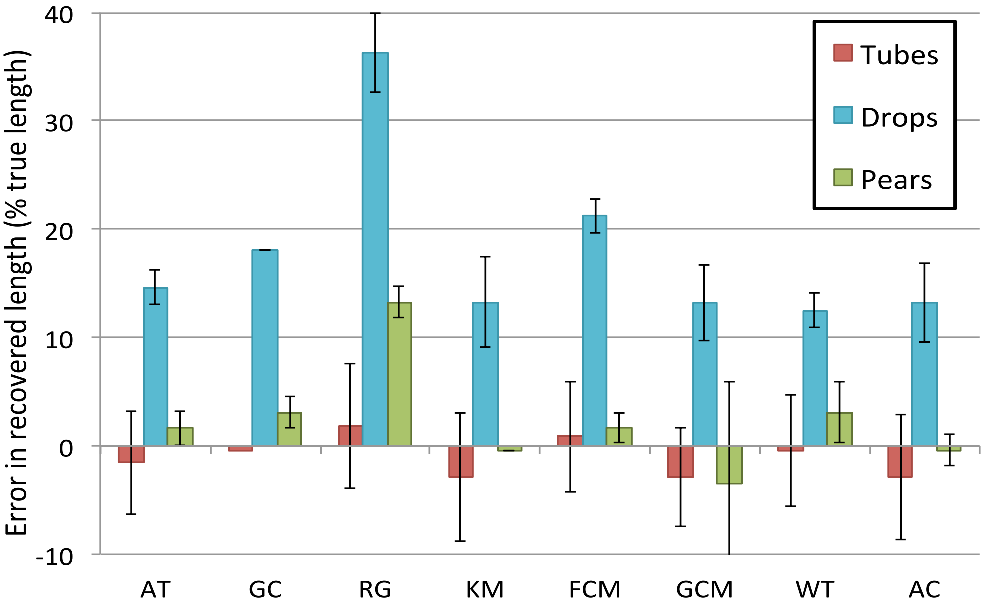


Figure 37. Error in the recovered length averaged on inserts of the same type, with error bars of one associated SD.

No correlation was found between the AR and DSC values for pear- or drop-shaped inserts. A decrease in DSC with increasing AR was observed for the delineation of tubes by some PET-AS methods. This was the case for AT, RG, FCM, GCM and AC, with differences between values for AR=2 and AR=3 of 1.7%, 1.3%, 3.8%, 3.6% and 4.8% of the value for AR=2, respectively.

II. B. 2. d. Discussion and conclusion

In this study, 16 different non-spherical inserts were specifically designed and imaged to better model clinical tumours. This type of data has not been used to date for the validation of automatic PET-AS methods. By using non-spherical test objects, this work extends the range of variation of image parameters on which PET-AS methods have been tested (e.g. FCM tested by Hatt *et al.* [91]), while remaining in clinically relevant conditions.

The small disparities observed between the targeted geometry and manufactured inserts were due to minor modifications of the object geometry during the manufacturing process, which were necessary for ensuring a good sealing and attachment of the inserts to the positioning rods. However, these were accounted of by modifying the reference contour's geometry accordingly, and therefore had little impact on the segmentation analysis.

Non-spherical inserts were introduced as more challenging test cases for PET segmentation, for which the ratio between surface and volume, and therefore the proportion of boundary voxels, is larger than for spheres. The thoroughness of the testing was confirmed by the lower performance obtained for the non-spherical inserts for most PET-AS methods, compared to previous results using spheres (cf. II. B. 1). The insert geometry most difficult to delineate was the torus, for which DSC values ranged from 0.31 to 0.82 (cf. Figure 35 a)). This can be explained by the presence of the "cold" centre, and small diameter of the torus "tube". A decrease of the delineation performance of the tubes with increasing ARs was also observed for some methods. This is due to the PVE, which affects a higher proportion of voxels in thin tubes, for which the boundaries are only a few voxels apart. However, the effect observed was small compared to the difference between spherical and non-spherical inserts (0.1-0.3 DSC compared to 0.1-0.7 DSC).

An iterative thresholding method, similar to the method AT presented in this work, was developed by Jentzen *et al.* [75], and validated with the NEMA IEC body phantom. The AT method presented in this work showed good performance (DSC>0.7) for spheres larger than 0.5 mL at all six TBRs, which is in agreement with their findings (error in recovered volume lower than 20% for spheres larger than 1 mL, at TBRs of 2.1-7.8). In addition the AT method produced the highest DSC score on the dataset used here (DSC>0.83, except for tori). The results are also in agreement with the good performance obtained by Drever *et al.* [80] evaluating an iterative threshold segmentation method on spheroids and two irregularly shaped inserts. In addition, the choice of test data allowed highlighting the robustness of AT on a range of geometrical inserts. The method's performance showed little variation with the geometry, as shown in Table 11 and Figure 35. AT was overall the best performing method on the dataset used, which included volumes ranging between 0.5 mL and 102 mL, and TBRs between 1.4 and 6.4. Voxel-based approaches may be more advantageous when segmenting heterogeneous structures as suggested elsewhere [75]. The detailed investigation of the performance of PET-AS methods on heterogeneous tracer distributions was outside the scope of this study. However, work is in progress at the Wales Research & Diagnostic PET Imaging Centre to test the methods in such conditions.

Region growing is used within current commercial image processing software in combination with fixed thresholding methods, because of its use of voxel connectivity, which guarantees that the result is a single connected region. In this work, however, a more complex algorithm was used, growing the region by including voxels with intensity within an interval around the region mean intensity value. This was based on work carried out by Day *et al.* [89], which showed a good agreement between their method and manual delineation of patient data by experts. The authors acknowledged the difficulty of validating methods on clinical images, for which the ground truth is difficult to obtain. The results complement the findings from Day *et al.*, as they provide images with known ground truth, designed to represent realistic clinical conditions. Although the RG method was superior in the delineation of spherical inserts and showed a good volume recovery for the delineation of complex geometrical test inserts (cf. Table 11), the measurement of the recovered length indicated a poor shape recovery in the specific case of inserts with narrow ends (cf. Figure 37). This is likely to be due to its underlying algorithm, which grows the region to delineate simultaneously in all directions. The performance of this method could be improved by adjusting the speed of the region growing in different directions according to its rate of expansion in each direction.

Geets *et al.* [98] found a systematic underestimation of the Spherical Lucite Phantom inserts (2.1-92.9 mL at TBRs of 1.5-15) by the watershed algorithm, which they attributed partly to the presence of glass walls. When using WT to delineate the thin plastic wall inserts, overestimation of the volume was observed for small spheres and for all non-spherical inserts (cf. Table 11). This could be explained by the use of plastic objects instead of glass in the present study, and/or by the absence of a pre-processing step in the present method compared to Geets *et al.* In addition to these results, the lack of accuracy of the method was highlighted for delineating inserts with thin ends, such as drops (cf. Figure 36). Similarly, the watershed-based method, evaluated by Drever *et al.* [120]., showed difficulties in recovering the diameter of large and small cylindrical inserts, although no systematic error was observed.

The GC method, which also segments the gradient of the target image, performed well for spherical inserts (cf. Figure 35), but showed much lower accuracy for small inserts and non-spherical geometries. The method failed to accurately recover the contour on some 2D images, which was noticeable for the drop-shaped inserts (cf. Figure 36). The difficulty of delineating complex geometries (especially when narrow ends are involved) could be inherent to the use of the image gradient. As Drever *et al.* [80], and Geets *et al*. [98] pointed out, the image gradient calculation is highly sensitive to noise and image blurring. GC and WT could therefore benefit from a partial volume correction, which Geets *et al*. used to improve the definition of the intensity gradient image [98]. It is also worth noting that both GC and WT methods are implemented as a slice-by-slice process initialised using a single seed voxel, which could explain why they only delineated part of the toroidal inserts (cf. Figure 35 c)).

Various versions of the K-means (KM), Fuzzy C-Means (FCM), and Gaussian Mixture Models Clustering (GCM) algorithms have been evaluated in the literature. Belhassen *et al*. [72] tested the FCM method on simulated data from the NCAT phantom and clinical PET images. They found large underestimation of the volumes in both cases, which was also observed in this study for both baseline and complex inserts (cf. Table 11 and Figure 35). It can be noted that the low performance of FCM due to its low sensitivity is more obvious for non-spherical inserts, as shown by Figure 35 a). However, the systematic error in volume obtained for the baseline study was not detected when developing the algorithm using thick-wall spheres from the NEMA phantom. This consolidates the results of previous work in which the importance of using realistic thin-plastic inserts for the evaluation and comparison of PET-AS methods was highlighted [131]. Hatt *et al.* [138] evaluated FCM (among others) on the IEC quality phantom, and suggested that the method is outperformed by region growing schemes on low contrast images, which the results of this study confirm (cf. Figure 35). Montgomery *et al*. [90] tested different clustering methods on spheroidal wall-less structures, and found that KM significantly underestimated the volumes, whereas the addition of Gaussian Mixture Modeling overestimated them. Such systematic errors can be explained by the choice of test inserts used to develop the methods (wall-less inserts compared to NEMA spheres in this study). In particular, the parameter used to assign voxels from the fuzzy region to tumor or background clusters is tuned according to the test data selected. In addition in this study, GCM performed well for both spherical and non-spherical inserts of volumes larger than 1.8 mL (cf. Table 11). The method models the spatial intensity distributions in both lesion and background, which could explain its good performance and robustness to complex shapes (cf. Figure 36).

The AC method developed in this study was based on methods developed primarily for MRI images, predicting a high accuracy in the recovery of complex geometrical shapes [114], [115], [139]. AC, implemented on a 2D basis, includes an "elasticity" parameter, controlling the length-to-surface ratio of the contour generated on each slice. The errors in object recovery (DSC) and insert length by AC, shown in Figure 36 and Figure 37, were relatively small compared to the other methods. However, the method failed to recover the toroidal geometry, which suggests that the elasticity parameter should be modified for targets with a central necrotic region.

This work shows that the thorough validation of automatic segmentation algorithms requires the use of test volumes of more complex geometries than spheres. The robustness and high performance of AT, observed for spherical inserts, was confirmed for the segmentation of non-spherical inserts. On the other hand, the use of non-spherical inserts added valuable information to the tests performed with spherical inserts, highlighting large errors obtained by the slice-by-slice gradient-based techniques, systematic under-segmentation from the fuzzy clustering method implemented in this work, and the lack of sensitivity of the region-growing algorithm for complex geometries. This work provides useful data for further optimisation of PET segmentation methods in clinically relevant reference conditions. One limitation of this study is that the tracer uptake heterogeneity was not modeled. Including imbricated or compartmented inserts would have added different FDG uptake regions, but making such inserts while maintaining good sealing and a regular wall thickness across the inserts would have required an industrial manufacturing process that was not available. The use of a printed subresolution sandwich phantom presented in the next chapter helped overcome this problem.

The experimental tests presented in this first chapter allowed the project to move forward by validating the PET-AS methods implemented, and showing their superiority to commonly used fixed thresholding methods FT42, FT50 and SUV2.5. In addition, the results have highlighted the most promising methods, such as AT and RG, as well as the weakest ones, such as FCM, GC and KM. The data also provided information useful for understanding and improving the methods, by using 3D implementation and pre-processing in some cases (cf. II. A. 2. b) or further tuning the methods (cf. II. A. 3). Some methods showed promising results in specific cases, such as GCM and AC in terms of robustness to image parameters, and shape recovery (in this section) and will benefit from further analysis. However, the approach of using geometrical plastic inserts with homogeneous uptake to simulate tumour regions remains highly unrealistic in a study investigating the segmentation of H&N tumours. Work presented in the following chapter will address these issues.

Chapter III. Development of an optimised segmentation framework

After ensuring optimal implementation of the different PET-AS methods, and comparing them in targeted challenging situations, further work focused on quantifying the accuracy of the methods for the delineation of realistic H&N PET images. The aim was to evaluate the accuracy of the different PET-AS method in various conditions modelling H&N background and tumour uptake as realistically as possible, and use the results of these studies to develop an optimised segmentation process applied to clinical data at Velindre Cancer centre. For this purpose, data from a printed sandwich phantom was used, as well as data from a PET simulator tool, both of which combined the advantages of a great flexibility in the definition of the FDG uptake to be modelled, and the availability of a known ground truth.

III. A. Evaluation of the PET-AS methods using a printed subresolution sandwich phantom

III. A. 1. Development and validation of the method

III. A. 1. a. Purpose

The previous work decsribed in this thesis highlighted several drawbacks of fillable phantoms, including fixed geometry, absence of heterogeneity modelling, and the presence of thick plastic walls around the target objects.

The use of printed uptake patterns has recently been investigated as a novel technique for generating radioactive sources for SPECT [140]–[143]. Work by Larsson *et al.* [140] and van Laere *et al.* [141] has taken this forward with the use of stacked radioactive printouts, applied to the generation of idealised SPECT images for experimental validation and comparison, and applications in neuroimaging activation studies respectively. Both studies suggest using similar techniques for PET. In their recent work, Sossi *et al.* [144] have applied the printing technique to the production of planar radioactive sources for PET, using conventional ink and ^{18}F nuclide printed on ordinary paper. A quantitative calibration study of the printing method was described in detail by Markiewicz *et al.* [145] for generating single-slice patterns with applications to brain imaging studies. However, the stacking of several printed patterns to produce a 3D object was not investigated in this study, which focused on the characterisation of the experimental setup. For this project, a collaboration was started with Robin Holmes (Bristol Royal Infirmary, Bristol, UK) who developed a similar technique with the purpose of generating 3D brain HMPAO images for SPECT imaging [146]. Such a printed phantom provided a great alternative to fillable phantoms, involving a physical 3D object to be scanned as well as the possibility of designing any given FDG uptake for modelling.

The work presented in the following sections therefore aimed at implementing a protocol for the use at Velindre Cancer Centre of such a technique, with FDG. The technique was given the name printed subresolution sandwich (SS) phantom because of the distance between two printout sheets, which is smaller than half the axial resolution of the scanner. The printed SS phantom technique was further used to produce realistic PET images of H&N lesions, for evaluating PET-AS methods.

This section focuses on the development of the protocol for using a printed SS phantom at Velindre Cancer Centre. This included:

- Demonstrating the feasibility of using a printed SS phantom for generating realistic PET images in a useful timeframe whilst minimizing the operator's radiation dose
- Characterising the performance of the technique in terms of printing homogeneity, reproducibility and accuracy

- Calibrating the phantom for accurate reproduction of the desired FDG uptake

In addition, the advantages and drawbacks of using a printed SS phantom for modelling PET target objects were investigated, by comparing the PET images obtained using the SS phantom and a cylindrical fillable phantom. This work focused on the phantom's ability to accurately reproduce spherical and non-spherical objects in terms of geometry and tracer activity.

III. A. 1. b. Methods

The phantom consists of 120 oval Polymethyl Metacrylate (PMMA) sheets of 2 mm thickness corresponding to axial slices, which can reach a maximum length of 24 cm when assembled with radioactive printouts. This is done using three plastic rods attached to a cylindrical PMMA support, held together with a PMMA sheet screwed at the top of the phantom. The phantom can then be scanned as a physical 3D object. A picture of the assembled 3D phantom is shown on Figure 38 a), along with the position of the phantom in the scanner on Figure 38 b).

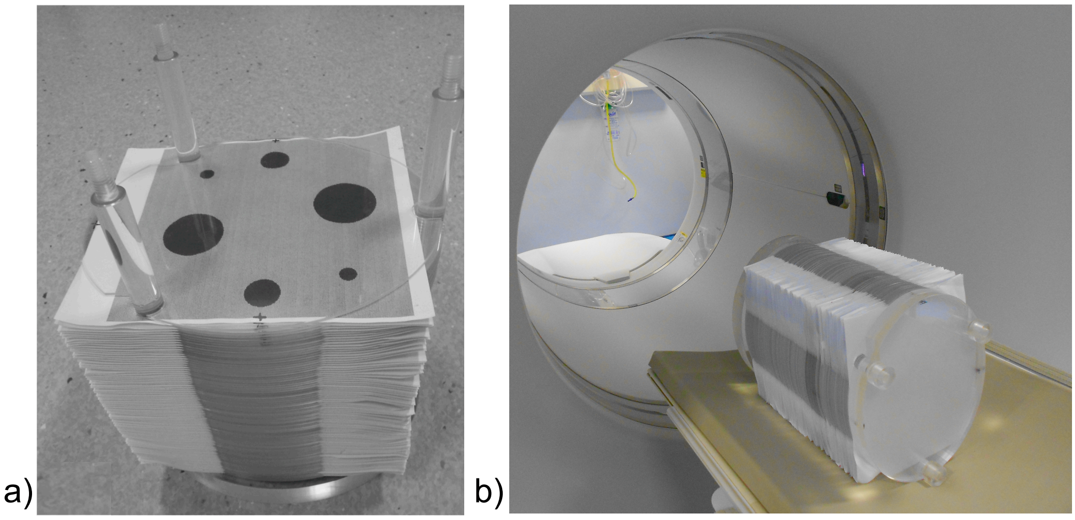


Figure 38. a) Half-assembled printed SS phantom and b) assembled phantom positioned on the scanner bed.

Plain A4 paper (80 mg per sheet) was cut to 168 mm x 197 mm and hole punched to fit into the phantom. Tracer uptake printouts were generated as grey level 3D images in Matlab, resampled to 2 mm slices, and printed on a HP deskjet 990 cxi

(Hewlett-Packard Limited, Berks, UK), using drop-on-demand thermal inkjet printing. The advantage of this type of equipment is its use of refillable ink cartridges, making it possible to add the desired quantity of radiotracer to the same cartridge before each set of experiments. The printing settings "normal" and "black & white" were chosen to minimise the printing time (and therefore the radiotracer decay and user exposure to gamma emissions) while ensuring a good printing quality. The corresponding printing speed is 6.5 pages per minute, with a resolution of 600 x 600 dpi.

The cartridge was filled with the desired FDG volume for each experiment and topped with black ink. The printing was done in a hot cell (Gravatom Engineering Systems Ltd, Southampton, UK), after leaving the cartridge upright with its dispensing head down for 20 minutes to homogenize its contents, as recommended by the manufacturer. All operations including filling the ink cartridge and assembling the phantom were done behind a lead-glass shield (Bright Technologies Ltd, Sheffield, UK). Any inaccuracy in the positioning of the pattern on the paper was corrected for by aligning cross-shaped markers printed at the left (L), right (R) and top (T), at a minimum distance of 10 mm to the uptake pattern, with reference markers drawn at a fixed position on the transparent PMMA sheet. The phantom was scanned in the available PET/CT scanner immediately after assembly with the protocol described previously (cf. II. A. 1. a). Operator exposure to the radioactive tracer was controlled using standard safety equipment (e.g. lead glass shields, shielded syringe carriers, hot cell) and monitored with electronic portable dosimeters (EPD) (RAD-60S, RADOS Technology, OY Finland) placed in the front pocket of the operator's laboratory coat.

III. A. 1. b. i. Evaluation of the printing homogeneity and reproducibility

The printing homogeneity was assessed with a flood field test using a nominal grey level rectangle of 210 mm x 280 mm printed on A4 with radioactive ink and imaged in the PET/CT scanner as a single sheet in the coronal plane. Intensity profiles of 1.3 mm (3 pixels) width were taken across the digital image obtained in vertical and horizontal direction. The reproducibility of the printing was evaluated with a similar flood field, printed 66 times with radioactive ink and scanned in the assembled phantom. The same 27 mm x 27 mm ROI was reproduced in the centre of each transverse slice on the PET imaged obtained, and the mean intensity value and SD were measured for each ROI.

In addition, the number of counts was measured along two 50 mm paper strips corresponding to homogeneous printing of the mixture of black ink and radiotracer in left-right and anterior-posterior directions, using the thin layer chromatography (TLC) equipment iScan (CANBERRA Nuclear Measurements Business Unit, Uppsala, Sweden) at a speed of 1 mm/s.

III. A. 1. b. ii. Evaluation of the accuracy of the phantom assembly

The accuracy of the paper positioning in the phantom was assessed using the cross-shaped markers described in the previous section. The markers were printed with the same radioactive ink as the printout, and were therefore visible on the PET image obtained. Their alignment was evaluated by determining their position on the resulting PET image for each slice, as the highest intensity voxel in a 5 x 5 voxel square drawn around the imaged marker. For each one of the L, R and T markers, the difference in positioning with the average marker position was measured on each slice.

III. A. 1. b. iii. Evaluation of the accuracy of grey level printing

The linearity of the grey level printing was first evaluated by investigating the relationship between the intensity specified and the quantity of ink printed. For 10 grey level values, corresponding to 5, 16, 26, 37, 53, 68, 80, 89 and 100 % of the maximum intensity, a 140 mm x 160 mm homogeneous rectangle was printed 5 times with a mixture of black ink and radiotracer. The paper was weighed before and after printing to measure the amount of ink added by the printer. The weight of ink printed for each grey level, averaged over the 5 instances, was then plotted against the grey level values specified.

Next, the relationship between grey level specified and number of photon counts was investigated, in order to derive a grey level calibration protocol for the phantom. For this purpose, 20 distinct homogeneous 30 mm x 30 mm squares of grey level values evenly spaced within 5-100% were printed one by one with the radioactive ink mixture. The number of counts detected across the different rectangles was then measured using the TLC equipment described previously. The average number of counts was measured via TLC measurement across each square, with correction for radioactive decay to compare all values at the same time point. This was done at three instances with different activity concentrations in the ink at the time of measurement corresponding to different volumes of black ink added to 2 mL of the same radiotracer solution. The results of the previous experiment were used to plot the count values obtained against the amount of ink printed on the paper, to focus on the imaging process.

III. A. 1. b. iv. Comparison to a thin-wall fillable phantom

The following set of experiments tested the capability of the assembled SS phantom in replicating a full 3D phantom. For this purpose, this work aimed at reproducing the spherical and non-spherical removable inserts used in II. B. 2 with the Raydose phantom. For the SS phantom, printout patterns representing the different inserts positioned in the fillable phantom were derived.

The PET images obtained were evaluated in terms of intensity distribution within the inserts and recovered object geometry. The true object contour and a background ROI, which was measured within a sphere of the same size as S58 placed in the background, were used to extract mean intensity values with associated SD and the recovery coefficient (RC, cf. II. B. 1), in both inserts and background. The true concentration was the filled-in activity for the fillable phantom. For the SS phantom, the amount of ink printed corresponding to sphere S58 was calculated, leading to the corresponding amount of tracer, which was then divided by the sphere volume. The COV was calculated for spheres and background as a measure of heterogeneity (cf. Eq. 19).

The recovery of the objects' dimensions in the image was done by segmenting all objects with a background-subtracted threshold of 50% of the maximum intensity, using the background mean calculated previously. The contours obtained were then compared to the true object contours, with metric HD described in II. A. 1. e) (cf. Eq.15). RC, COV and HD obtained on the different inserts were compared pairwise between values obtained for fillable and SS phantom, using the Mann Whitney U-test to detect any significant differences. The level of significance was set to *p=0.05*.

In addition, the diameters of the contours obtained with segmentation were measured in superior-inferior (for a subject positioned head first supine on the scanner bed, perpendicular to PMMA sheet for the phantom), and left-right directions (parallel to PMMA sheet). The use of a threshold value of 50% of the maximum intensity, with subtracted background intensity provided an evaluation of the FWHM of the objects' intensity profiles. The values obtained were compared to the true object dimensions, and the Mann-Whitney U-test was used again to identify significant differences in the error made in the dimension of the inserts in both directions.

Finally, the effect of the PMMA sheets on the imaging quality was assessed by printing out nine spheres of 20 mm diameter, named S1-S9, with homogeneous uptake corresponding to the grey levels used previously placed in a cold background. Printouts for these small spheres were generated for 0.1 mm and 2 mm slice gaps, allowing a comparison between the spheres printed and assembled between the 2 mm PMMA sheet, and the same spheres modelled with paper printouts only. The continuously printed spheres and 2 mm spaced printouts were assembled in the same phantom and scanned simultaneously.

III. A. 1. c. Results

III. A. 1. c. i. Evaluation of the printing homogeneity and reproducibility

Figure 39 shows the number of counts measured by TLC along two stripes with the same homogeneous grey level printed in left-right and anterior-posterior directions. Measurements started at 20 mm from the edge of the printout. The printout was 2.6 mm wide in left-right direction and 3 mm thick in anterior-posterior direction, leading to a slight difference in average number of counts across the two stripes (11 and 14 respectively). The highest variations observed were 240% and 198% of the average value in anterior-posterior and left-right directions respectively.

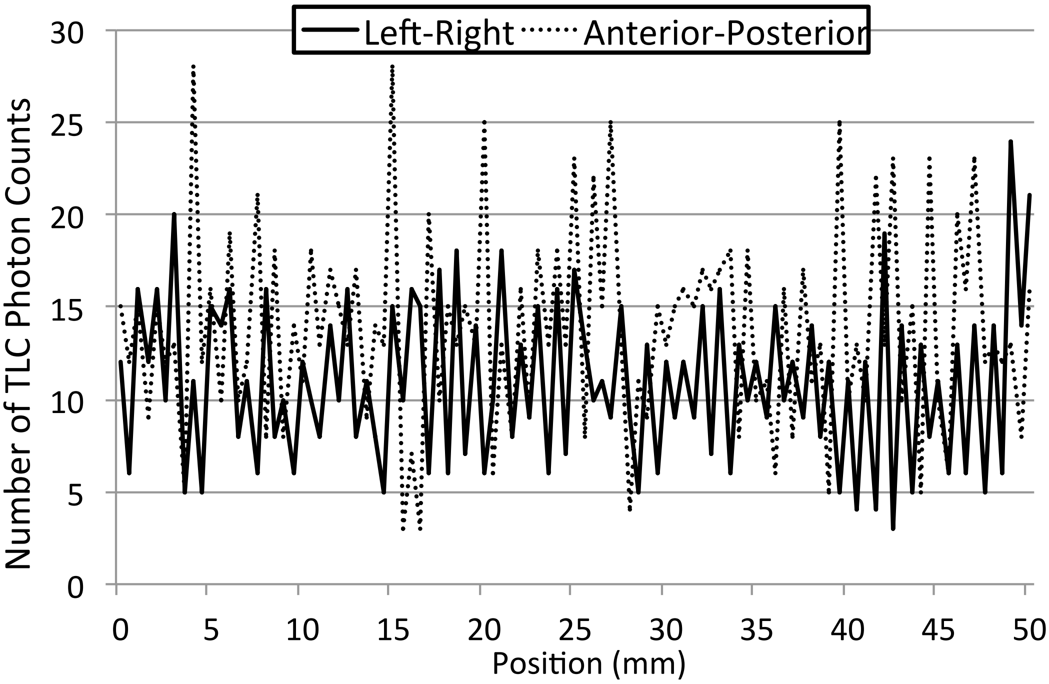


Figure 39. Results of TLC measurements of the number of counts detected along 50 mm of grey level stripes printed in left-right and anterior-posterior directions.

Intensity values measured through the printed homogeneous grey level in both directions of the paper showed slightly higher intensity levels at the extremities of the sheet. The average difference to the mean value was 2.3% and 3.0% of the mean value in horizontal and vertical directions, and ranged between 0.5-8.7% and 0.6-7.0% respectively. For the 66 consecutive printouts of the same homogeneous grey level pattern, the average difference to the mean ROI value was 4.2%, with a variation range of 0.27 - 12.8%.

III. A. 1. c. ii. Evaluation of the accuracy of the phantom assembly

Figure 40 a) shows the error in mm of the positioning of the alignment markers measured on the resulting PET image at three different locations in the image. The markers are shown on one slice of the printout template on Figure 40 b). Although measurements were limited to the voxel size, errors values were systematically smaller than 2.3 mm, which corresponds to a displacement of one voxel. In addition, no systematic error was observed. The alignment of the markers, corresponding to accurately aligned printout sheets, can be visualised on Figure 40 c) as the vertical stripes on either side of the PET image.

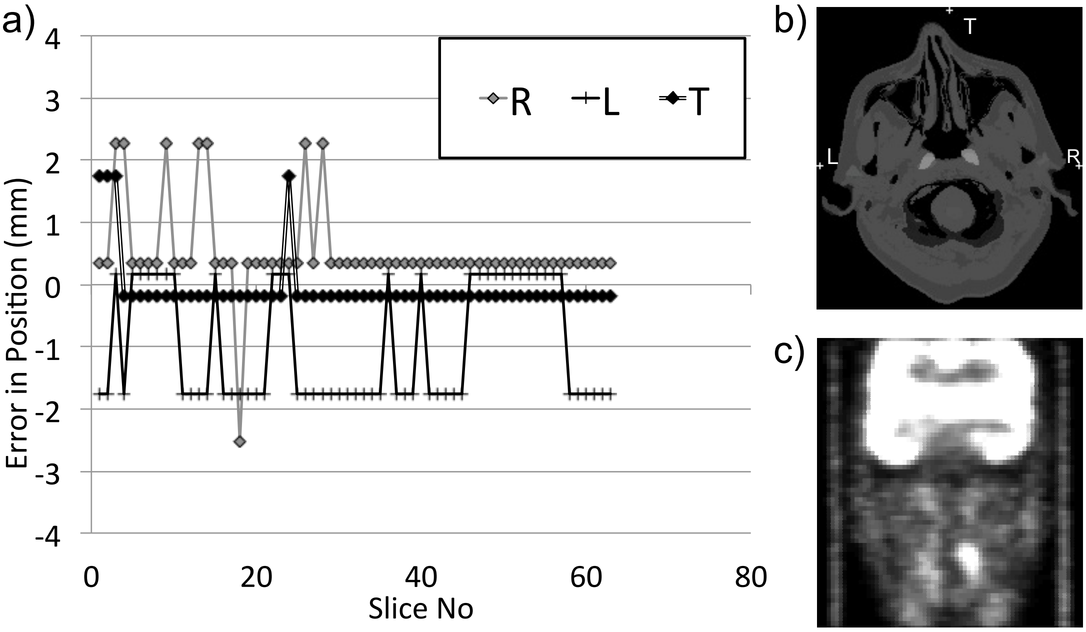


Figure 40. a) Error in positioning of the alignment markers at the left (L), right (R) and top (T) of the printout, b) cross-shaped markers shown on single printout template, c) sagittal view of PET image obtained with axially aligned markers L and R.

III. A. 1. c. iii. Evaluation of the accuracy of grey level printing

Figure 41 a) shows the grey level patterns printed and measured with TLC. A non-linear relationship was obtained between the specified grey levels and the amount of ink deposited on the paper when printing with a mixture of black ink and radiotracer, as depicted in Figure 41 b). The curve was best fitted to a third degree polynomial (R^2>0.99). The equation of this curve was used to transform the grey levels specified in the following experiment into the amount of ink deposited on the paper.

Figure 41 c) shows the relationship linking the amount of ink deposited on the paper and the number of counts measured in each square with TLC, for different activities in the cartridge. The combined data for all activities, using values normalised to the intensity obtained at the highest level for each case, was fitted to a proportional relationship with R^2=0.98 as shown in Figure 41 c).

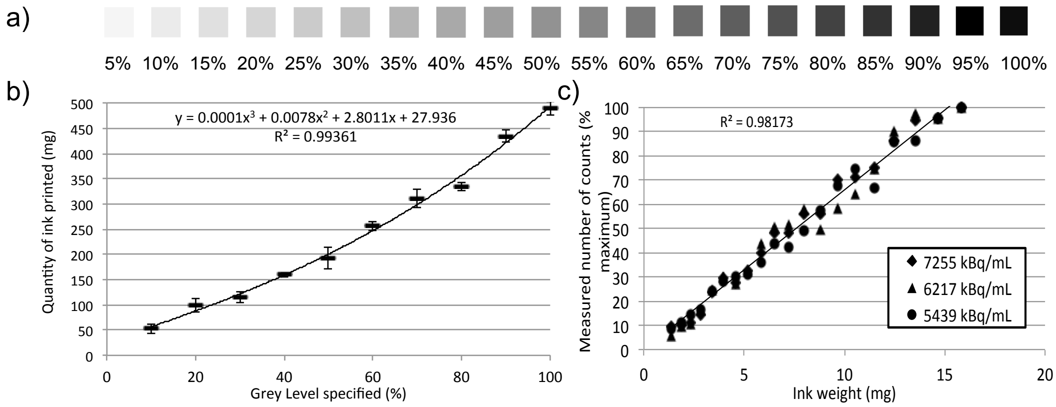


Figure 41. a) Printout used for the calibration, b) average measured weight of deposited ink and associated SD displayed as vertical error bars, c) number of counts measured via TLC for printing with three different radiotracer concentrations in black ink.

III. A. 1. c. iv. Comparison to a thin-wall fillable phantom

Table 12 shows the different phantoms scanned, together with the objects they contained and their targeted TBRs. Some experiments combined two types of objects in the same printed SS phantom, which was large enough to ensure at least 2 mm spacing between the boundaries of different objects. In the case of the fillable phantom, spheres S10, S20, S38 and S15, S37, S58 were scanned in two different instances. Differences between the fillable insert volumes and the volumes of the printed objects, did not exceed 9% of the fillable volume, and were due to the interpolation of the true contour to the uptake map grid, which had a pixel dimension of 2 x 2 mm in the transverse plane. The TBRs obtained for the SS phantom ranged within 6-6.8 (ratio of activities printed), and were within 5% of the values obtained for the fillable phantom (ratio of activities filled-in), which allows a direct comparison of the two cases. The radiotracer volume was approximately 260 mL for the SS phantom (163 mm x 159 mm pattern printed on 100 sheet of 0.1 mm thickness), and 6080 mL for the filled Raydose phantom with an inner height and diameter of 160 mm and 220 mm respectively.

Fillable Phantom			SS phantom		
Scan	**Objects**	**TBR**	**Scan**	**Objects**	**TBR**
F1	S10, S20, S38	3	SS1	All spheres	3
F2	S15, S37, S58	3	SS2	All spheres	5
F3	S10, S20, S38	5	SS3	All spheres	8
F4	S15, S37, S58	5	SS4	Ellipses and Pears	6
F5	S10, S20, S38	8	SS5	Tori and Drops	6
F6	S15, S37, S58	8			
F7	E15, E20, E30, E38	6			
F8	To20, To28, To38	6			
F9	Tu38 a-c	6			
F10	P38 a-c	6			
F11	D38 a-c	6			

Table 12. Summary of scans acquired for both fillable and SS phantom.

PET images were obtained for both fillable and SS phantoms modelling the six spheres described above and the non-spherical objects presented in II. B. 2. Figure 42 shows coronal views of the images obtained for the spheres S15 and S58 at TBR=8 for both phantoms. The dashed lines represent the intersection with the transverse plane, from which the left-right profiles across the spheres were drawn. The dash-dotted lines represent the superior-inferior profiles taken across S58.

Figure 43 shows the superior-inferior and left-right profiles obtained for both phantoms at TBR=5 for the smallest sphere S10, for which differences were the most visible. The intensity values for the fillable phantom were slightly higher than for the printed SS phantom in both spheres and background, so that the TBR remained very close. The presence of plastic rods holding the inserts, visible under the spheres on Figure 42 b), results in a lower background intensity on one side of the inserts, which is visible at the right of the peak corresponding to the insert on the superior-inferior profiles shown on Figure 43 a), drawn through the insert and rod. Differences can be observed at background intensities, for which profiles for the fillable phantom show more fluctuations than for the SS phantom. Finally, the images also show slightly larger FWHM for the SS phantom.

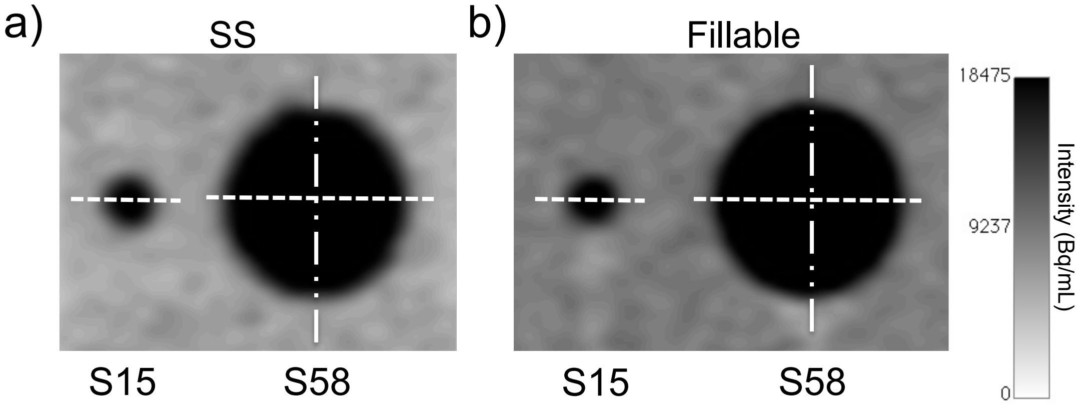


Figure 42. Coronal views of S15 and S58 spheres at TBR=8, for both types of phantom. Dashed lines and dash-dot lines represent transverse and superior-inferior profiles taken respectively.

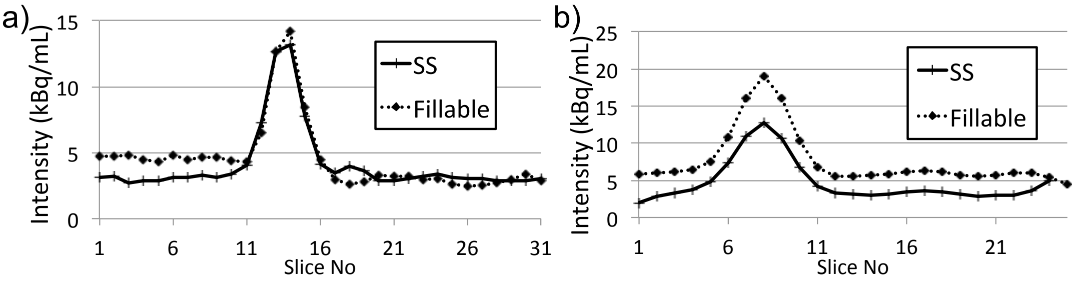


Figure 43. Comparison of profiles obtained for both phantoms at TBR=5 for sphere S10 in a) superior-inferior and b) left-right directions.

The PET images obtained for both fillable and SS phantoms, modelling different geometrical objects, were then compared. Table 13 provides a comparison of the COV, given in % of the mean intensity, the geometry recovery (HD, in cm), and the RC, in % of the true activity concentration, in the inserts and background. Values are reported in terms of average (SD) for each sphere across the different TBRs, for each insert geometry across inserts and for the background ROIs for scans at TBR=5. COV values for the fillable phantom were slightly higher in the background, and systematically higher inside the inserts compared to the SS phantom. This was found statistically significant with the one-tailed Mann-Whitney U-test for spheres (*p=0.03*) and non-spherical inserts (*p=0.004*). HD values were larger for the fillable phantom in some cases including the three smallest spheres, indicating a lower accuracy in modelling the object shape for smaller inserts. There was no overall significant difference between HD values for both phantoms. The RC values were systematically higher for the fillable phantom for both spherical and non-spherical inserts

	COV (%)		HD (cm)		RC (%)	
	SS	**Fillable**	**SS**	**Fillable**	**SS**	**Fillable**
S10	0.19 (0.00)	0.26 (0.00)	1.37 (0.34)	1.50 (0.26)	46.7 (3.2)	63.1 (3.6)
S15	0.28 (0.08)	0.48 (0.02)	0.86 (0.02)	0.93 (0.13)	69.7 (5.2)	82.4 (3.7)
S20	0.58 (0.00)	0.76 (0.01)	0.96 (0.84)	1.20 (0.20)	73.5 (4.9)	94.5 (4.5)
S30	0.77 (0.26)	1.40 (0.01)	1.41 (0.13)	1.20 (0.18)	81.3 (6.5)	102 (4)
S38	1.60 (0.00)	2.00 (0.00)	1.45 (0.00)	1.30 (0.19)	84.9 (5.3)	112 (5)
S58	1.50 (0.69)	3.60 (0.00)	1.53 (0.19)	1.10 (0.29)	90.1 (3.4)	113 (4)
Drops	1.60 (0.01)	2.00 (0.10)	2.00 (0.01)	1.30 (0.01)	71.7 (1.1)	109 (3)
Pears	1.60 (0.01)	1.90 (0.01)	2.00 (0.01)	1.20 (0.01)	73.6 (0.8)	107 (0)
Ellipses	1.10 (0.33)	1.20 (0.33)	3.50 (0.08)	1.10 (0.020)	74.2 (6.3)	108 (6)
Tori	0.67 (0.16)	0.72 (0.19)	1.10 (0.00)	1.60 (0.037)	74.7 (4.9)	108 (7)
Tubes	1.60 (0.01)	2.00 (0.01)	1.60 (0.01)	1.10 (0.010)	71.4 (0.4)	105 (1)
B	3.20 (0.29)	5.0 (1.0)	-	-	-	-

Table 13. Comparison of COV values, Hausdorff Distance and activity concentration recovery RC in the background (B) and insert regions, averaged TBRs for spheres, on the different ARs for other inserts, and on all scans at TBR=5 for the background.

In addition, measurements of the objects' FWHM in superior-inferior and left-right directions showed systematically higher values for the SS phantom (except for S10), which was statistically significant with the Mann Whitney U-test (*p<0.01).* For non-spherical objects, with results given in Table 14, absolute errors in dimensions were lower for the SS phantom, or equal for both phantoms, for 14 out of 16 and for 12 out of 16 inserts in superior-inferior and left-right directions respectively. This included all ellipsoids and drop-shaped objects. The errors in superior-inferior dimensions of the drops and tube-shaped objects were smaller for the SS phantom, but reached 15.8% of the object length for D38b. The one-tailed Mann-Whitney U-test results comparing absolute errors in apparent dimensions across non-spherical objects showed significantly higher values for the fillable phantom in the left-right direction (*p=0.019*). Differences in the superior-inferior direction were not significant.

	Ellipsoid				Torus			Tube			Pear			Drop		
Name	**E38**	**E28**	**E20**	**E15**	**To20**	**To28**	**To38**	**Tu38a**	**Tu38b**	**Tu38c**	**P38a**	**P38b**	**P38c**	**D38a**	**D38b**	**D38c**
								Superior-inferior dimensions (mm)								
True	16	24	37	46	18	22	35	56	64	70	56	70	82	56	66	80
SS	13	23	36	46	19	22	31	56	62	72	56	62	79	49	56	69
F	13	20	36	43	17	24	32	52	59	69	56	65	75	49	56	62
								Left-right dimensions (mm)								
True	14	20	28	37	5.0	8.0	11	28	27	24	37	31	29	38	36	34
SS	15	19	27	33	5.9	6.9	10	27	25	28	26	23	26	28	26	31
F	13	14	25	32	8.1	8.8	12	27	23	25	26	24	26	28	23	30

Table 14. Comparison of dimensions measured with calliper on the insert (cf. II. B. 2) and on SS and Fillable (F) phantom PET images, in superior-inferior and left-right directions.

Sphere	S1	S2	S3	S4	S5	S6	S7	S8	S9
2 mm spacing	41.9	80.9	81.7	84.9	94.6	92.4	78.3	81.8	87.6
Continuous	28.5	85.9	106	101	107	110	102	98.5	109

Table 15. Activity Recovery Coefficient RC (% true activity) for 2 mm-spaced and continuous printouts of spheres S1 to S9, corresponding to grey levels of 6-100%.

Table 15 shows up to 22% higher RC when using continuously printed patterns instead of 2 mm PMMA sheet, except for sphere S1 which corresponded to a very low grey level value. Values obtained with continuous printing and assembly are close to the values obtained for the fillable phantom (cf. Table 13).

III. A. 1. d. Discussion and conclusions

This work investigated the advantages and drawbacks of using the printed SS phantom for generating realistic PET images, with the purpose of generating realistic PET images for testing the PET-AS methods. This has taken forward the work published by Markiewicz *et al.* [145] to the generation of a 3D object using a large number of printed sheets. The printer was fully characterised and a custom calibration procedure was developed for it. The amount of time necessary for a single operator to prepare the phantom, after optimising the sequence of tasks, was approximately 80 min, including a) filling the cartridge (10 min), b) leaving the contents of the cartridge to homogenize (10 min), c) printing (30 min) and d) assembling (20 min) and e) scanning (10 min). This allowed scanning the phantom within one half-life of the ^{18}F decay. The total exposure to the radioactive tracer for one session with a single scan was 4 μSv on average, which is comparable to the exposure of manipulating a conventional fillable phantom, with the EPD in the same position. Aerosol exposure was made negligible by printing all sheets in a closed hot cell.

A good homogeneity and reproducibility of the grey level printing were obtained with the equipment used for this work (cf. Figure 39). The results of the TLC showed large variations within a 3 mm stripe of homogeneous grey level, which are likely to be due to background noise at the low activities used. However, these small variations only led to variations on the PET image lower than 8.7% of the average. Higher quality printing equipment and automatic paper cutting systems would further improve the method in terms of accuracy as well as speed. The effect on the PET images of blurring in the phantom due to the imaging process is expected to be minimal, since the distance between two printouts (2 mm) is smaller than half the Full Width at Half Maximum of the PET imaging system in the axial direction. This also means that using planar printout sources closer than 2 mm from each other would not improve the blurring effect. However, the presence of plastic as a medium (instead of water for a fillable phantom) between the radioactive sources may lead to shorter positron range for the printed SS phantom. This has a potential for slightly better image quality, although it may not be as relevant for modelling patient data.

The results given on Figure 40 show the satisfactory accuracy achieved in aligning the printout patterns with each other. The maximum error in positioning obtained, corresponding to one voxel difference between slices. The absence of correlation between errors obtained for the three markers on the same slice may indicate uncertainties in the measurement on the PET image rather than a misalignment of the printouts. Although the results suggest that some inaccuracies in the paper alignment exist, the level of accuracy obtained was considered satisfactory for this study. Further work on the technique will address this issue.

The results published by Markiewicz *et al*. [145] for a similar two-dimensional study showed a non-linear relationship between grey level specified and obtained on the PET image, which was best fitted to a third degree polynomial. In the present work, the phantom process was broke down into the printing process, and the imaging process. Results on Figure 41 show that the non-linear effect observed by Markiewicz *et al*., as well as in this work, is due in this case to the printing process, with a non-linear relationship between the amount of ink and radiotracer mixture printed and the grey levels specified.

The focus of the quantitative analysis was to evaluate the ability of both phantoms to accurately reproduce the objects described II. B. 1 and II. B. 2, rather than reproduce the whole Raydose phantom, which is made impossible by the different geometry of both phantom structures. The spherical and non-spherical objects were modelled in a homogeneous background, and the geometry recovery and homogeneity of the objects with both SS and Fillable phantom was evaluated. The total activity present in the phantom, however, and therefore the total number of photon counts was different for both phantoms. This is due to the smaller quantity of tracer included in the printed SS phantom, representing approximately 3.7% of the fillable phantom radiotracer volume (cf. III. A. 1. c). Exact photon statistics of a fillable phantom cannot be reproduced with the present version of the printed SS phantom, because of the difference in material encountered by the photons (water for the fillable phantom, PMMA and paper for the printed SS phantom). However, the printed SS phantom allows modelling any given uptake pattern, which is its largest advantage on fillable phantoms. This work aimed at investigating additional advantages in terms of control over image heterogeneity and object geometry.

Coronal views on Figure 42 and profiles given on Figure 43 showed no effect of the presence of the 2 mm gaps on the PET scans obtained with the SS phantom. On the contrary, the background intensity profiles obtained for the SS phantom in superior-inferior direction (perpendicular to the PMMA sheets) showed less variation compared to the fillable phantom. The good homogeneity observed when printing large homogeneous grey level regions confirmed this finding. The data also showed that the SS phantom provided systematically smaller COV than the fillable phantom, especially inside the spherical objects modelled, where values were up to 58% lower for the SS phantom (cf. Table 13). The higher COV obtained for the fillable phantom in the background regions could be due to insufficient mixing of the radiotracer and water solution, which was made difficult by the large phantom size, and limited in time to minimise the operator's exposure to radiation. The solution used to fill the spheres, however, was prepared in a 1 L sealed vial behind leaded glass and was carefully shaken to ensure accurate mixing. Therefore, the higher COVs obtained in the spheres are more likely to be due to the presence of the inactive plastic walls causing lower voxel values at the sphere boundaries, due to the PVE and scatter inherent to PET imaging.

Both phantoms showed similar HD values for the recovery of the object shape using a 50% of the maximum intensity background-subtracted threshold. HD values obtained with the fillable phantom were higher for the three smallest spheres, indicating lower accuracy in the shape recovery. In addition, the comparison of the dimensions measured in superior-inferior and left-right directions showed that the measured insert diameter was closer to the true value for the SS phantom compared to the fillable phantom for all spheres except S10. The benefit of using printed patterns instead of plastic inserts was also highlighted for the left-right dimensions of non-spherical objects, for which the SS phantom achieved a better recovery of the object shape, yielding significantly lower errors in the left-right direction (cf. Table 14). This shows that at present, the technique developed with the materials available can model geometrical objects with similar and even higher accuracy than the fillable phantom used for comparison, which contained inserts with very thin plastic walls. However, it should be noted that the presence of the 2 mm spacing between printouts makes it challenging to model details smaller than 2 mm in the superior-inferior direction, such as the thin drop-shaped objects used in this work. This could explain the non-statistically significant improvement on the fillable phantom observed in the superior-inferior direction.

Although the object geometry recovery was higher for the SS phantom, the recovery of the activity present in the imaged objects was systematically lower compared to the fillable phantom, with differences up to 26% (cf. Table 13). This is likely to be due to the thick cold plastic sheets separating the different printouts, and lowering the recovered activity due to the PVE. A further experiment involving nine homogeneous spheres showed that using continuous assembly of 0.1 mm printouts allows increasing the RC up to 22% compared to 2 mm PMMA spacing, reaching values similar to the ones obtained with the fillable thin plastic objects (Table 15).

The SS phantom technique showed great potential for modelling realistic biological tracer distributions. The flexibility in the design of tracer uptake patterns allows lesions to be represented with any geometry or uptake distribution, modelling heterogeneities, necrotic regions and, theoretically, microscopic tumour extension. The absence of a sufficient volume of radiotracer still represents the main limitation of the approach described in this investigation, as it does not at present allow for an accurate reproduction of the scatter properties for a 3D object or the total number of counts. Tri-dimensional printing could combine the advantages of the SS phantom to the benefit of having a 3D volume of radiotracer available. This promising idea was investigated by Miller *et al.* [147], but requires high level and costly equipment. In addition, more work is needed to overcome the difficulties of incorporating radiotracer molecules within the printing material, and this was therefore not investigated in this work. Alternatively, the use of continuous printing showed a potential for good activity recovery, combined with the higher geometry recovery and heterogeneity control of the printing technique. Another limitation of the phantom investigated in this work is the absence of a CT component, which would improve its usefulness for a number of PET studies. These ideas will be taken forward in a future project.

III. A. 2. Evaluation of PET-AS methods for realistic H&N data

III. A. 2. a. Purpose

Previous work (cf. II. A. 3 and II. B) focused on evaluating and comparing the delineation accuracy of a number of segmentation methods, using controlled fillable phantoms and varying different parameters, to reveal the different relative accuracies for the PET-AS methods tested in different conditions. The present work aims at evaluating the PET-AS methods with realistic images of H&N cancer lesions, which corresponded to the clinical objective for this project. Realistic oropharyngeal PET data, covering a wide range of clinical scenarios encountered for oropharyngeal cancer, was generated using the printed subresolution sandwich (SS) phantom described in III. A. 1. and the help of a radiologist expert in H&N cancer. This data was then used to cross-compare the ability of the different PET-AS tested for accurately delineating H&N lesions, and conclude on the optimal PET-AS process to use.

III. A. 2. b. Methods

The printed SS phantom described in III. A. 1. was used to generate a range of clinically relevant 3D H&N models with tumour uptakes of known ground truth. This work aimed at reproducing typical normal H&N FDG uptake, with the addition of realistic H&N lesions for oropharyngeal cancer patients.

The printout template was generated from an available clinical PET/CT scan, by segmenting a selected number of anatomical structures on CT or PET, and assigning to each one a grey level value corresponding to its mean FDG uptake on the PET image. The resulting image was resampled to 2 mm slices, representing transverse slices of the H&N, in order to fulfil the requirements for the SS phantom. The choice of the PET scan and the design of the final template were both reviewed by a radiologist. Table 16 lists the anatomical structures present in the H&N template as well as the activities assigned, and slices of the final template obtained are shown on Figure 44.

Structure delineated	Delineation method	Intensity (Bq/mL)
Skin	Threshold on CT	1500
Fat	Subtraction of other structures from outline	1500
Soft tissue	Manual on CT	2000
Bone	Threshold on CT	0
Grey matter	Threshold on PET	18000
White matter	Threshold on PET	6000
CSF, air cavities	Threshold on CT	0
Tonsils & vocal cords	Manual on CT	6000
Extra ocular muscle	Manual on PET	10000
Eyeballs	Manual on PET	4000
Spinal cord	Manual on CT	3000
Parotids	Manual on CT	2500
Submandibular glands	Manual on CT	3500

Table 16. Values assigned to the different anatomical structures of the H&N printout template.

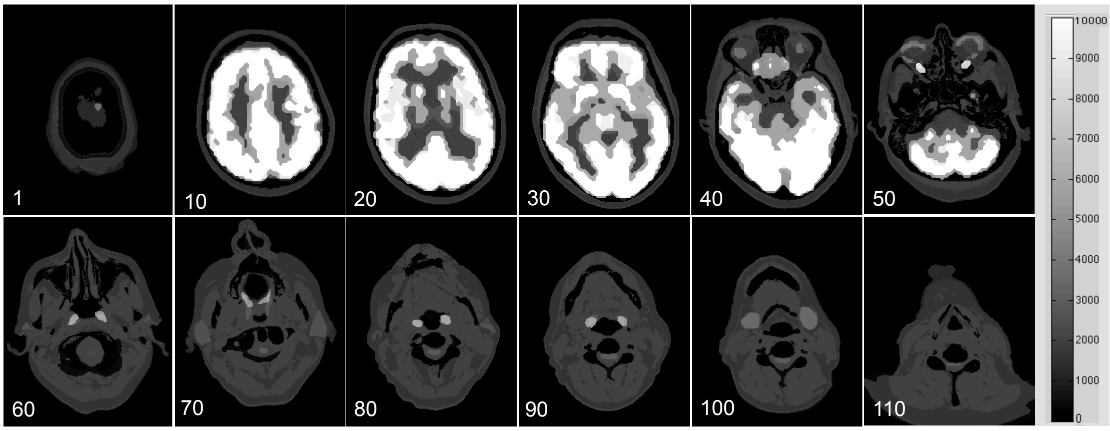


Figure 44. Selection of single 2D slices (with corresponding slice number) of the FDG uptake map generated from an existing PET/CT scan, with associated colour bar for Matlab matrix values.

For the different experiments, irregularly or spheroid-shaped tumours, drawn on the original CT image, were added to the background FDG uptake map, as shown on Figure 45. The geometry, size and location of the tumour printouts were reviewed by a radiologist as relevant for the modelling of tumours in this work. Tumour locations chosen included the base of tongue, tonsils and parotid space. Various tumour uptake distributions were modelled, as well as different TBRs. Table 17 summarises the different scans and corresponding lesions modelled. Heterogeneous tumour uptakes modelled included:

- Gaussian smoothed: homogeneous uptake smoothed with a Gaussian filter to model higher uptake at the centre.
- Necrotic: homogeneous high uptake with a homogeneous region of typical soft tissue uptake at the centre of the tumour
- Gaussian necrotic: necrotic uptake smoothed with a Gaussian filter
- Noisy: random distribution of intensity values across the tumour, using the Matlab function rand, with a SD of 20% of the mean intensity.

		Irregular lesions			**Spheroidal lesions**
Scan	**TBR**	**Uptake**	**Scan**	**TBR**	**Uptake**
1	2	Homogeneous	9	2	Homogeneous
2	4	Homogeneous	10	4	Homogeneous
3	6	Homogeneous	11	6	Homogeneous
4	8	Homogeneous	12	8	Homogeneous
5	5	Gaussian smoothed	13	10	Homogeneous
6	5	Necrotic	14	5	Homogeneous
7	5	Necrotic smoothed	15	5	Homogeneous (larger size)
8	5	Noisy	16	5	Homogeneous (smaller size)
			17	5	Necrotic
			18	5	Necrotic smoothed
			19	5	Noisy
			20	5	Gaussian smoothed

Table 17. Summary of scans acquired, each containing a tongue, tonsil and parotid lesion, for different TBRs, sizes and FDG uptakes.

Figure 45 a) shows a transverse slice of the original CT with contours corresponding to the tongue, tonsil and parotids irregular lesions. The tonsil, tongue and parotid lesions had volumes of 4.03, 4.35 and 3.99 mL respectively, while the spheroids had volumes of 11 mL. For the purpose of this series of experiments, the complexity of each irregular lesion's geometry was evaluated by calculating the HD comparing the lesion contour to a sphere of the same volume centred on its centre of mass. The different uptake patterns modelled are shown on Figure 45 b).

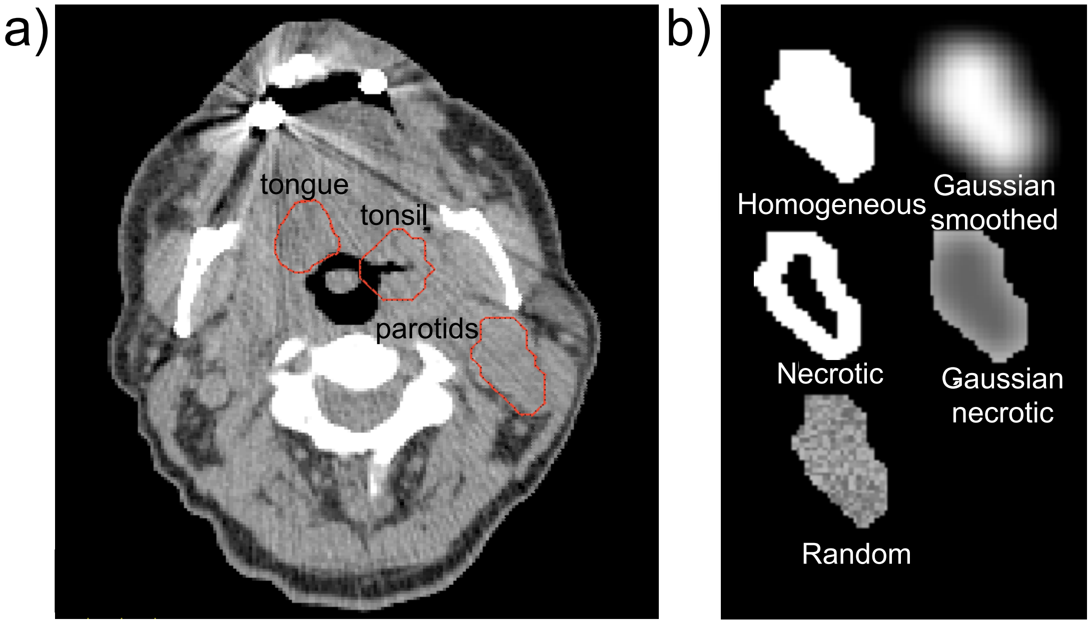


Figure 45. a) Contours drawn on original slice No 134 for tongue, tonsil, and parotid lesions added to the FDG uptake map, and b) tumour patterns modelled on the different uptake maps shown for the parotid lesion.

The phantoms obtained for each case were scanned with an activity concentration in the cartridge of about 1500 kBq/mL, as this provided a PET image with activities corresponding to the original PET scan.

All target lesions modelled were segmented for all PET-AS methods and evaluated with the accuracy metrics described in II. A. 1. e). The ground truth was obtained from the printout template containing the lesion stored in Matlab, which was resampled to the PET image grid. The methods' accuracy was compared across irregular and spheroidal lesions, for heterogeneous, homogeneous and necrotic uptake. The accuracy obtained was also compared for the different irregular lesions drawn.

III. A. 2. c. Results

A total of 60 H&N test lesions were obtained (3 lesions for each of 20 scans). Figure 46 shows the average accuracy obtained across the irregular lesions for each method, compared between the three lesions drawn. The contours drawn for tonsil, tongue and parotid lesions returned HD values of 3.2 mm, 5.6 mm and 3.1 mm respectively when compared to a spherical contour of equivalent volume. The method showing the highest robustness (least variation in DSC or RVE) to the type of lesion delineated (tonsil, tongue or parotid) was RG, whereas FCM and GCM were highly affected by the lesion type (up to 90% and 173% difference in RVE between sites respectively). The largest RVE was reached for most methods for the parotid lesion, but the lowest DSC was obtained with the tongue lesion for most methods.

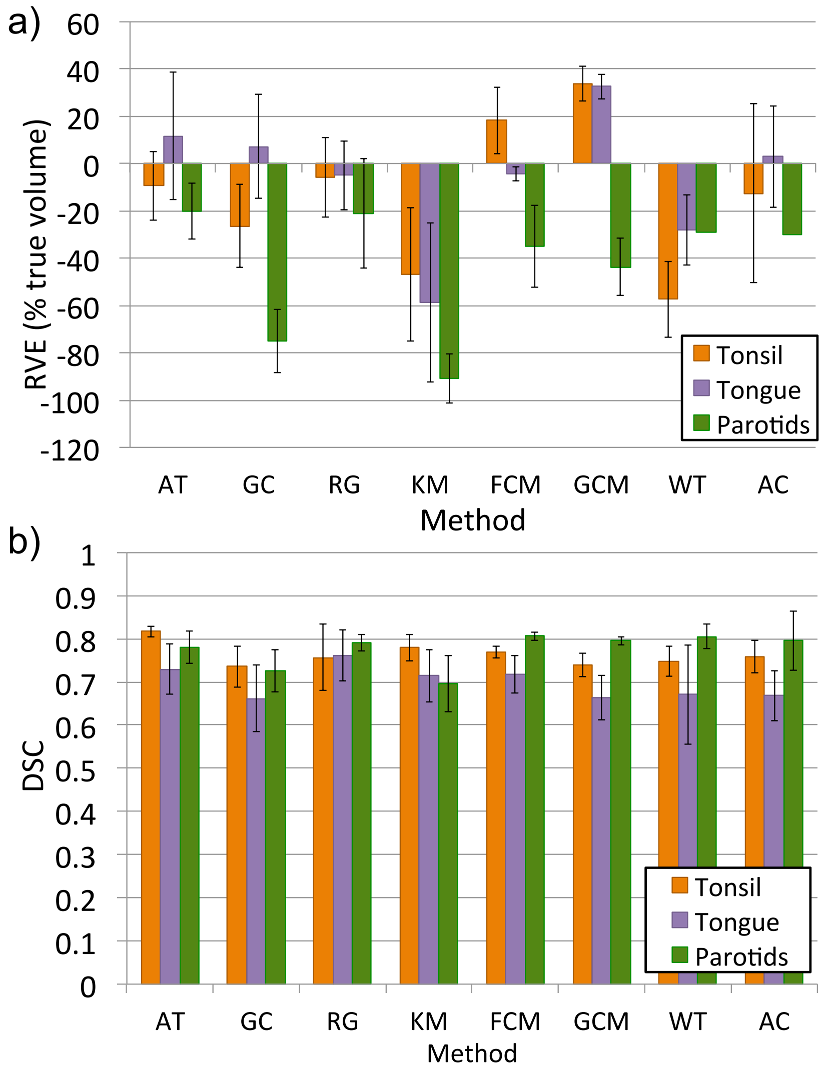


Figure 46. Average a) RVE and b) DSC obtained with the different PET-AS methods for irregular tonsil, tongue and parotid lesions. Error bars correspond to one SD.

Figure 47 and Figure 48 show the RVE and DSC values obtained for regular (spheroids) and irregular lesions respectively, separated for homogeneous, heterogeneous (including Gaussian smoothed and noisy) and necrotic regions (including necrotic and Gaussian necrotic). Higher accuracy and robustness (small error bars) were obtained for spheroidal lesions compared to irregular lesions for all PET-AS methods. AT, RG, GCM and AC reached the highest DSC for homogeneous spheroids (DSC of 0.914, 0.901, 0.905 and 0.904 respectively), with the lowest RVE obtained with AT (RVE=-5.37%). GC showed the best accuracy in delineating heterogeneous spheroids, with a RVE of 1.31% and a DSC of 0.915, followed by AT and AC. Necrotic spheroidal lesions were systematically overestimated by all methods, with the lowest errors achieved by clustering methods KM, FCM and GCM. KM also reached the highest DSC for necrotic lesions (DSC=0.914).

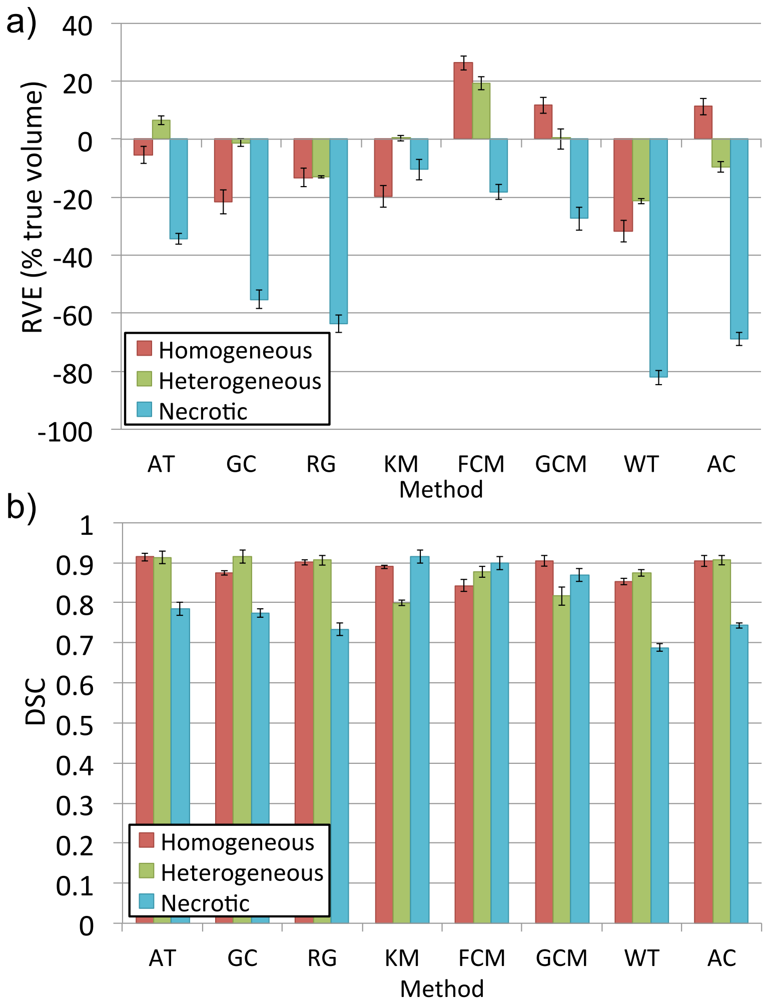


Figure 47. a) Average RVE and b) average DSC obtained with the different PET-AS methods for homogeneous, heterogeneous and necrotic spheroidal lesions. Error bars correspond to one SD.

For irregular lesions, the best accuracy in delineating homogeneous uptake lesions was achieved by RG (highest DSC of 0.821) and AC (lowest RVE of 1.56%). Average RVEs for AT, FCM, GCM and AC did not exceed ±30% of the true volume, and were lower than 13% for FCM. Methods GC, KM and WT appeared most affected by the lesion heterogeneity, with differences in DSC up to 10.6% and 8.61% between homogeneous and heterogeneous or necrotic irregular lesions respectively.

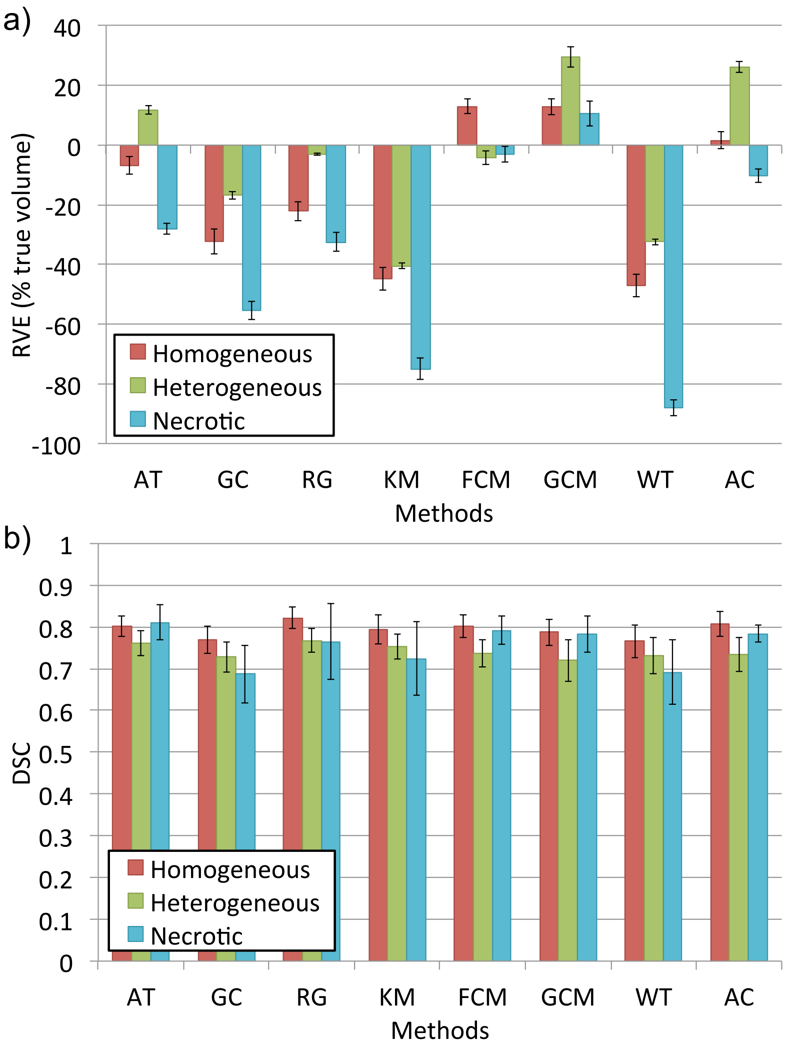


Figure 48. a) Average RVE and b) average DSC obtained with the different PET-AS methods for homogeneous, heterogeneous and necrotic irregular lesions. Error bars correspond to one SD.

The necrotic area inside the spheroidal lesion was excluded from the contour for AT but was included in the final contour for RG and the different versions of FCM, as shown on Figure 49. This was not the case for irregular lesions, which were smaller in volume. The true contours used for evaluation did include the necrotic area. Necrotic spheroidal lesions were systematically overestimated by all methods except FCM2, which achieved the lowest absolute RVE (note that it included the heterogeneous region). FCM2 also reached the highest DSC for necrotic spheroids (DSC=0.887).

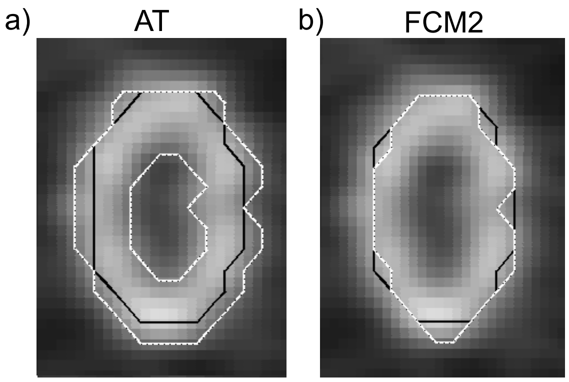


Figure 49. Sagittal slices showing ground truth contour (black) and PET-AS contours (white) for a) AT and b) FCM2 for the necrotic spheroid located in the parotid space.

III. A. 2. d. Discussion and conclusions

These experiments have validated the accuracy of the implemented PET-AS methods on realistic H&N data. The use of the SS printed phantom allowed modelling H&N background and lesion heterogeneity as well as typical lesion location and geometry.

The stratification of the methods' accuracy with the irregular lesion allowed a comparison of the effect on the delineation of the lesion geometrical complexity, and the lesion background. HD values for comparison with an equivalent sphere showed that the tongue lesion corresponded to the most complex geometry. Parotid and tonsil lesion had similar HD values. Average DSC values were lower for the tongue lesion for all methods, except RG and KM (cf. Figure 46). RG reached similarly high DSC values for all three different types of lesions, which shows its robustness to lesion geometry and location. All other methods were affected by the lesion type, in particular gradient-based methods AC and WT, which reached 10% and 12% lower average DSC respectively for the tongue than for the other lesions. In addition, the parotid lesions modelled were surrounded by typical heterogeneous soft tissue uptake, whereas the tongue and tonsil lesions boarded the oral cavity, with no FDG uptake (cf. Figure 45 a)). This is likely to explain the large negative RVEs visible for the parotid lesion in Figure 46, especially for KM and GC, as it was otherwise the same volume as the tonsil lesion.

The results confirmed the high accuracy and robustness of AT seen in Chapter II, even in the case of irregular and heterogeneous lesions. However, the low accuracy obtained by AT for necrotic spheroids, which were larger than the irregular lesions, shows a weakness of the method for delineating highly heterogeneous targets. Figure 49 shows that some methods PET-AS methods such as AT did not include the necrotic area in the delineation, whereas methods such as FCM2 did. This result can be important when delineating heterogeneous tumours, for selecting the methods satisfying the outlining protocol, which should specify if necrotic areas should be included or not. In this work, necrotic areas were included in the GT contours because such tumour regions are included in the GTV for the clinical protocol currently used at Velindre Cancer Centre for H&N RT planning.

Results obtained with ellipsoids confirmed the high accuracy of RG, although the method performed less well for heterogeneous, and particularly necrotic lesions (cf. Figure 47 and Figure 48). RG however still largely outperformed GC and WT, for which a low accuracy was confirmed in this work (negative RVEs larger than 20% for homogeneous lesions). Method GC was discarded in the rest of this project, due to low accuracy and large errors obtained in all the different studies carried out. In addition, the stratified analysis highlighted the strengths of clustering methods KM (for spheroidal lesions), and FCM, GCM and AC for realistic irregular H&N lesions. The high accuracy observed for FCM and GCM on Figure 47 and Figure 48 could be due to an averaging effect of underestimation for tonsil and tongue lesions, and overestimation of the parotid lesion suggested by the RVEs on Figure 46. However, the robustness in DSC values obtained for the different types of heterogeneities (cf. Figure 47 and Figure 48) shows the potential of these methods for realistic H&N lesions.

III. A. 3. Segmentation of heterogeneous lesions

III. A. 3. a. Purpose

Work presented in the previous section highlighted the robustness of clustering methods tested on highly heterogeneous lesions. Structures consisting of a wide range of intensity values can include several regions of different mean intensities. Binary methods classify voxels into two categories corresponding to the tumour and the background. With such methods, tumour voxels of low intensity will therefore be classified as background, even if their intensity level is higher than the mean background value. As a consequence, a reduced sensitivity is expected for binary segmentation methods when delineating highly heterogeneous structures. Clustering methods have the potential of overcoming this by identifying the multiple regions in a heterogeneous structure. This is illustrated on Figure 50. For this study, the following hypotheses were formulated:

- Multiple clustering methods are more accurate than binary methods for the delineation of heterogeneous tumours of large volume.
- The optimal number of clusters for methods KM, FCM and GCM is related to the number of homogeneous uptake regions in the tumour.

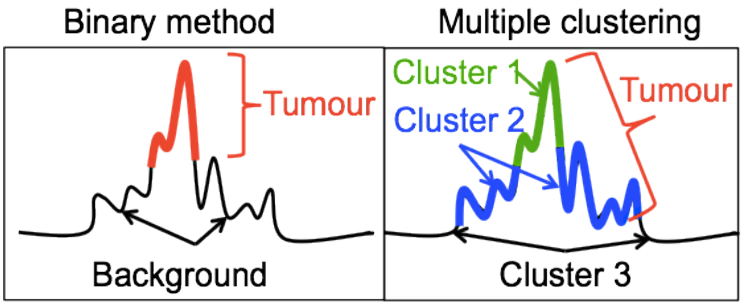


Figure 50. Comparison between binary segmentation and multiple clustering in the case of a heterogeneous tumour with high SUV peak

The work described in the present section aimed at testing the hypotheses listed above, by modelling highly heterogeneous test objects with controlled heterogeneity levels using the SS printed phantom. These experiments focused on evaluating and comparing the accuracy of clustering methods when applied to the detection of different numbers of clusters. This work concludes on the benefits of applying clustering methods to the detection of more than two clusters for heterogeneous target objects, on the influence of the number of clusters on the segmentation accuracy of objects with different levels, on the optimal number of clusters to use, depending on the method and heterogeneity level.

III. A. 3. b. Methods

The six spheres contained in the NEMA phantom (cf. II. A. 1. b), named S10, S13, S17, S22, S28 and S38, were modelled with the following uptake (cf. Figure 51):

- Homogeneous: uniform uptake
- Heterogeneous: 2 homogeneous uptake regions modelled as concentric spheres
- Highly heterogeneous: 4 homogeneous uptake regions modelled as concentric spheres.

The different spheres were modelled with a TBR of 5. A total of 18 images were obtained for the different uptakes modelled for each one of the six NEMA spheres. The reference contour was derived for each sphere by generating a spherical contour from the sphere's known diameter, and positioning this contour on the PET image. Methods KM, FCM and GCM were applied to all spheres modelled for the detection of 2, 3, 4, 5, 6, 7 and 8 clusters. Binary methods AT, RG, WT and AC were applied to all images for comparison. In addition for each clustering method, the accuracy of the different versions applied was evaluated for each sphere size and number of underlying homogeneous uptake regions.

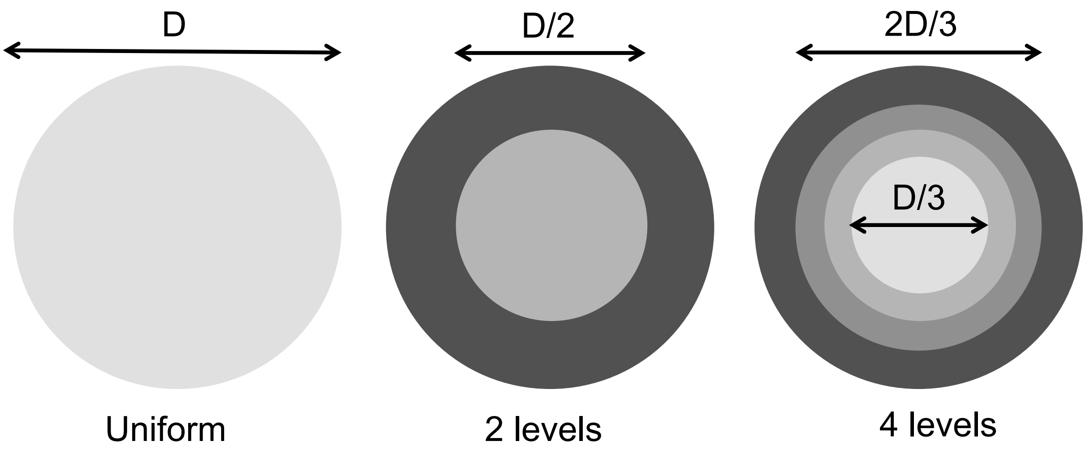


Figure 51. Schematic of the different uptake patterns, which were modelled for each of the six spheres S10-S38.

III. A. 3. c. Results

Table 18 gives the average DSC values obtained across homogeneous spheres, and heterogeneous spheres with two or four homogeneous uptake regions, using binary methods AT, RG, WT and AC, and all clustering methods evaluated in this section. Binary methods AT, RG and clustering method (used as a binary segmentation) GCM2 reached the highest average DSC across homogeneous spheres. For heterogeneous spheres, the best performing methods were KM2, followed by RG, GCM4 and FCM4. For highly heterogeneous spheres, the best performing method was GCM4, followed by KM2 and FCM4.

Uptake regions	AT	RG	WT	AC	KM2	KM3	KM4	KM5	KM6	KM7	KM8	FCM2	FCM3
1	*0.888*	*0.811*	0.785	0.785	0.786	0.574	0.557	0.431	0.373	0.310	0.290	0.503	0.745
2	0.635	0.748	0.684	0.652	0.772	0.630	0.664	0.529	0.460	0.381	0.364	0.223	0.617
4	0.661	0.753	0.690	0.707	*0.804*	0.670	0.529	0.437	0.324	0.276	0.242	0.313	0.658

	FCM4	FCM5	FCM6	FCM7	FCM8	GCM2	GCM3	GCM4	GCM5	GCM6	GCM7	GCM8
1	0.670	0.543	0.435	0.367	0.318	*0.817*	0.744	0.655	0.503	0.407	0.345	0.302
2	0.732	0.730	0.678	0.615	0.557	0.717	0.686	0.744	0.706	0.660	0.592	0.526
4	0.795	0.787	0.723	0.633	0.547	0.766	0.732	*0.809*	0.779	0.691	0.602	0.510

Table 18. Average DSC values obtained for all methods in the case of spheres with 1, 2 and 4 homogeneous uptake regions (values above 0.8 are shown in italic).

Figure 52 a), b) and c) show the DSC values obtained for KM applied to the detection of 2, 3, 4, 5, 6, 7, and 8 clusters for delineating homogeneous spheres, heterogeneous spheres with two regions of homogeneous uptake, and four regions of homogeneous uptake respectively. In the case of homogeneous spheres, KM performed best for all sphere sizes when applied to 2 clusters. For heterogeneous spheres, KM2 reached the highest DSC in all cases except S37 modelled with two homogeneous uptake regions, where KM3 performed better. Although KM2 performed best in most cases, KM versions applied to a higher number of clusters reached DSC values much closer to KM2 for heterogeneous spheres. This effect also increased with sphere size.

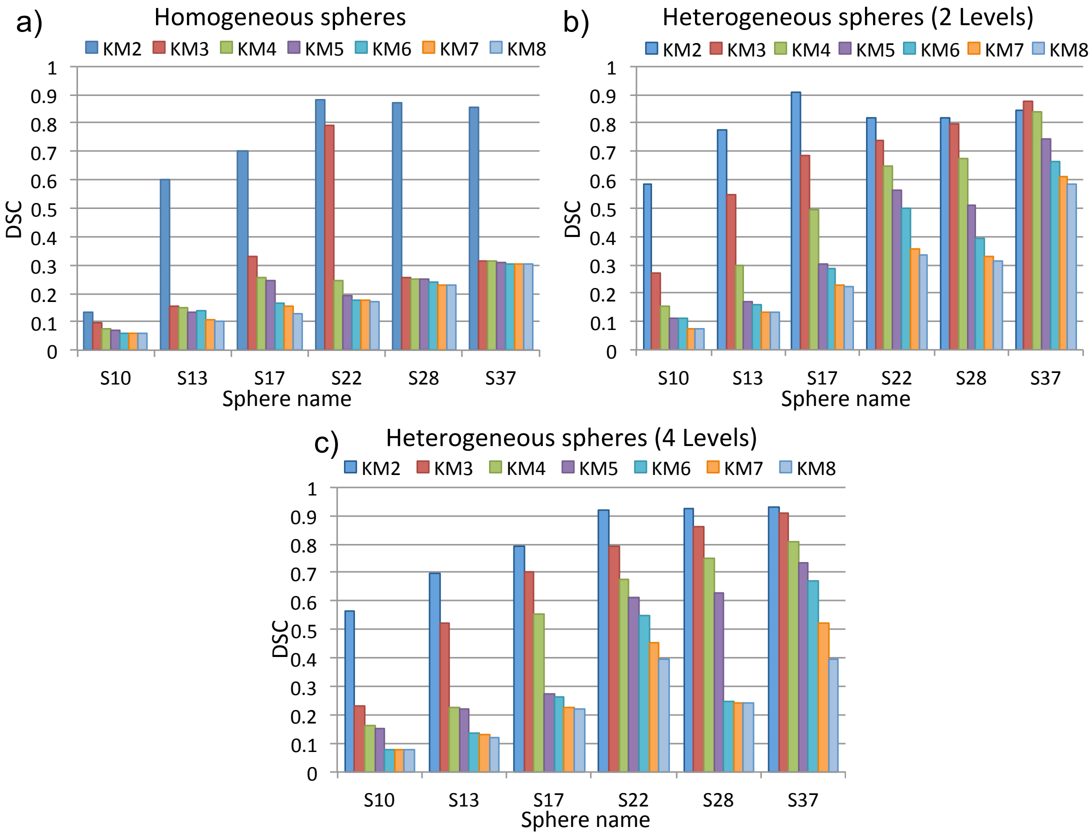


Figure 52. Accuracy of the segmentation by KM applied to 2, 3, 4, 5, 6, 7 and 8 clusters for a) homogeneous spheres, b) heterogeneous spheres with 2 homogeneous uptake regions, and c) heterogeneous spheres with 4 homogeneous uptake regions.

Figure 53 a), b) and c) show the DSC values obtained for FCM in the same situation. For homogeneous spheres, FCM2 and FCM3 perform largely better than versions of FCM applied to a higher number of clusters. However, FCM2 did not perform well for heterogeneous spheres (DSC<0.51). For a heterogeneous uptake of two homogeneous regions, the best performing version of FCM was FCM3 for the two smallest spheres, FCM5 for the two intermediate spheres, FCM6 for S28 and FCM7 for

S37. In the case of a four-region heterogeneous uptake, FCM4 was the best method for S10, and FCM5 for larger spheres.

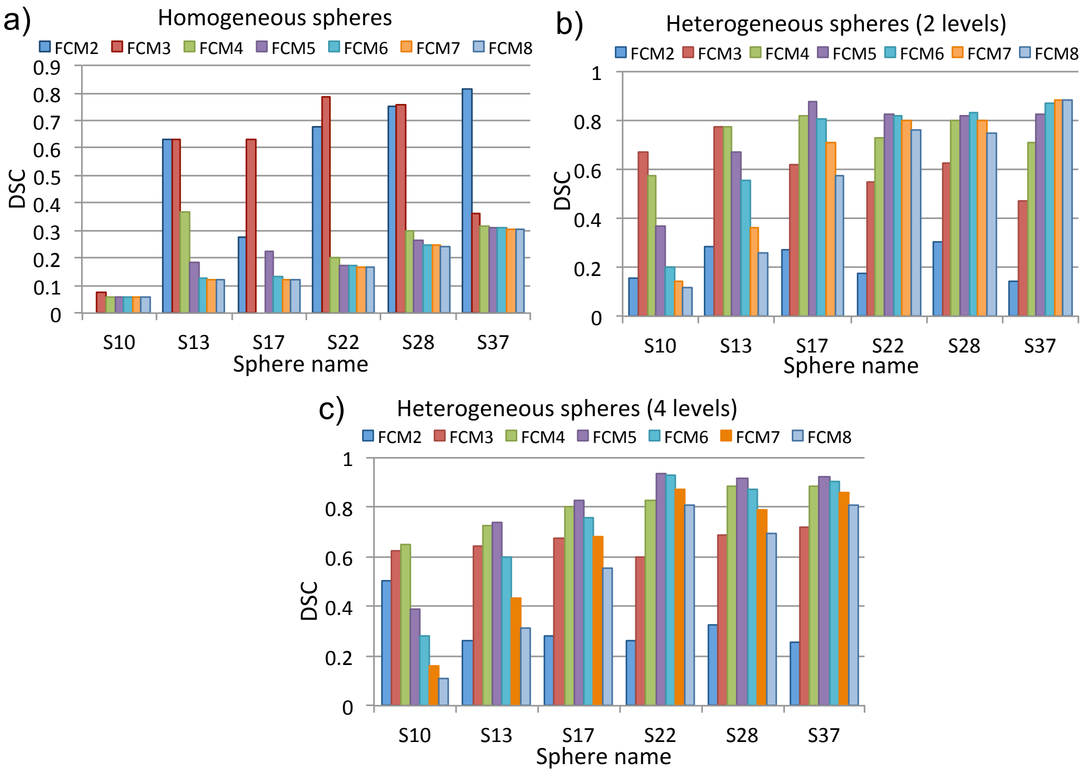


Figure 53. Accuracy of the segmentation by FCM applied to 2, 3, 4, 5, 6, 7 and 8 clusters for a) homogeneous spheres, b) heterogeneous spheres with 2 homogeneous uptake regions, and c) heterogeneous spheres with 4 homogeneous uptake regions.

Results for GCM, shown on Figure 54 a), b) and c) were similar to the ones obtained for FCM. GCM2 and GCM3 were by far the best methods for homogeneous spheres, while the detection of a higher number of clusters provided better accuracy for spheres larger than 13 mm diameter in the case of two homogeneous uptake regions, and spheres larger than 10 mm diameter for a sphere with four homogeneous uptake regions.

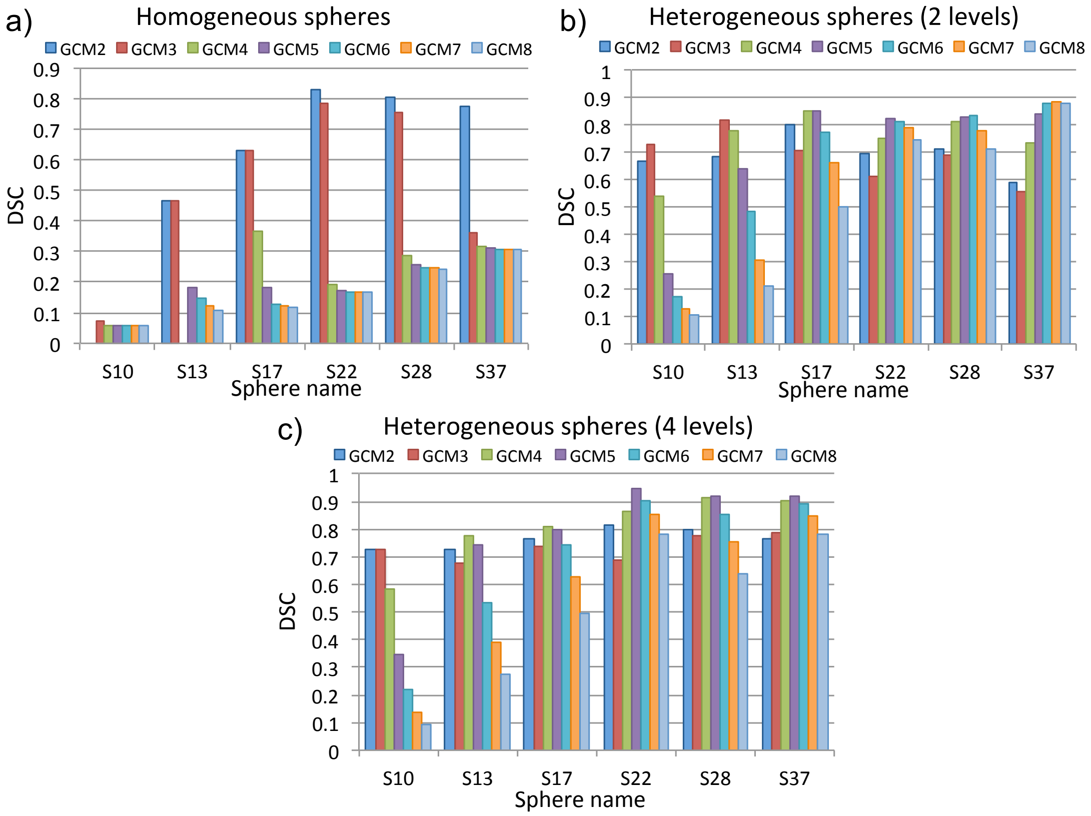


Figure 54. Accuracy of the segmentation by GCM applied to 2, 3, 4, 5, 6, 7 and 8 clusters for a) homogeneous spheres, b) heterogeneous spheres with 2 homogeneous uptake regions, and c) heterogeneous spheres with 4 homogeneous uptake regions.

Figure 55 a) shows one transverse slice of the printed pattern used to model a four-region uptake in a 37 mm diameter sphere. Figure 55 b) shows a transverse slice of the PET image obtained by scanning this same printed pattern within the printed SS phantom, together with the contours provided by the GCM method applied to the detection of 2, 4, 6 and 8 clusters. The volume delineated increased with increasing number of clusters, leading to optimal accuracy when it was closest to the true volume. The best delineation accuracy for the 37 mm diameter sphere (DSC of 0.94) was reached by GCM for 5 clusters. The binary methods AT, RG, and methods FCM and GCM applied to the identification of two clusters, all had DSC values lower than 0.80 in comparison. This is due to the fact that they recovered only the most intense part of the spheres, as illustrated Figure 55 b). KM reached a high DSC when used for 2 (DSC=0.93) and 3 clusters (DSC=0.89).

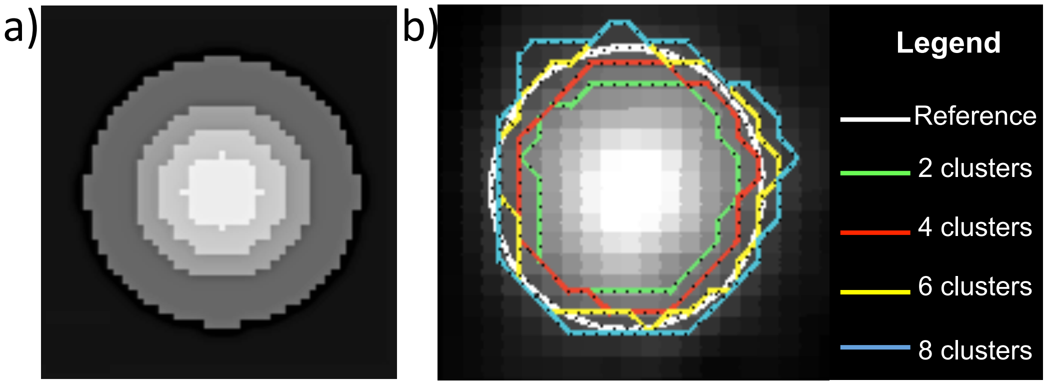


Figure 55. a) printout pattern for a 37 mm diameter sphere with four different homogeneous uptake regions and b) resulting PET image and reference contour (white) and contours from GCM applied to 2, 4, 6 and 8 clusters.

III. A. 3. d. Discussion and conclusions

The investigation of the performance of the different PET-AS on heterogeneous structures is particularly important when evaluating methods aimed for GTV delineation in the challenging case of H&N tumours. Recent work has shown that the impact of the PET-AS method used on the dose distribution delivered to the patient during radiotherapy treatment is particularly important in the case of heterogeneous lesions [148]. This work aimed at evaluating the segmentation accuracy of three different clustering-based PET-AS methods for heterogeneous delineation targets, using to binary segmentation for comparison. For this purpose, the number of clusters used by the clustering methods in the segmentation process was increased from 2 to 8 clusters. The printed SS phantom (cf. III. A. 1) was used to acquire PET images of spheres of different volumes modelled with homogenous, heterogeneous (2 homogeneous uptake regions) and highly heterogeneous intensity distribution (4 homogeneous uptake regions).

Average DSC values calculated across sphere sizes showed that binary methods performed best for homogeneous uptake, but that clustering methods reached higher accuracy for spheres of heterogeneous uptake, especially for spheres modelled with 4 uptake levels (cf. Table 18). RG was the only binary method reaching a DSC higher than 0.71 for heterogeneous spheres, which shows that it is more robust to heterogeneities than AT, WT and AC. However, RG was less accurate than clustering methods for highly heterogeneous spheres. The best performing methods in this case were clustering PET-AS applied to two (for KM) or more (for FCM and GCM) clusters. These results confirmed the necessity of using methods identifying more than two regions in the image, for the delineation of heterogeneous target objects.

The results also suggested that there may be an optimal number of clusters to use for each heterogeneous case. This can be seen on Figure 53 and Figure 54 for FCM and GCM, for which the optimal number of clusters to use increased with heterogeneity and sphere size. This result is intuitive and corresponds well to the hypotheses previously formulated. However, it seems that the number of clusters to use does not necessarily correspond to the number of underlying homogeneous uptake regions in the tumour, nor did it follow a strictly proportional relationship in this work. For example, even for the spheres with the largest number of homogeneous KM2 out-performed all other versions of KM. A thorough investigation of the relationship between the number of homogeneous uptake regions and the optimal number of clusters to use for KM, FCM and GCM would still require further investigation. However, the number of homogeneous uptake regions in a given real tumour cannot be known, and such a study would therefore only provide qualitative indications as to the number of clusters to use.

In comparison to FCM and GCM, the optimal number of clusters for KM was less affected by the number of homogeneous uptake regions in the tumour, as KM2 was the best performing method in most cases. This could be due to the fact that KM uses a binary membership function for each cluster involved (i.e. 1 if the voxel belongs to the cluster, 0 otherwise), whereas FCM and GCM use a continuous membership function, assigning to each voxel a set of values representing the probability of belonging to each cluster, according to intensity distribution criteria.

These results are in line with work published by Hatt *et al.* [138] showing higher delineation accuracy of their clustering-based method FLAB when using it to identify 3 regions instead of 2. However, the results have also shown that the optimal number of clusters to detect may be extended to values higher than 3, which are therefore used in the rest of this work. The use of a higher number of clusters in the segmentation could also improve the geometry recovery of such methods (cf. II. B. 2). For example, thin areas in a target object may yield low intensity values due to PVE, and could therefore be identified by clustering algorithm as lower intensity homogeneous regions, which would only be included in the final contour if a larger number of clusters are detected. Clinical lesions, in particular are likely to be highly heterogeneous with various uptake patterns observed, such as necrotic centres. Work by Belhassen *et al.* [149] uses the Bayesian Information Criterion to determine the optimal number of clusters to use based on the image intensity distribution complexity [72]. The authors also acknowledge that their fuzzy clustering method underestimates the volume of heterogeneous tumours when it is limited to detecting 2 regions. However, in the case of relatively homogeneous regions, binary methods may still outperform multiple clustering schemes. In the rest of this work, the clustering methods are therefore extended to the detection of 2 to 8 clusters, and focused on combining binary PET-AS and multiple clustering approaches in order to provide an optimal segmentation applicable to the variety of lesion types observed clinically.

III. B. Development of ATLAAS: an optimised decision tree segmentation method

III. B. 1. Purpose and description

The number of published and validated advanced PET-AS methods is currently growing, as a number of centres are aiming at implementing such methods into their planning protocol. However, the focus of the literature remains on individual experience of different centres with single PET-AS methods. 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. The previous chapters have addressed this issue by comparing a selection of promising state-of-the art PET-AS methods in a wide range of clinically relevant conditions. The results have highlighted strengths and weaknesses of different approaches in different situations, and shown the necessity to combine the advantages of the most promising PET-AS methods, instead of selecting a single one for segmentation.

A small number of publications in the field of medical image segmentation have suggested the use of machine learning techniques, for continuously improving algorithms, which "learn" from the data they are used on. Machine learning allows a given algorithm to be built and optimised using existing data for which the ground truth is known, in order for it to achieve optimal performance in cases where the ground truth is unknown. Machine learning techniques include methods such as *K Nearest Neighbours* [150], [151], *Support Vector Machine* [152], [153] and *Artificial Neural Networks* [111], [154], 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. In addition, training methods can be continually improved or updated by modifying the training dataset, which implies that a method developed for data acquired in one centre could be easily transferred to a different centre, provided training data is available. Machine learning methods are commonly applied within one test image, to classify the voxels into different categories [155], or for diagnostic purposes [156]. However such methods could also be applied to the classification of test cases into groups for which a particular segmentation algorithm would perform best. In particular, *Decision Tree Learning* is another supervised learning method providing a set of classification rules for the training dataset learned during the training process (the tree). The advantage of such a technique is that it provides a way of implementing these rules into an optimised decision process. *Decision Tree Learning* could therefore be a powerful tool in the exploration of a wide range of data cases representative of clinical tumours, to achieve optimal segmentation in routine clinical practice.

This section describes the implementation of ATLAAS: Automatic Tree-based Learning Algorithm for Advanced image Segmentation, a segmentation framework built using decision tree learning techniques, aimed at achieving optimal segmentation accuracy for H&N PET data.

III. B. 2. Materials and methods

III. B. 2. a. Design of the model

The flowchart for building the ATLAAS method is shown on Figure 56. ATLAAS extracts from the target image (Clinical data on Figure 56) a number of parameters, which are used to select and apply the best among a number of available PET-AS methods. This is done using decision trees calculated from a large training dataset (Training data on Figure 56). For each PET-AS method and for a given set of image parameter values, the decision trees provide the expected accuracy of the method in generating an optimal contour. ATLAAS then applies to the target image the algorithm with the highest expected accuracy.

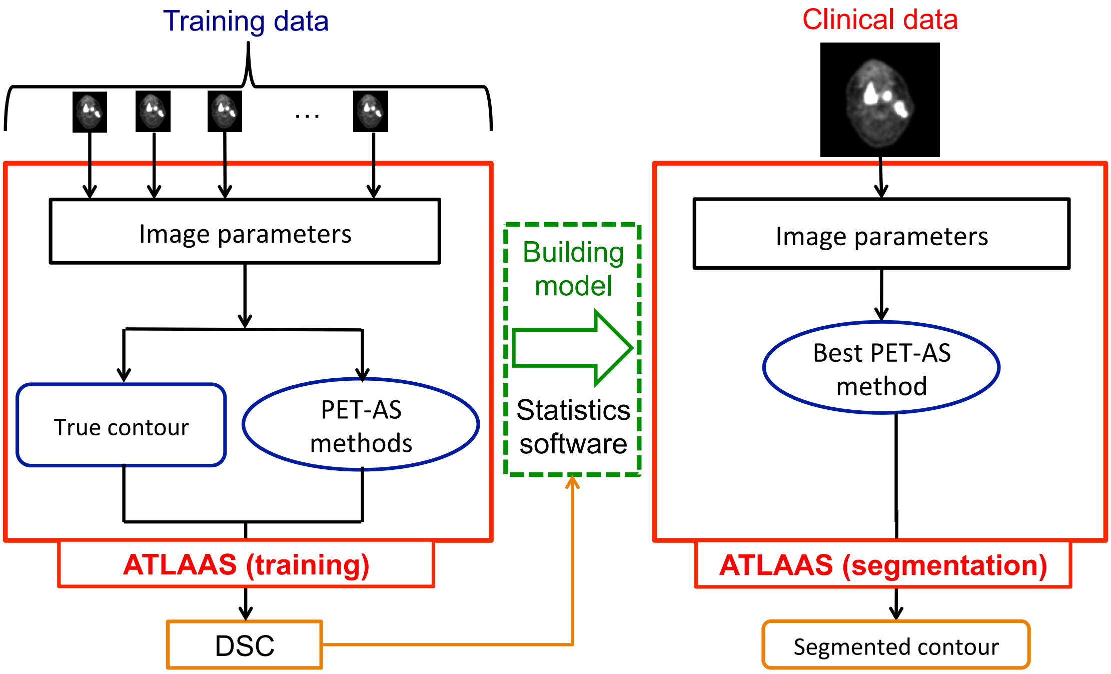


Figure 56. Steps in the training and use of ATLAAS.

The following metrics describing the simulated tumours were identified:

- Vol: tumour volume (mL)
- TBR_{peak}: Ratio between the tumour SUV_{peak}, calculated as the mean value in a 1 cm^3 sphere centred on the maximum SUV in the tumour and the background SUV, calculated as the mean intensity in a 5 mm thick extension of the contour.
- COV: Coefficient of Variation (cf. Eq. 19 in II. A. 3)

Regional texture features were extracted to investigate the influence of the number of intensity levels in the tumour. These metrics rely on the identification of regions of connected voxels with the same intensity value, after resampling the tumour to 64 discrete intensity levels as described by Haralick *et al.* [9]. The following texture metrics were calculated:

- Iv: Intensity Variability:

$$Iv = \sum_j \sum_i \left(n_{(i,j)}\right)^2 \qquad \text{Eq. 24}$$

With $n_{(i,j)}$ the number of voxels in regions of intensity level i and size j.

- Sv: Size-zone Variability

$$Sv = \sum_i \sum_j \left(n_{(i,j)}\right)^2 \qquad \text{Eq. 25}$$

- NI: number of intensity levels
- NR: number of homogeneous regions
- NRS: number of homogeneous region sizes.

III. B. 2. b. Description of the PET simulator tool PETSTEP

A fast simulator tool, developed at the Memorial Sloan Kettering Cancer Centre (New York, USA), and implemented in CERR for this project was used for the generation of the training dataset. The tool is named PETSTEP: a Positron Emission Tomography Simulator of Tracers via Emission Projection. PETSTEP was calibrated for generating PET images equivalent to the ones obtained at Velindre Cancer Centre in Cardiff. The description of the PETSTEP and outputs of this calibration work have been submitted to a peer reviewed journal and were presented at international meetings [157].

III. B. 2. c. Building of the model

Realistic simulated PET images were generated using PETSTEP for training the decision trees to be generated for each method. The FDG uptake map defined for the printed SS phantom in III. A. 1 was used as a simulation template, to model typical background H&N activity levels. A large dataset was generated, by adding tumours of varying sizes and activity distributions to the background FDG uptake map. This was done automatically from initial tumour contours drawn manually on the FDG uptake map, to model different irregular shapes and locations.

The data were aimed at covering the range of tumour metrics observed for clinical H&N data at Velindre Cancer Centre. For this purpose, two types of lesions were identified on available clinical data:

- Primary lesions, located around the oral cavity, tonsils and base of tongue,

- Nodal lesions, located in the parotid and submandibular spaces.

A number Nc=10 of different initial contours was used a for generating the dataset. Five of the ten initial contours corresponded to typical primary lesions locations, and were used to generate synthetic tumours with targeted volumes and intensities ranging within 7-75 mL and maximum SUV values of 5-40, corresponding to the ranges clinically observed for primary tumours. The remaining contours represented nodal lesions and corresponding synthetic tumours were similarly generated with volumes ranging within 0.5-40 mL and maximum SUV values of 2-25. For each initial contour, tumours were modelled for Na=5 uptake values spanning the specified SUV range, and for Nv=5 values of the tumour volume within the specified range. For each tumour geometry, uptake value and volume, Nt=4 different uptake textures were simulated:

- homogeneous uptake,
- 2 homogeneous regions with the highest uptake level in the centre,
- 3 homogeneous regions with the highest uptake level in the centre,
- necrotic uptake in the tumour centre.

A total of 1000 lesions were modelled ($Nc * Na * Nv * Nt = 1000$). Global tumour uptake noise was modelled in each modelled tumour by randomly assigning to the voxels uptake values extracted from a Gaussian distribution centred on the targeted mean uptake value, with SD of 100% of the mean.

The statistical analysis software SPSS (cf. II. A. 1. c) was used to derive decision trees for the selection of the optimal segmentation approach among AT, RG, AC and WT and clustering methods KM, applied to the detection of 2 or 3 clusters, and FCM and GCM applied to the detection of 2 to 8 clusters. For each segmentation approach, a tree was grown for the prediction of the DSC score obtained on different tumour types. The Classification and Regression Tree (CRT) growing method [158] was used with all tumour and image characteristics entered into the model as prediction variables. The impurity measure “Gini” was used to ensure homogeneity of the cases classified into the same groups. The maximum tree depth was set to 10 with a minimum number of 50 cases per node, as a good trade-off between tree accuracy and tree growth. The software returned the trees obtained for each segmentation approach, together with the risk estimate RE, which represents the average error made on the whole dataset if assigning each case to the DSC value predicted by the tree, rather than the actual DSC obtained. The importance to the model was also calculated for each variable, as the improvements brought to the model when using the given variable (summed across all steps of the classification tree). It quantifies the amount by which the use of a given predictive variable (image metric) in the tree reduces the error in prediction of the outcome (the DSC). The trees were pruned back to the smallest size possible for a minimum classification risk allowed of 1 SD of the average risk, to avoid over-fitting the model to the training dataset. ATLAAS was finally built using the decision trees derived for each PET-AS method. It calculates for each new case the image parameters using a single estimate of the tumour contours, followed by the predicted DSC score for each approach based on these parameters. The method then selects the algorithm reaching the highest predicted DSC value to be applied.

III. B. 2. d. Evaluation with simulated data

A first validation dataset was generated by simulating 100 new cases of tumour in H&N background. Each case was generated with volume value, maximum intensity value and uptake pattern randomly chosen from the range of values obtained in the training dataset. Only the images for which these fell into the range used in the training dataset were kept. The ATLAAS model built as described in the previous paragraph was applied to each test case of this validation dataset, on the basis of image parameters calculated from an initial estimate of the contour. The contour estimate was obtained using KM2, selected as the PET-AS with the lowest SD of DSC values across the training dataset, so as to provide a robust estimate of the image parameters. For comparison purposes, ATLAAS was also built using parameters extracted from the true (GT) contour. This model is named ATLAAS_GT in the following sections.

For the evaluation of ATLAAS, the mean, median and minimum returned by the method were calculated across the validation dataset. In addition, the percentage of cases where ATLAAS returned the best DSC (Pbest) and a DSC value within 10% of the best (P10), were determined and compared to the single PET-AS methods.

In addition, the distribution of DSC values obtained for ATLAAS was compared to the single PET-AS methods using the Mann Whitney U-test. The test was also used to compare the results from ATLAAS for the array of DSC values corresponding to the best DSC achieved across the single PET-AS methods for each case. These values correspond to the highest segmentation accuracy achievable by a model such as ATLAAS using the same single PET-AS methods. It is named HS in the rest of this work.

III. B. 2. e. Evaluation with phantom data

A total of 115 cases were used for the validation of ATLAAS, including:

- 58 H&N homogeneous and heterogeneous (random, Gaussian smoothed and necrotic uptakes) cases generated with the subresolution sandwich printed phantom,
- 39 fillable phantom cases using thin-wall spherical and non-spherical plastic inserts,
- 18 cases of heterogeneous spheres (1, 2 or 4 homogeneous regions of different uptakes) obtained with the printed subresolution sandwich phantom.

These validation cases corresponded to the ones obtained for testing the different PET-AS methods, as described in II. A. 3, II. B. and III. A. 2., excluding spherical

NEMA inserts, and cases for which TBR$_{peak}$ or volume values fell outside the range of values used for training ATLAAS.

The ATLAAS method was applied and evaluated similarly as for the simulated validation dataset.

III. B. 3. Results

III. B. 3. a. Building of the model

Simulated training cases for which the TBR$_{peak}$ or Vol values were outside the observed clinical range were discarded. In particular, low TBRs were sometimes achieved due to the proximity of intense normal uptake (tonsils, brain, etc.), which hampers the delineation process. Overall, a total of 845 cases were used to build the model. Table 19 presents the range of values obtained for the different metrics considered, for both simulated and clinical images. The parameter values simulated closely matched the clinical values observed, with minor differences in texture feature values (e.g. lower values for Iv in the simulated cases). Figure 57 shows the coverage of the range of clinically observed TBR$_{peak}$ and Vol values achieved by the training dataset. Small gaps observable between cases of similar volume values are due to the fact that the data were generated by choosing evenly spaced values for the maximum tumour intensity and volume, spanning the range of values targeted.

Parameter	TBR$_{peak}$	Vol (mL)	COV	Iv	Sv	NI	NR	NRS
Clinical	1.1 – 8.4	0.44-67	10-51	153-12x10^{5}	19-72x10^{3}	13-63	15-1756	2-344
Simulated	1.1-9.7	0.62-76	9.7-53	194-79x10^{4}	19-47x10^{3}	17-63	18-1304	1-402

Table 19. Comparison of the range of values measured for the parameters considered for clinical, simulated and validation phantom data.

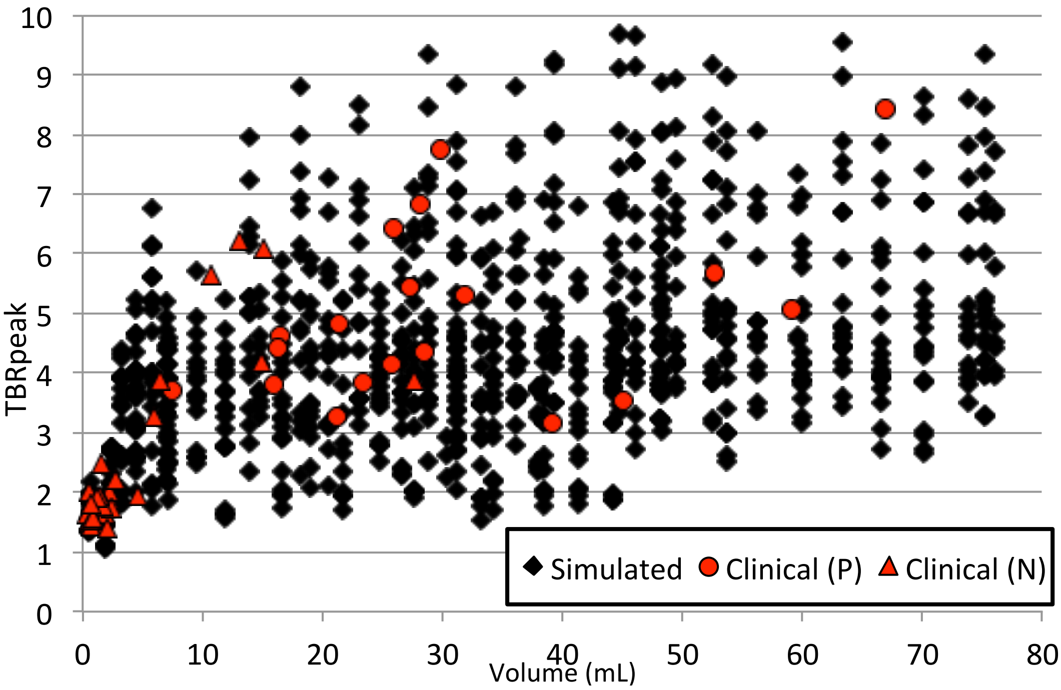


Figure 57. Range of TBR_{peak} and volume values obtained for simulated cases, compared to clinical primary (P) and lymph node (N) cases recorded at Velindre Cancer Centre.

Methods FCM and GCM, applied to 6 or more clusters, did not reach the highest DSC value for any of the training cases, and were therefore discarded in the rest of this study. A total of 14 trees were generated using all parameters described previously as classifiers, corresponding to predictions for the delineation accuracy of AT, RG, AC, WT, KM2, KM3, FCM2, FCM3, FCM4, FCM5, GCM2, GCM3, GCM4 and GCM5. Table 20 provides for each one of the methods the mean DSC obtained on the training dataset, the percentage of cases for which it returned the highest DSC (Pbest), the risk estimate of the tree obtained, and the normalised importance of each one of the variables included in the tree. All trees were built with an estimated classification risk of less than 5%, demonstrating good classification accuracy. TBR_{peak} was included in all of the decision trees generated but one, and was the most important variable to the model for 10 out of 14 methods. COV was also included in 10 out of 14 trees. Sv was included in none of the trees generated, which shows that no classification of the data allowed predicting the DSC of the PET-AS methods with enough accuracy to be included in the tree building process. For this reason, Sv was excluded from further analysis.

III. B. 3. b. Evaluation with simulated data

The 100 simulated images generated for the model validation had $\mathrm{TBR}_{\mathrm{peak}}$ and Vol values ranging within 1.3-9.3 and 0.6-68 mL respectively. The results of the validation with the phantom data are given in Table 21 for the validation dataset, showing the mean, median and lowest DSC obtained on the dataset as well as the percentage of cases for which the method achieved the best DSC across all methods tested (Pbest), or a DSC within 10% of the best DSC value (P10). These results are compared to the results obtained for PET-AS methods AT, RG, KM2, and WT, which were the only methods achieving mean or median values higher than ATLAAS across the different datasets. For the simulated validation dataset, ATLAAS_GT reached the same mean DSC and P10 value as the best performing PET-AS method on this dataset (KM2), with a slightly higher minimum DSC value. Slightly lower values were obtained for ATLAAS but the results were still very close to the best performing method. ATLAAS returned exactly the best DSC in 7% cases, and a DSC within 10% of the best in 89% cases. The largest error (difference to the best DSC obtained for single PET-AS methods) was 21%. These results are illustrated on Figure 58 a), which shows the DSC values obtained for ATLAAS (in black) and the selected single segmentation approaches for each case of the validation dataset. It can be noted that the black curve corresponding to ATLAAS covers the highest DSC values in most cases.

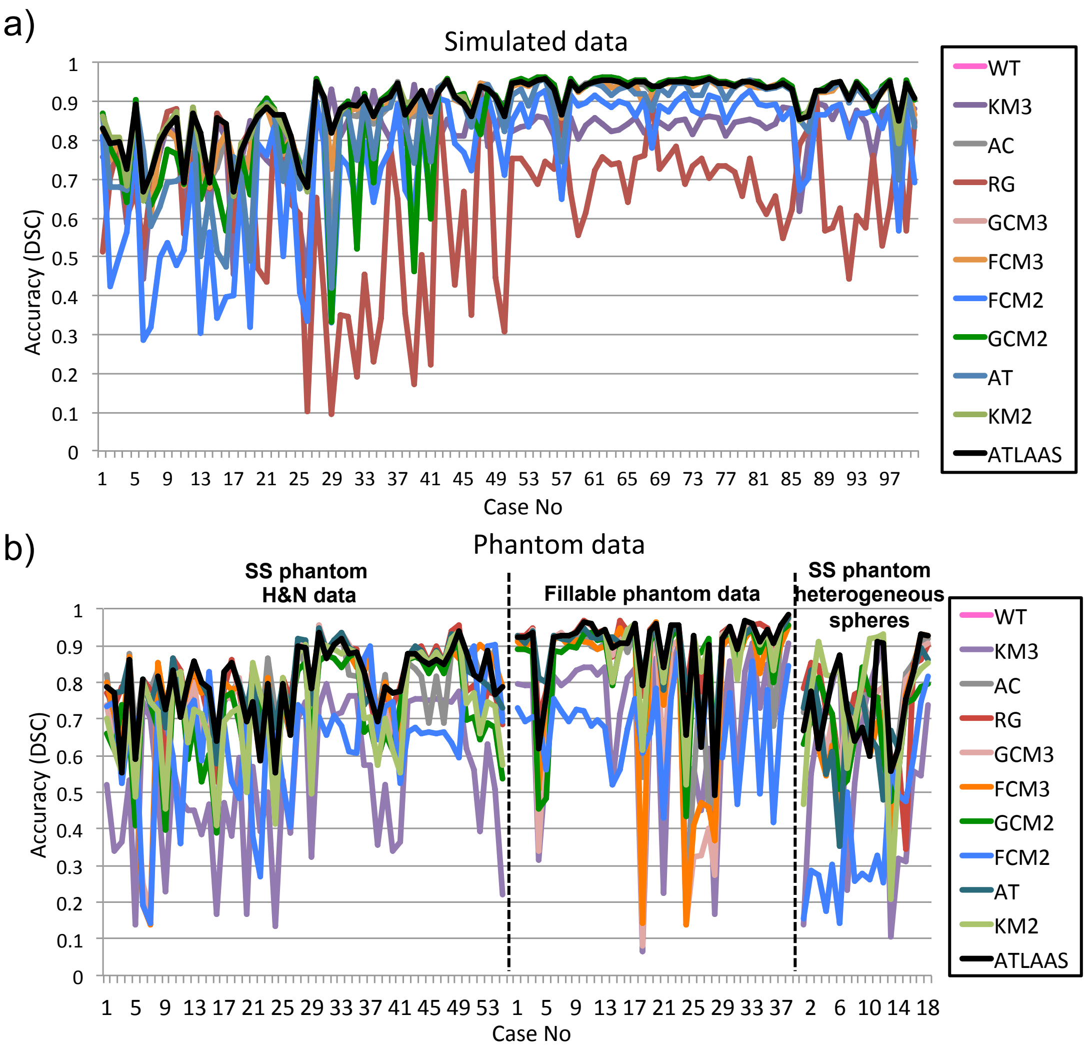


Figure 58. Accuracy of ATLAAS (black curve) compared to the single segmentation approaches over a) the simulated and b) the phantom validation datasets.

PET-AS Method	AT	RG	WT	AC	GCM2	GCM3	GCM4	GCM5	KM2	KM3	FCM2	FCM3	FCM4	FCM5
Mean DSC	0.752	0.724	0.722	0.805	0.826	0.823	0.782	0.732	0.863	0.784	0.676	0.824	0.793	0.744
Pbest (%)	7.6	5.2	21.4	3.2	2.8	1.9	1.2	<1	18.2	18.3	0.2	0.0	1.3	<1
RE	0.018	0.049	0.031	0.01	0.004	0.005	0.008	0.007	0.007	0.008	0.017	0.005	0.007	0.007
							Relative importance (%)							
TBR_{peak}	71	97	100	85	100	100	100	100	n.i.	100	100	100	100	100
Vol	25	44	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	100	n.i.	12	n.i.	n.i.	n.i.
COV	100	84	77	100	86	46	n.i.	24	n.i.	n.i.	100	52	n.i.	21
Iv	37	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	55	n.i.	n.i.	n.i.	n.i.
Sv	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.
NR	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.	24	n.i.	ni.	n.i.	n.i.	n.i.	n.i.	24
NRS	n.i.	100	n.i.	n.i.	n.i.	17	n.i.	n.i.	77	n.i.	24	17	n.i.	n.i.
NI	72	n.i.	99	ni	ni	n.i.	n.i.	ni.	n.i.	ni.	n.i.	n.i.	ni.	n.i.

n.i.: not included in model

Table 20. Description of the decision trees training using all image parameters for the different PET-AS methods selected (with mean DSC, Pbest: % of cases where the method returned the highest DSC, and risk estimate RE) and importance of the image parameters to the tree models.

	HS	AT	RG	KM2	WT	ATLAAS_GT	ATLAAS
Mean DSC	0.906	0.847	0.640	0.890	0.862	0.890	0.878
Median DSC	0.932	0.905	0.672	0.915	0.921	0.915	0.915
Min DSC	0.716	0.419	0.095	0.643	0.332	0.667	0.667
Pbest (% cases)	100	3.0	5.0	13	51	11	7.0
P10 (% cases)	100	80	22	96	89	96	89

Table 21. Results of the evaluation of ATLAAS_GT and ATLAAS with simulated data, including mean, median and minimum DSC obtained, Pbest and P10 values (cf. III. B. 3. b). Segmentation results are given for AT, RG, KM2, WT and the highest achievable segmentation accuracy HS for comparison.

III. B. 3. c. Evaluation with phantom data

The results of the evaluation of ATLAAS with phantom data are given in Table 22 for the H&N printed phantom, fillable phantom, and heterogeneous spheres printed phantom separately, and for the phantom data overall. The results show that the different single PET-AS methods performed differently on the three types of phantom data used: AT was the best method for H&N subresolution sandwich phantom data, RG for fillable phantom data, whereas KM2 reached the highest mean DSC for the heterogeneous spheres. Overall, ATLAAS reached mean and median DSC higher than any of these methods, as well as a higher minimum DSC value (when built using KM2). ATLAAS provided the best DSC value in 20% cases, and DSCs within 10% of the best value in 77% cases. ATLAAS reached slightly lower DSC values than the best single PET-AS method for H&N printed and fillable phantom data (AT and RG respectively), but still performed better than all other methods. For the highly heterogeneous spheres generated with the printed phantom, it was again more accurate on average than the best single PET-AS (KM2). The largest error (difference to the best DSC obtained for single PET-AS methods) was 32%, and only 5 cases reached errors higher than 20%. Figure 58 b) shows the DSC values obtained for ATLAAS (in black) and the selected single segmentation approaches for each case of the three phantom datasets separately. The curve for ATLAAS is again covering the highest DSC values obtained by single PET-AS for most cases.

		HS	AT	RG	KM2	WT	ATLAAS_GT	ATLAAS
H&N SS printed	Mean DSC	0.844	0.806	0.787	0.731	0.716	0.788	0.799
	Median	0.854	0.799	0.784	0.730	0.723	0.814	0.809
	Min	0.698	0.579	0.530	0.411	0.389	0.137	0.552
	Pbest (% cases)	100	17	22	6.9	1.7	14	14
	P10 (% cases)	100	79.3	65.5	39.7	41.4	72	47
Fillable phantom	Mean DSC	0.917	0.891	0.901	0.879	0.851	0.897	0.889
	Median	0.918	0.921	0.938	0.929	0.896	0.929	0.927
	Min	0.640	0.640	0.500	0.492	0.436	0.492	0.492
	Pbest (% cases)	100	17.9	30.8	30.8	5.1	33	33
	P10 (% cases)	100	89.7	89.7	84.6	79.5	92	87
Heterogeneous spheres	Mean	0.842	0.686	0.728	0.765	0.690	0.771	0.758
	Median	0.851	0.714	0.760	0.817	0.743	0.775	0.756
	Min	0.642	0.353	0.345	0.207	0.476	0.556	0.556
	Pbest (% cases)	100	3.4	1.7	5.2	1.7	3.5	1.7
	P10 (% cases)	100	17.2	15.5	20.7	5.2	19	19
Overall phantom data	Mean	0.869	0.817	0.817	0.787	0.758	0.823	0.824
	Median	0.891	0.824	0.840	0.824	0.789	0.854	0.851
	Min	0.640	0.353	0.345	0.207	0.389	0.137	0.492
	Pbest (% cases)	100	16.5	22.6	16.5	3.5	20	19
	P10 (% cases)	100	23.5	29.6	24.3	5.2	77	77

Table 22. Mean, median and minimum DSC, Pbest and P10 values (cf. III. B. 3. b) obtained by ATLAAS_GT and ATLAAS on the phantom datasets. Results are given for AT, RG, KM2 and WT for comparison, and for the highest achievable accuracy segmentation HS.

The Mann Whitney U-test comparing the distribution of DSC values obtained for ATLAAS and the single PET-AS methods showed significant differences ($p<0.05$) with all methods except WT on the simulated dataset, and with WT and AC on the phantom dataset. When the values obtained for ATLAAS were compared to the highest DSC value obtained for each case, the differences were not significant.

III. B. 4. Discussion and conclusions

The aim of this study was to develop a model able to select and apply the best performing method among a selection of promising segmentation algorithms. The combination of several algorithms has been investigated in other studies using other combination processes. McGurk *et al.* [108] compared two voxel-wise methods for the combination of 5 different segmentation algorithms, in order to limit inconsistencies of the segmentation accuracy on a large dataset. In the present work, a different approach was used with the aim of optimising the accuracy of the PET-AS segmentation by considering only the predicted best performing method. This optimisation approach relies on the training of a prediction model to apply the best PET-AS method out of a selected range. This differs from the approach of McGurk *et al.* in that it does not include the results of all PET-AS methods in deriving the final contour, but only uses the results of the best among the PET-AS methods used. This work aimed at providing the highest possible accuracy, whereas the method from McGurk *et al.* is focused on avoiding large errors observed for some methods in specific cases. The final method, ATLAAS, showed excellent accuracy across a wide range of simulated and phantom data (mean DSC of 0.878 and 0.824 respectively), and achieved a prediction of the best or near-best segmentation (DSC within 10% of the best) in a large number of cases (96% and 77% for simulated and phantom data respectively), as shown in Table 22. These results, together with the data shown on Figure 58, also demonstrate the robustness of ATLAAS, which systematically returned DSC values higher than 0.67 and 0.49 on the simulated and phantom validation datasets respectively. The mean and median DSC values obtained for ATLAAS were within 7% of the hypothetical values achieved if the model was perfect, corresponding to the highest segmentation accuracy among single PET-AS methods for each case (HS), for both simulated and phantom data (cf. Table 22). In addition, the results of the Mann-Whitney U-test showed that the distribution of values obtained for ATLAAS was not significantly different from the distribution of HS values.

In this work, a full protocol was also developed for the building and application of the ATLAAS method, including the automatic generation of a large dataset representative of the clinical situation targeted. This was made possible by using PETSTEP, a fast and flexible PET simulation tool, and building an automatic process for the generation of images covering a wide range of image parameters. However, since the quality of the model depends largely on the quality of training dataset, it is expected that any improvement to the training image quality, such as with future releases of PETSTEP, would further increase the accuracy of the ATLAAS model.

The data obtained in the evaluation of PETSTEP also suggests some possible improvements for the method. The simulated validation dataset showed a very similar accuracy for KM2 and ATLAAS, both methods reaching values very close to the HS situation of an ideal model (cf. Table 22). This is due to the very high accuracy of KM2 for the simulated validation dataset, as the PET-AS achieved DSCs within 10% of the best in 96% of the cases. For the phantom dataset, KM2 showed lower accuracy than AT and RG on the H&N phantom and fillable phantom datasets, corresponding to cases with largely irregular and relatively small tumours. This suggests that although the methodology used aimed at generating a training dataset representative of observable clinical cases, the introduction of several extreme cases, more accurately delineated by a method other than KM2, could improve the building and evaluation of the ATLAAS model.

The comparison of ATLAAS to ATLAAS_GT shows little difference in the accuracy, which is a sign of a low sensitivity of the method to the initial contour estimate. However, the use of KM2 for estimating the tumour parameters may bias the model toward selecting and applying KM2. Additional work could therefore investigate the use of a different method for estimating the tumour parameters, or preferably, focus on tumour parameters little influenced by the exact contour estimate. Future work could also aim at assessing the importance of different aspects of the tumour on the model accuracy. For example, additional parameters describing the tumour geometry could be included, such as the shape indices used by Tixier *et al.* for predicting response to therapy [159]. A thorough study determining the metrics most adequate for the extraction of such parameters would also greatly benefit the development of ATLAAS, and the number of parameters describing the tumour heterogeneity may be reduced to maintain a better balance between the different tumour characteristics. However, such analysis would require a project on its own, and it was not the aim of this present study.

The robustness observed for ATLAAS across phantom datasets (cf. Table 22) is also due to the fact that the single PET-AS performed differently on the three phantom datasets used (cf. Table 22), suggesting that images from these datasets are different. This is likely to be due to the difference in image quality between simulated images and printed subresolution sandwich phantom data, including differences in the image scatter and photon statistics. The results showed high accuracy of ATLAAS for all of these different datasets, which shows potential robustness of the method to data from different sites.

This work describes the development of ATLAAS, a model for automatic selection and application of a range of PET-AS method, providing high segmentation accuracy for realistic H&N data and challenging phantom images. ATLAAS was validated on a selection of the highest quality phantom data currently available the Wales Research & Diagnostic PET Imaging Centre. This work showed that it presents many advantages on the use of a single PET-AS algorithm in terms of accuracy, reliability and robustness to both tumour and data type. In addition, a full framework for the automatic training and building of such an optimised segmentation method was applied, which could be applied to data from any centre, positron emission imaging system and selection of automatic segmentation algorithms. Throughout this work the future work necessary for improving the method was also identified, such as the use of a higher quality PET simulator, or the inclusion of more complex cases in the training dataset. The use of ATLAAS is expected to be highly beneficial in the RT planning process, as it provides rapid, reliable and accurate GTV segmentation. Future work will therefore also aim at implementing the method in routine clinical practice at Velindre Cancer Centre, and making it available as a package for other centres.

Chapter IV. Application of the PET-AS to clinical data

IV. A. Development of a clinical protocol

IV. A. 1. Purpose

POSITIVE was designed as a pilot study for the implementation of PET-AS methods within the routine clinical RT process at Velindre Cancer Centre, and with the aim of recruiting 20 H&N cancer patients for evaluating the methods in clinical conditions. The involvement of patients in the study set the requirement for a well-defined clinical protocol, allowing efficient and accurate patient scanning and segmentation, while remaining within the time constraints due to the treatment timeline. The definition of such a protocol was also required as a key step to implementing the use of PET into the RT process at Velindre Cancer Centre, for further studies involving PET-AS. The recruitment of patients started after a clinical protocol was finalised together with the clinicians involved in the study. This section describes the clinical protocol derived for POSITIVE, and followed for the studies presented in IV. C and IV. D.

IV. A. 2. Definition of the patient population

The aims of the study included implementation of a PET auto segmentation protocol for routine delineation of H&N GTVs for patients at Velindre Cancer Centre. A single subsite was identified a within H&N cancers, so as to limit the study population and maintain homogeneity within the study. The oropharynx was chosen, as it is the most common subsite for H&N cancer at Velindre Cancer Centre. In order to recruit a consistent patient population, it was also agreed to focus the study on patients undergoing non-surgical therapy, with neo adjuvant chemotherapy followed by concurrent chemotherapy with a curative intent. The patients received 66 Gy in 30 fractions delivered over six weeks by IMRT with concurrent chemotherapy.

The study protocol was accepted by the National Health Service (NHS) Research Ethics Committee and given the number 12/WA/0083. It is part of the UK Cancer Research Network portfolio database, under number 13769.

IV. A. 3. Image acquisition

Since the study was open to patients undergoing neoadjuvant chemotherapy prior to the RT treatment, the planning PET/CT scan was acquired just before the start of chemotherapy. This was done to avoid planning the patients with tumour volumes changed by the chemotherapy, so as to target the whole initial tumour burden. The planning scan was performed at PETIC following patient consent, with an immobilisation shell to maintain a reproducible position during radiotherapy. Time to start of treatment after the scan ranged from 6 to 9 weeks. Therefore the immobilisation shell was checked before starting the treatment, to make sure that it still provided a good fit to the patient contour. If this was not the case, e.g. following weight loss or weight gain due to chemotherapy, the mask was refitted, and the patient was re-scanned and re planned if necessary. All patients were scanned using the scanner and settings described in II. A. 1. a) using 6-8 bed positions of 3 min each, ranging from the top of the head to the sternum. All CT scans were resampled to 2.5 mm slice and interval thickness for the RT planning.

The scanning protocol was defined as follows:

- 90 minutes uptake for the patient, after injection in the back of the hand preferably,
- routine whole body CT scan with a noise index of 16 and 700 mm FOV. This scan was resampled to 3.75 mm slice thickness to be used for attenuation correction of the PET data.

- Contrast Enhanced CT (CECT) with Niopam 300 (75 mL at 2 mL/s, from Bracco UK limited, Bucks, UK) followed by a Saline flush (10 mL at 2 mL/s) for 2-3 bed positions to include the whole of the head and neck.
- Whole body PET scan, 3 min per bed position
- CECT resampled to 2.5 mm slice and interval thickness.

Reporting was done by trained radiologists following the planning scan.

IV. A. 4. Data transfer

The following series were sent in DICOM[1] format to VCC via the Picture Archiving and Communication System (PACS), together with the radiologist report:

- Whole Body CT resampled to 2.5 mm slice and interval thickness
- CECT resampled to 2.5 mm slice and interval thickness, used by the clinicians for planning
- Whole body PET scan

The series were then retrieved from the PACS for use on workstations in VCC, including VelocityAI (version 2.7, Velocity Medical Solutions, Atlanta, USA) and ProSoma (Version 3.1, MedCom GmbH, Darmstadt, Germany).

In addition, all series acquired were anonymised via the PET Xeleris (version 3.0, GE Healthcare, Milwaukee, USA) workstation linked to the scanner and copied onto a DVD for use on the Matlab research workstation.

The data transfer was verified before the start of the project by comparing the SUV within defined regions of routine scan of the NEMA body phantom on the different workstations used. This was done to ensure that the DICOM information was not affected by the transfer, and check that any difference in SUV values calculated at the two centres (due to different software settings) did not exceed 10%.

[1] http://medical.nema.org/standard.html

The data transfer workflow is shown on Figure 59.

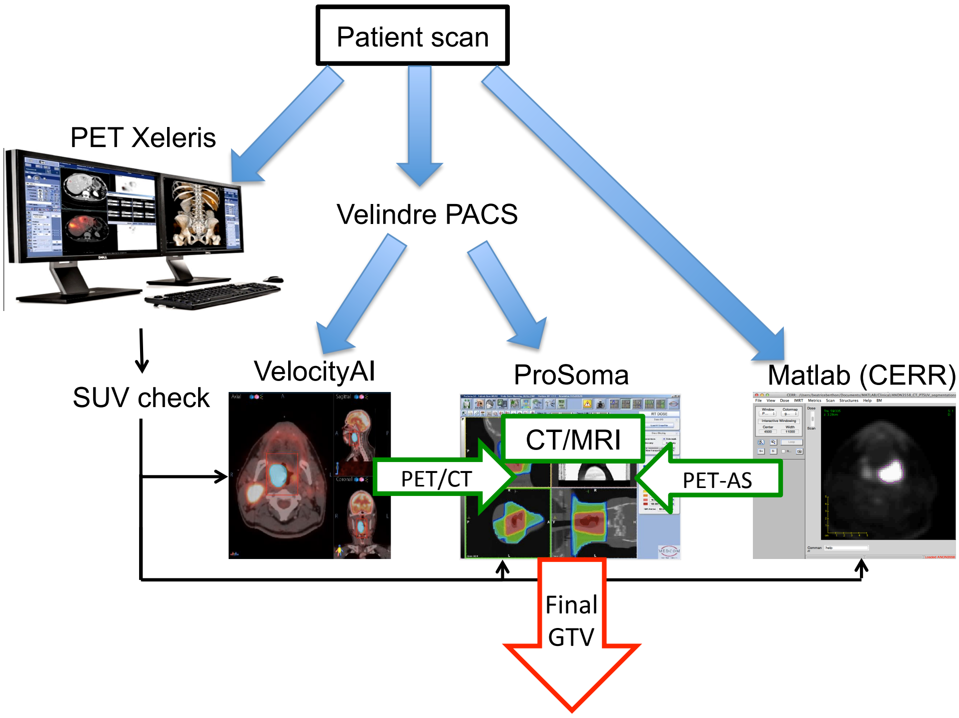


Figure 59. Workflow for the transfer of data across centres and workstations.

IV. A. 5. Outlining of GTV

GTVs were outlined by three experienced clinicians prior to starting the radiotherapy. Imaging data included the latest available diagnostic MRI (axial T1-weighted post contrast) scan for the patient, which was acquired at the referring hospital (data from various institutions), the resampled contrast CT acquired in planning position at PETIC, and the registered planning PET. For the first 10 patients recruited, the following volumes were outlined in this order:

i) Primary GTV (GTV_p), nodal GTVs (GTV_n) and OAR volumes were manually outlined on fused CT/MRI, registered with Mutual Information using ProSoma

ii) A GTV_p and nodal GTV_n were manually outlined on registered planning contrast CT and planning PET in VelocityAI. PET/CT manual delineation was done using a fixed window level of 6, recommended by expert radiologists at Velindre Cancer Centre.

iii) PET-AS contours were generated in Matlab using CERR, within an initial VOI drawn by a non-experienced user.

For the next 10 patients, step (ii) was not carried out, and the PET-AS contours were provided to the clinicians after CT/MRI (i) outlining.

The PET-AS volumes were exported to VelocityAI and all volumes were then sent to the ProSoma workstation, where the final planning outlines were derived by the clinicians using the contours obtained in (i), (ii) and (iii). The implementation of the PET-AS method is presented in details in the following section.

IV. B. Implementation of the PET-AS tool

IV. B. 1. Purpose

One of the aims of the POSITIVE project was to develop one or several PET-AS methods for optimal delineation of the GTV within the RT planning process at Velindre Cancer Centre. Following discussions with the clinicians involved in the planning process the following requirements were identified:

- The segmentation tool should provide accurate contours for the type of tumours targeted.
- The segmentation process should be rapid and simple.
- It should be accessible to non-experts in Matlab or CERR.
- CT-based thresholding should be available, to allow using anatomical information such as the presence of bone or air.

In addition, the observation of existing cases highlighted the fact that high intensity structures, such as the tonsils or malignant lymph nodes could be located close to the target tumour, and therefore included in the GTV delineated by the PET-AS. This led to an additional requirement that the initial region in which to perform the segmentation could be manually selected, so as to ensure that the GTV contains only the primary tumour burden.

IV. B. 2. Methods

IV. B. 2. a. PET-AS algorithm

At the time of planning the patients recruited for the study, the optimised method ATLAAS (cf. III. B) was not yet fully developed. For the clinical cases described in this chapter, a simplified version of ATLAAS was used, based on the analysis of a set of 60 H&N images simulated using PETSTEP as for ATLAAS. The best performing binary method on this dataset was AT, while the best multiple clustering method was GCM5. The data were used to determine an optimal cutoff value for selecting the most accurate method from AT or GCM5, based on TBR_{peak}, calculated as for ATLAAS (cf. III. B). TBR_{peak} was chosen as a classifying parameter because of its low dependency on both contour estimate (obtained using KM2) and absolute SUV values. The decision process used for applying the optimal segmentation method is illustrated on Figure 60. The best performing method for each case considered is shown on Figure 61 together with the cutoff chosen of TBR_{peak}=5. The method's accuracy was evaluated with the phantom data used for validating ATLAAS (cf. III. B).

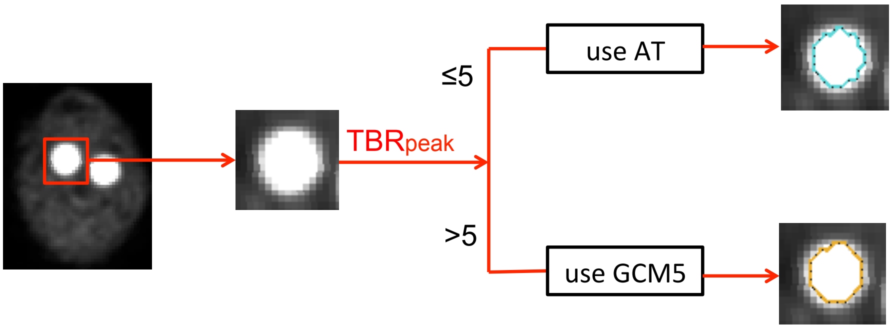


Figure 60. Description of the PET-AS method used in this chapter.

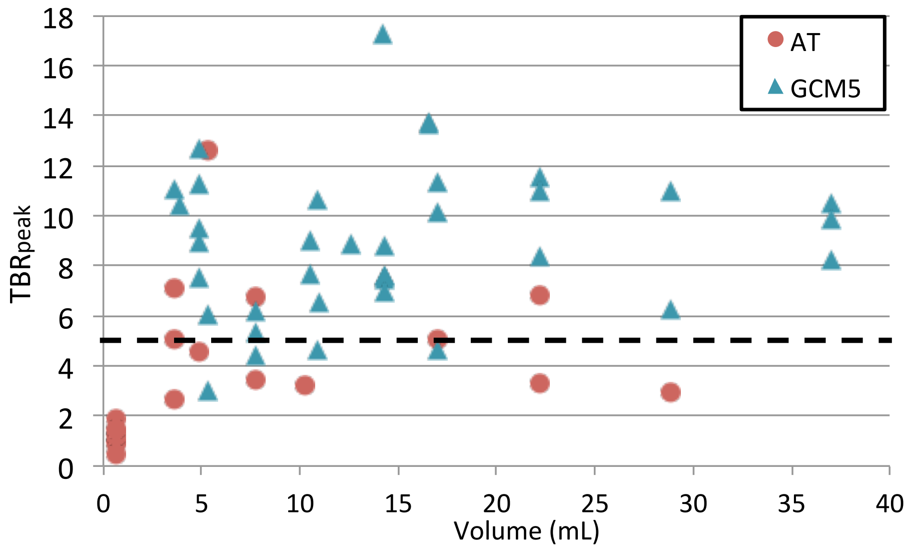


Figure 61. Best performing method of AT and GCM5, for 60 phantom cases, shown by TBR_{peak} and volume, with the TBR_{peak} cutoff value chosen (dashed line).

IV. B. 2. b. PET-AS implementation

The methods were implemented in Matlab, for use with the CERR platform. The code was optimised so as to reduce the processing time. The PET-AS methods can be called from the command line, only specifying:

- the number of the scan to be segmented
- the name to give to the generated contour,
- the VOI in which to perform the segmentation.

The VOI is specified by the coordinates of the six sides of a rectangular 3D box (if the VOI is a 3D rectangle), or in the form of a mask of the same size as the target scan, consisting of voxels with a value of 1 at the location of the initialisation VOI and zeros otherwise. The initialisation VOI coordinates (or mask) could be selected manually by the user on the CERR viewer, but this would require a good knowledge of the software and make this a time consuming process. In addition, this would not allow selecting a non-rectangular VOI. For these reasons, a graphical user interface (GUI) was built to enable the following actions:

- visualisation of the image on three views (transaxial, sagittal and coronal),
- zooming back and forth each view,

- scrolling through each view using crosshairs and a slider bar,
- drawing a rectangle on each view,
- drawing an irregular contour on each view,
- propagating a contour on a selection of slices to make a 3D contour on the current view,
- editing a contour,
- resetting the views.

The GUI returns a mask corresponding to the VOI drawn by the user, obtained as the intersection of the 3D contours drawn on the three different views. In cases where the mask is not a 3D rectangular box, the segmentation is applied to the rectangular bounding box of this non-rectangular initialisation VOI, but only considers the voxels contained in the non-rectangular VOI image for identifying maximum SUV values. The final segmented volume is then masked with the non-rectangular initialisation mask.

A CT-based editing tool was also implemented with a user interface, allowing the user to select a Hounsfield Unit (HU) value as a threshold for removing corresponding voxels from the PET contour, based on the fact that CT and PET images are co-registered.

IV. B. 3. Results

IV. B. 3. a. PET-AS algorithm

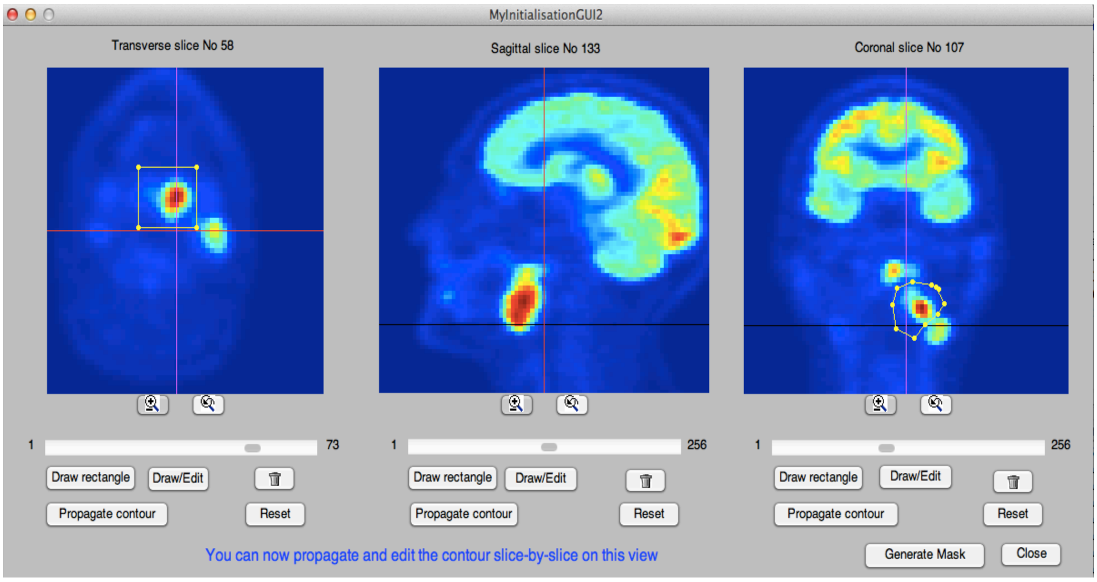
The results given in Table 23 show that the method choosing between AT and GCM5 based on $\mathrm{TBR}_{\mathrm{peak}}$ was an improvement on AT alone, in particular for the case of the largely heterogeneous spheres included in the “Heterogeneous spheres” dataset.

Dataset	H&N SS printed			Fillable phantom			Heterogeneous spheres			Total		
	AT	GCM5	Comb	AT	GCM5	Comb	AT	GCM5	Comb	AT	GCM5	Comb
Mean	0.806	0.648	0.806	0.891	0.681	0.891	0.686	0.398	0.716	0.817	0.621	0.821
Median	0.800	0.814	0.800	0.921	0.695	0.921	0.714	0.303	0.733	0.825	0.670	0.827
Min	0.579	0.137	0.579	0.640	0.419	0.640	0.353	0.142	0.48	0.353	0.142	0.480

Table 23. Comparison of the accuracy (DSC) obtained for AT, GCM5 and the combination of the two methods (Comb) on the phantom dataset used for evaluating ATLAAS.

IV. B. 3. b. PET-AS implementation

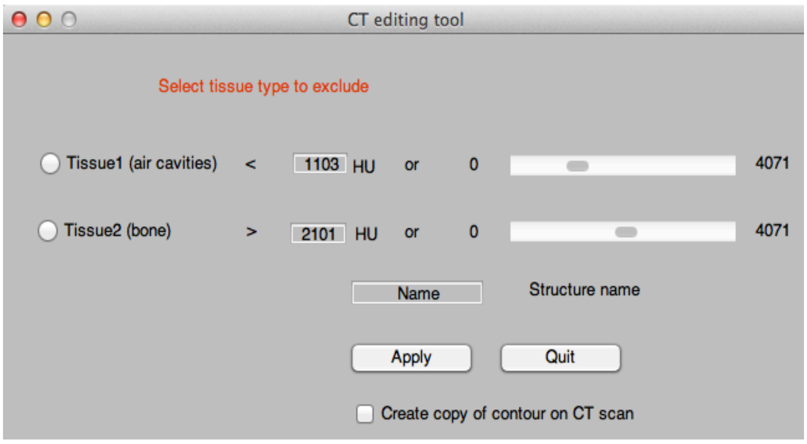
Figure 62 shows a snapshot of the initialisation GUI. Contours were drawn with the rectangle tool on the left (transaxial view) and with a manual drawing tool on the right (coronal view). The button “Generate Mask” closes the window and returns a mask variable in the variable space corresponding to the intersection of the 3D contours drawn on each view.



You can now propagate and edit the contour slice-by-slice on this view

Figure 62. Initialisation interface for selecting the cropped image in which to perform the segmentation.

The GUI developed for the CT-based editing is shown Figure 63. It is called from a single command line with the identifier in CERR of the contour to edit as only input argument. The user can select a high Hounsfield Unit (HU) value, for which all voxels with a higher HU are removed from the corresponding PET contour, or a low HU value, for which all voxels with a lower HU are removed from the corresponding PET contour. This can be done by entering the value in a box or via a slider, automatically positioned between the minimum and maximum, which is updated when the editing box value is changed. The value in the box is also updated according to the slider position. Suggested values, corresponding to arbitrary thresholds for removing air cavities and bone tissue appear as default for the boxes and sliders. The “Apply” button generates a new contour on the PET scan from the original PET contour, using the CT-based editing specified by the user. The user is prompted to enter the name of the new contour, which is then generated on the corresponding data file opened in CERR. A checkbox also allows creating a copy of this new contour onto the registered CT image.



CT editing tool

Select tissue type to exclude

○ Tissue1 (air cavities) < 1103 HU or 0 4071

○ Tissue2 (bone) > 2101 HU or 0 4071

Name Structure name

□ Create copy of contour on CT scan

Figure 63. User interface allowing modifications of the PET-AS contour based on the corresponding CT values.

IV. B. 4. Discussion and conclusions

The tools implemented are simple and easy to use, and allow a good navigation through the image to be segmented. A clinical case can be segmented within one minute, the limiting factor being the selection of the initial cropped image. This is often trivial if the tumour can be encompassed in a rectangle with 10 mm margins, without including any other high intensity structure. At present the PET-AS still heavily relies on the availability of Matlab and CERR (although CERR is freely available for download), and on a basic knowledge of both language and software. Although the current implementation was easily used for the studies described in the following sections, future work will consider implementing the tool in a different environment more accessible to clinicians and non-Matlab experts.

IV. C. Retrospective analysis of outlines for ten oropharyngeal patients

IV. C. 1. Purpose

Following the development and implementation of an optimised PET-AS method, its use was assessed within the clinical RT planning process at Velindre Cancer Centre. This first study, which was retrospective, aimed at testing the workflow and methodology described in IV. A., and at evaluating the differences between PET-AS and clinician outlines, so as to determine the usefulness of PET-AS contours in the process.

IV. C. 2. Methods

Ten oropharyngeal cancer patients were recruited and scanned at VCC following the protocol described in IV. A. The outlines were generated following the same protocol, leading to CT/MRI manual outlines named $GTVp_{CT/MRI}$ and $GTVn_{CT/MRI}$ for primary and nodal volumes respectively, and PET/CT manual outlines named $GTVp_{PET/CT}$ and $GTVn_{PET/CT}$ for primary and nodal volumes respectively.

The PET-AS contours were automatically edited based on the HU values of the corresponding CT image to exclude air cavities and bone tissue. This was done using the interface described in IV. B. with threshold values of 600 and 1700 in CERR respectively, corresponding to values around -600 and +900 HU. The outlines obtained were named $GTVx_{PET-AS}$, with “x” corresponding to the letter p for the primary tumour volume, and n1-nz for lymph node volumes numbered as 1 to z.

The final planning GTV contours were drawn by the clinicians on the basis of the CT/MRI and PET/CT manual outlines, and named $GTVx_{final}$. The interpretation of the PET information in the definition of the final planning GTV was recorded for each case.

All outlines generated were copied onto the planning CT scan. The outlines obtained for each patient with the PET-AS method ($GTVx_{PET-AS}$) were compared to both manual outlines ($\mathrm{GTVx}_{\mathrm{CT/MRI}}$ and $\mathrm{GTVx}_{\mathrm{PET/CT}}$). The volumes corresponding to the different outlines were calculated in Matlab. The conformity of each PET-AS outline with the corresponding PET/CT and CT/MRI outlines was calculated with the DSC.

In addition, the DSC compared to both $\mathrm{GTVp}_{\mathrm{CT/MRI}}$ and $\mathrm{GTVp}_{\mathrm{PET/CT}}$ was calculated separately for each slice of the primary tumour containing one of PET-AS, PET/CT or CT/MRI outline. For each slice with a DSC lower than 0.7, the difference between the reference and PET-AS outline was analysed visually, to check for differences due to the proximity of bone tissue, air cavities and to the absence of outline on one or several slices in the superior-inferior (sup-inf) direction.

IV. C. 3. Results

IV. C. 3. a. Primary volumes

The PET-AS primary volumes were smaller than the CT/MRI volumes for 7 out of 10 cases (cf. Table 24), with absolute differences ranging from 8.9% to 66% of the $\mathrm{GTVp}_{\mathrm{CT/MRI.}}$ The comparison to the PET/CT outlines showed smaller PET-AS volumes for 8 out of 10 cases, and absolute differences in volume ranging between 8.5% and 38% of the $\mathrm{GTVp}_{\mathrm{PET/CT}}$. The PET-AS contours showed a good agreement with CT/MRI contours (median DSC of 0.67) and an excellent agreement with PET/CT contours (median DSC of 0.85). The PET-AS showed greater overlap with PET/CT than with CT/MRI outlines in all cases, but the PET-AS volume was closer to the CT/MRI for 2 out of 10 patients.

	Patient No	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Median**
Volume (mL)	**CT/MRI**	44.8	16.9	20.2	13.7	27.4	13.7	55.0	16.5	109.9	65.4	-
	PET/CT	33.1	11.7	25.7	20.9	22.8	20.9	44.6	17.4	78.1	39.1	-
	PET-AS	30.3	9.00	33.6	17.8	18.5	17.8	33.7	23.8	48.3	71.2	-
DSC(PET-AS vs CT/MRI)		0.77	0.59	0.70	0.74	0.66	0.64	0.70	0.47	0.59	0.68	0.67
DSC(PET-AS vs PET/CT)		0.92	0.72	0.85	0.84	0.87	0.86	0.85	0.73	0.72	0.85	0.85

Table 24. Comparison of PET-AS with manual outlines for primary GTVs, in terms of volume and DSC. The lesion volume given corresponds to the CT/MRI outline.

Figure 64 shows selected slices with manual CT/MRI and PET/CT, and PET-AS outlines for different patients. Differences between automatic and manual outlines were observed in particular in terms of the sup-inf extent of the tumours (cf. Figure 64 a), b) and c)) and around bone tissue (cf. Figure 64 d) and e)) and air cavities (cf. Figure 64 f) and g)). Other differences were due to different soft tissue extent identified in the transverse plane (cf. Figure 64 h) and i)). Table 25 gives the percentage of slices with dissimilar outlines (DSC<0.7), for which the differences between manual and PET-AS outline are due to:

- Different sup-inf extent
- Differences around air cavities or bone tissue
- Other differences in soft tissue extent

Across patients, different sup-inf extent accounted for 10% to 64% of the large differences (DSC<0.7) observed between CT/MRI and PET-AS outlines, and up to 76% of the large differences observed between PET/CT and PET-AS outlines. The differences occurring around bone tissue and air cavities accounted for up to 36% of the large differences with the CT/MRI outline and up to 46% of the large differences with the PET/CT outline.

		Patient No									
Reference	**% of slices with DSC<0.7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CT/MRI	**Different sup-inf extent**	41	29	34	50	61	60	11	10	64	48
	Air/bone	-	-	5.7	-	-	4.3	36	25	4.9	12
	Other	59	71	60	50	39	36	53	65	31	40
PET/CT	**Different sup-inf extent**	60	21	37	-	46	50	27	22	76	33
	Air/bone	-	-	5.3	-	-	-	46	44	-	13
	Other	40	79	58	100	54	50	27	34	24	44

Table 25. Percentage of the dissimilar tumour slices (DSC<0.7) for which the difference between outlines is due to different sup-inf extent, to the proximity of air or bone regions, or other differences.

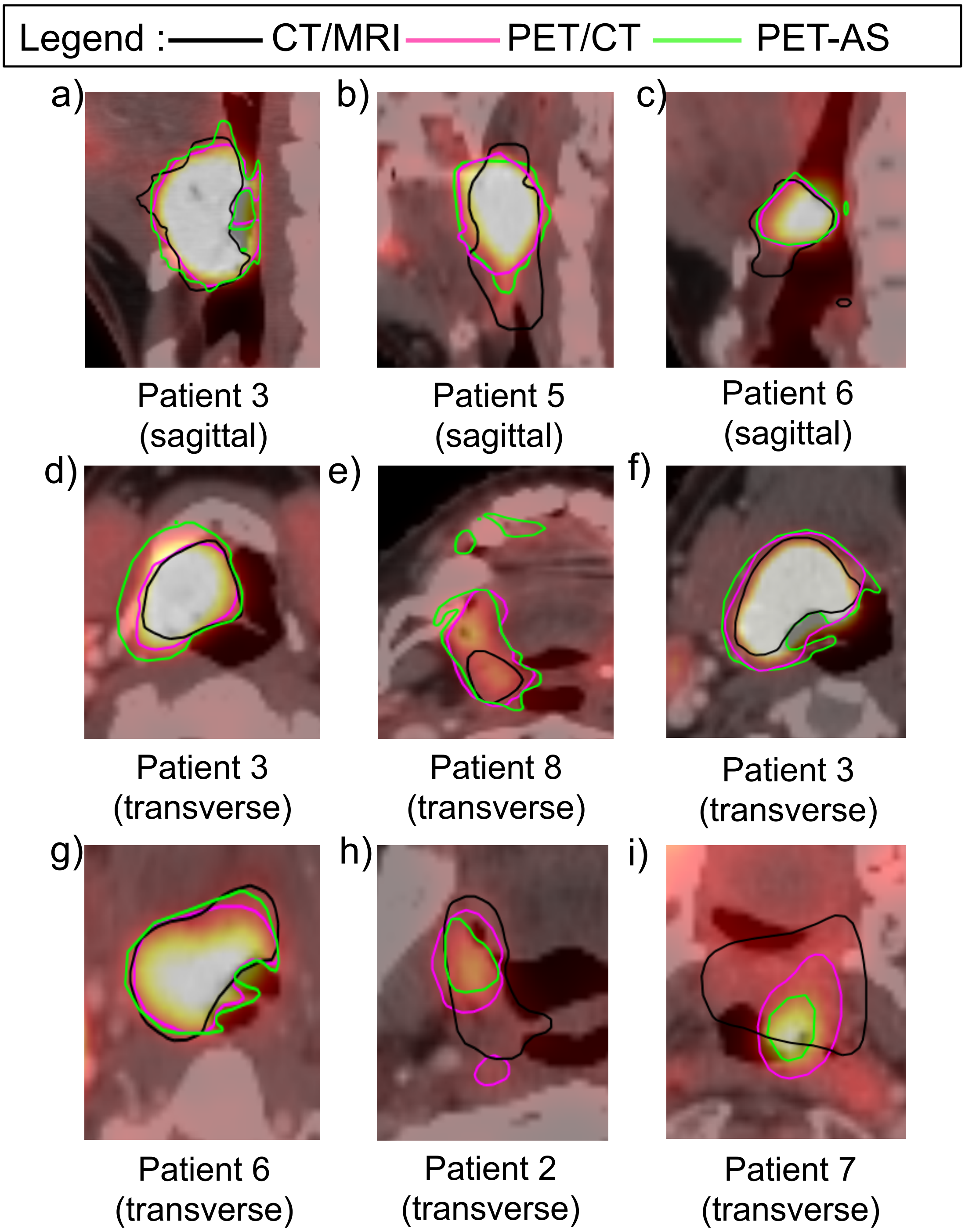


Figure 64. Example of outlines obtained for different patients in the transverse and sagittal directions.

IV. C. 3. b. Lymph nodes

A total of 22 lymph nodes were considered in this study, and are listed in Table 26 for each patient. Differences between the outlining processes were observed in terms of the total nodal volume. In particular, 2 lymph nodes detected on CT/MRI were

PET-negative, while 3 additional malignant lymph nodes were detected when PET was included in the outlining process, one of which was a contralateral lymph node.

The PET-AS nodal volumes were smaller than the CT/MRI volumes for 8 cases, and the comparison to the PET/CT outlines showed smaller PET-AS volumes in 10 cases. The median geographical overlap [160] with the clinician outlines was 0.65 for CT/MRI and 0.64 for PET/CT.

Patient	1	2	3	4	5	6	7	8	9	10
CT/MRI	LN1	LN1	LN1	LN1	-	LN1	LN1	LN1	LN1	-
	LN2	LN2	-	-	-	-	LN2	LN2	LN2	-
	-	-	-	LN3	LN3	-	LN3	LN3	LN3	-
	-	-	-	LN4	-	-	-	-	-	-
PET/CT and PET-AS	LN1	LN1	LN1	LN1	LN1	LN1	LN1	-	LN1	-
	-	LN2	-	LN2	LN2	-	LN2	LN2	LN2	-
	-	-	-	LN3	LN3	-	LN3	LN3	LN3	-
	-	-	-	LN4	-	-	-	-	-	-

Table 26. Lymph nodes outlined for each patient using CT/MRI and with the inclusion of PET (PET/CT or PET-AS outlines).

IV. C. 4. Discussion and conclusions

The results of this analysis show a high similarity between PET-AS outlines and PET/CT outlines obtained by manual delineation. The PET-AS outlines were highly similar (DSC>=0.7) to manual CT/MRI outlines for 4 out of 10 patients. Differences with manual outlines were observed around air cavities and bone tissue, due to a different use of the registered CT information. The PET-AS volumes outlined were smaller than manual outlines in most cases, due to a different extent in the superior-inferior direction. However, the use of FDG-PET provided additional information to manual CT/MRI outlines, in particular in terms of superior-inferior extent, soft tissue extent and lymph node status. This is in line with findings published in the recent literature [24]. The current study shows the potential of using, within the radiotherapy treatment planning process, PET-AS methods, which combine FDG-PET information with fast and reliable delineation of the GTV. Nevertheless, this study does not yet show if notable differences between PET-AS and manual outlines would be translated into a modification of the final contour, used for RT planning. This is the aim of the next study.

The results of this first study were promising, in particular because the clinicians provided excellent feedback on the quality of the contours generated. Their confidence in the PET-AS contour, acquired during this retrospective analysis, as well as the quantitative results obtained led to planning of a second prospective study for which the PET-AS contours would replace clinician manual PET/CT outlines.

IV. D. Prospective analysis of outlines for ten oropharyngeal patients

IV. D. 1. Purpose

The results of chapter IV. C have shown that PET-AS contours provided additional information to manual contours derived by clinicians on PET-CT or CT/MRI data. However, the patients in the previous study were still planned using manual PET/CT outlines. The work presented in this section aimed at demonstrating the feasibility of using PET-AS in replacement of manual PET/CT delineation within RT planning at Velindre Cancer Centre. In addition, the usefulness of the PET-AS contours on clinical H&N data was evaluated by determining in detail the impact of the PET-AS information on the final planning contour of primary and nodal volumes.

IV. D. 2. Methods

Ten oropharyngeal cancer patients were recruited and scanned at VCC following the protocol described in IV.A. Manual CT/MRI outlines and PET-AS contours were generated following the same protocol as for the previous study (cf. IV. C.).

In this case however, the final planning GTV contours were drawn by the clinicians on the basis of the CT/MRI, modified when necessary using the PET-AS contour, with the help of other relevant clinical information. The final planning GTV was named GTVx$_{\text{final}}$. These contours were then used for planning the RT for each patient.

A thorough analysis of the GTVs produced in this process was carried out a. The final planning GTV was compared to both CT/MRI and PET-AS outlines, to identify the cases in which the PET-AS volume was used for producing the final planning contour. The overlap between the final and CT/MRI outlines, as well as the overlap between final and PET-AS outlines were quantified using the DSC (cf. II. A. 1. e). The number of slices included in each contour was also recorded to identify any growth or reduction of superior-inferior extent due to the PET-AS.

Furthermore, a slice-by-slice comparison of the outlines provided the following information:

- Number of voxels in the CT/MRI contour not included in the final contour, corresponding to a shrinkage of the CT/MRI contour,
- Number of voxels outside the CT/MRI contour included in the final contour, corresponding to a growth of the CT/MRI contour,
- Number of voxels in the PET-AS contour not included in the final contour, corresponding to areas where the additional PET-AS information was ignored,
- Number of voxels outside the PET-AS contour included in the final contour, corresponding to areas where the additional PET-AS information was ignored,

For each slice for which a modification (growth or shrinkage) of the CT/MRI contour was found, visual examination was used to determine if the modification was due to the inclusion of the PET-AS data, or to other clinical considerations. For each slice on which additional information brought by the PET-AS was ignored, the reason (i.e. spill-out in bone region, in or around air cavities, or different soft tissue extent) was assessed visually.

IV. D. 3. Results

IV. D. 3. a. Primary volumes

Table 27 provides, for each patient, the volumes of the primary outlines and the DSC quantifying the conformity of the final volume to CT/MRI and PET-AS volumes, as well as PET-AS to CT/MRI volumes. The PET-AS volumes were smaller than the CT/MRI volumes for all patients, with absolute differences ranging from 3.1% to 61% of the $GTVp_{CT/MRI}$. The final contour was larger than both CT/MRI and PET-AS contours for 7 out of 10 patients, with differences of up to 16.5% of the CT/MRI volume. In the three remaining cases, the final volume was intermediate between the CT/MRI and PET-AS volumes for patient No 11, closer to the CT/MRI for patient No 12 and closer to the PET-AS volumes for patient No 16. PET-AS contours showed very good agreement with CT/MRI contours (DSC>0.70) for 7 out of 10 patients. Low agreement for patients No 13, 15 and 20 was due to a much smaller volume outlined on PET-AS.

Patient No		11	12	13	14	15	16	17	18	19	20	Median
Volume (mL)	**Final**	33.1	45.9	21.5	36.6	27.5	54.7	33.1	19.0	33.1	17.4	-
	CT/MRI	27.1	47.5	19.8	32.3	26.9	60.8	28.5	16.6	30.1	15.6	-
	PET-AS	27.3	41.9	7.8	24.5	15.6	52.5	29.0	16.1	23.3	8.6	-
DSC(final vs CT/MRI)		0.90	0.92	0.96	0.91	0.99	0.84	0.91	0.85	0.96	0.94	0.92
DSC(final vs PET-AS)		0.78	0.79	0.53	0.81	0.68	0.97	0.84	0.92	0.83	0.58	0.80
DSC (PET-AS vs CT/MRI)		0.77	0.76	0.43	0.73	0.67	0.82	0.74	0.73	0.76	0.51	0.74

Table 27. Comparison of PET-AS with manual outlines for primary GTVs in terms of volume and DSC. The lesion volume given corresponds to the CT/MRI outline.

Table 28 provides the results of the slice-by-slice analysis, describing for each clinical case the changes made to the CT/MRI on the basis of the PET-AS to make the final contour. The top row of the table shows that the whole PET-AS volume was included in the final GTV in 4 cases, and more than 83% was included for all patients. The second row provides the proportion of the final GTV volume that was modified using the PET-AS contour, which ranges from 0.4% to 33.3% across patients. The changes made to the CT/MRI contours included superior-inferior growth of up to 15 mm for six patients, and a reduction in superior-inferior extent for one patient. For six patients, additional differences in superior-inferior extent did not lead to any modification of the CT/MRI volume. The CT/MRI volume was grown locally based on the PET-AS information for all patients, with up to 11.6 mL added to make the final volume. This accounts for the fact that the final contour was often derived as the union of the CT/MRI and PET-AS volumes. This was the case for patients No 11, 13, 14, 17, 18, and 19, for which all CT/MRI and PET-AS information was included on most slices, as can be seen with the examples of patients No. 11 and 17 in Figure 65 a) and b) respectively. However, in some specific slices for these patients, the final contour differed from the union. This was the case for example on some slices for patient No 11, when some bone tissue was included in the PET-AS contour (cf. Figure 65 c)). This is also quantified in row 9 of Table 28. A compromise between CT/MRI and PET information was sometimes derived when the contour was modified on the basis of the PET-AS information, but following the edge of anatomical structures seen on CT or MRI, as shown for patients No 15 and 12 on Figure 65 d) and e) respectively.

A non-negligible amount of information provided by the PET-AS contour was not considered in drawing the final volume (cf. row 7 of Table 28). This includes both areas where the PET-AS was smaller than the CT/MRI contour, and areas where it was larger, which explains that differences can be larger than 100%. For patient No 13 and No 20, these large differences are due to the fact that the PET-AS was 61% and 44% smaller respectively than the CT/MRI outline, but the final contour was not reduced based on the PET-AS, as can be seen on Figure 65 f) and g) respectively. For patient No 12, the PET-AS included large areas (8.4 mL in total) around air cavities, as shown on Figure 65 h), which were not always suitable to be included in the planning volume. In some cases, additional clinical information led to inclusion or exclusion of the PET-AS information. For patient No 17, the PET-AS contour included tumour extension in the soft palate, which was not shown clinically. The CT/MRI contour was not extended superiorily following the PET-AS in that case. Similarly for patient No 13, clinical examination had shown abnormal mucosal extent, which is not visible on PET data. The volume was therefore not shrunk on the basis of the smaller PET-AS volume. A similar situation was encountered for patient No 20, for which the inferior extent of the PET-AS was much smaller, but this was ignored in the final contour. For patient No 18, the clinical report confirmed extension of the tumour to the midline, which was observed on the CT data but not on the MRI scan. The contour was extended to the midline on some slices following the PET-AS contour, which agreed with the CT data, confirmed by the clinical finding, as shown on Figure 65 i) and j) for patients No 14 and 20 respectively.

		Patient No									
		11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
% PET-AS included in GTV$_{\mathbf{final}}$		99.6	83.2	100	100	94.0	99.1	91.8	100	100	88.0
% of change in GTV$_{\mathbf{final}}$ due to PET-AS information		0.9	16.3	7.7	0.8	2.5	33.3	0.8	28.2	0.4	10.8
Superior-inferior extent (mm)	Grown	8	15	-	4	3	6	-	-	-	8
	Shrunk	-	-	-	-	-	-	-	4	-	-
Transverse using PET-AS (mL)	Grown	7.4	1.5	1.9	5.8	1.3	11.6	5.7	4.5	2.9	1.2
	Shrunk	-	-	-	0.5	0.1	24.6	0.6	1.1	-	-
% PET-AS not considered in GTV$_{\mathbf{final}}$		13.7	43.3	100.6	45.8	89.0	5.3	35.0	18.4	42.4	126.1
Superior-inferior extent (mm)		4	-	38	-	3	-	4	-	4	11
Bone regions (mL)		0.14	-	-	-	-	-	-	-	-	-
Air cavities or vicinity (mL)		-	1.0	-	-	1.7	0.9	1.3	-	-	1.9
Transverse soft tissue extent (mL)		6.7	7.1	8.9	13.8	25.4	1.7	9.3	3.4	11.5	13.8

Table 28. Quantification of the use of PET-AS data in the outlining process given as a percentage of the final planning GTV volume.

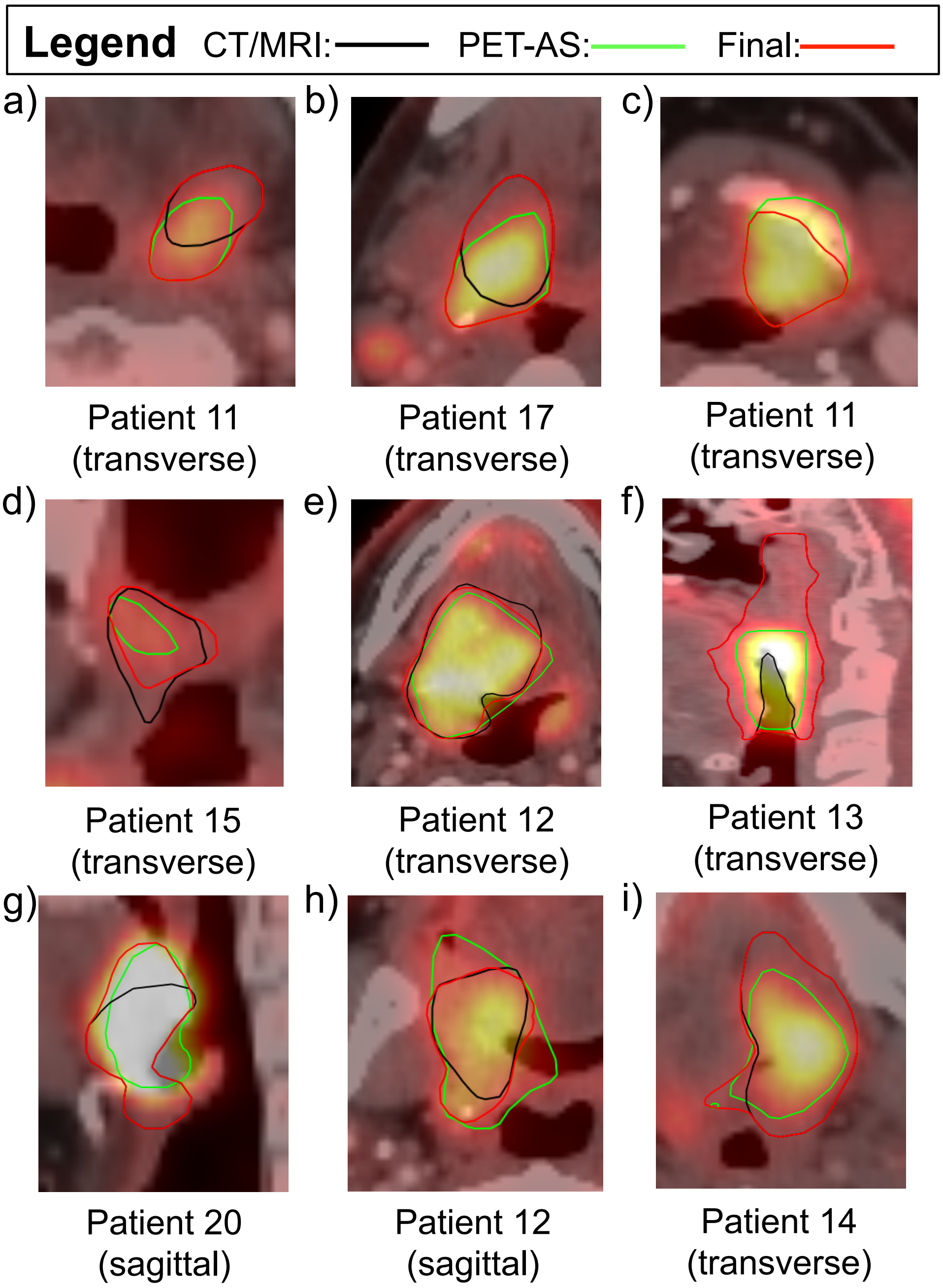


Figure 65. Examples of transverse slices showing the use of PET-AS contour in deriving the final volume from the CT/MRI volume. The GTV$_{\text{final}}$ is overlayed on the GTV$_{\text{CT/MRI}}$, which is itself overlayed on the GTV$_{\text{PET-AS}}$.

IV. D. 3. b. Lymph nodes

The PET-AS outlines were generated for all patients, but were not used for deriving the final volumes. This was due to the very high similarity between the GTVn volumes for CT/MRI and PET-AS in some cases, such as for patient No 20, as shown on Figure 66 a). However, in 4 out of 10 patients, the PET-AS lymph node volumes included large parts of adjacent blood vessels, and were therefore not considered in deriving the final volume. This is shown for patients No 11 and 17, on Figure 66 b) and c) respectively. Two lymph nodes for patient No 20, which were not identified as malignant on CT/MRI information, were outlined by the PET-AS method and included in the treatment.

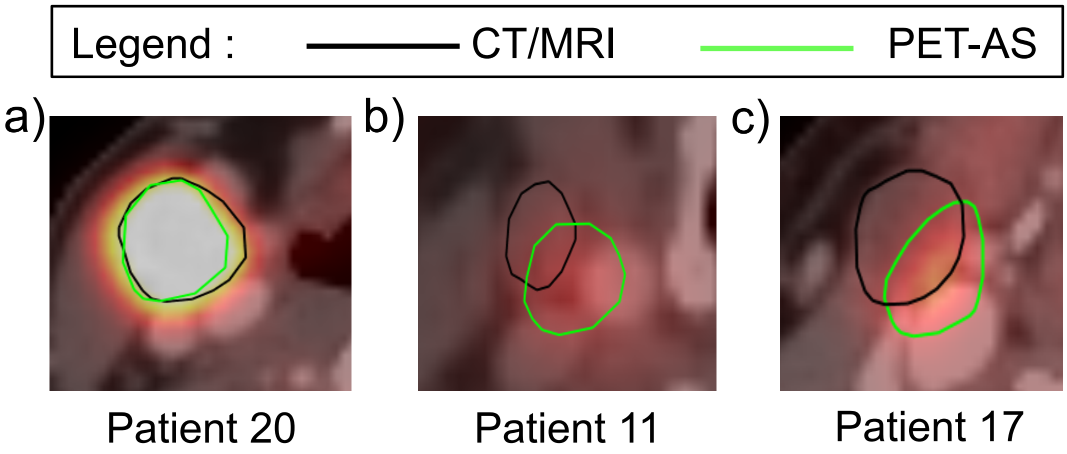


Figure 66. Transverse slice of CT/MRI and PET-AS outlines of lymph nodes for a) patient No 20, b) patient No 11 and c) patient No 17.

IV. D. 4. Discussion and conclusions

This work follows a retrospective study during which the clinicians Velindre Cancer Centre familiarised with the PET-AS methods and workflow. The retrospective study discussed in the previous section was very important in growing the team's confidence in using a PET-AS method for routine clinical practice. In this prospective study, a complete workflow was evaluated including patient recruitment, scanning, reporting, outlining and planning for the RT management of H&N oropharyngeal cancer patients. This included quality testing of the different steps in the workflow, training of the radiographers for acquiring planning PET/CT images of the H&N, including contrast CT, which had not been used at Velindre Cancer Centre previously in combination with FDG-PET. Most importantly, manual PET-CT contours were not needed in this study, which allowed reducing the clinicians workload by about a third for one GTV, leaving to them only the manual CT/MRI and final contours to outline. This represents an important change in clinical practice, and the positive results of this work show that the inclusion of PET-AS into the RT planning process at Velindre Cancer Centre was achieved for oropharyngeal primary tumours.

The results of this study highlight the similarity between PET-AS outlines and CT/MRI outlines obtained by manual delineation. In all cases, some additional information provided by the PET-AS contours contributed to the delineation of the final target volume. This positive result shows the confidence of the clinicians in using the PET-AS volumes clinically. The role of the clinicians remained of course paramount in making the final decision of including or excluding the PET-AS information. In particular, the analysis highlighted some limitations of the PET data, such as the absence of signal in abnormal mucosa (for patient No 13) and signal spill-out in air cavities or bone tissue (cf. Figure 65 and Table 28).

The decision of including the PET information or not highly depends on the clinician's judgment and expertise, and on the availability of additional clinical data, such as endoscopy or other clinical examinations. It is important to note that even when the PET-AS contours did not significantly differ from the CT/MRI contours, they were still very useful in reassuring the clinician that no additional malignant regions were missed. In addition, in cases for which conflicting or inconclusive data from CT and MRI were available (e.g. due to different patient positioning or poor image registration), the PET-AS data was useful in guiding the clinician and confirming the findings of one or the other imaging modality.

As far as lymph nodes are concerned, the PET-AS did not provide information that would change the planning process. This is because lymph nodes are often well defined on CT data, particularly if contrast is added to the imaging like it was in this study. The contrast agent allows a better visualisation of blood vessels, which are not well discriminated in the PET scan. However, this work has shown that the PET information provided by PET-AS delineation can play a crucial role in the detection of malignant lymph node, with potential consequences on patient management (cf. IV. C).

The results of this study have shown that including in the RT planning process FDG-PET information provided by PET-AS outlines can lead to significant changes to the final planning volume. A fully automatic outlining process was used within a simple workflow to provide PET-AS contours to the clinicians. The PET-AS is a very rapid process, lasting no more than 1 minute for a single outline. Therefore the time required for the patient GTV outlining was reduced by about a third compared to a process requiring both CT/MRI and PET/CT manual delineation. This represents a major advantage beyond the use of manual PET information in the RT planning process. As well as a reduction in time, other advantage of PET-AS compared to manual PET delineation are its low inter-observer variability, and reproducibility. The use of PET-AS could prove extremely useful for treatment involving dose escalation or volume boosting and this will be further investigated at Velindre Cancer Centre.

Chapter V. Discussion and conclusions

V. A. Future work

Throughout this work, a range of advanced PET-AS methods were evaluated and compared, and the knowledge acquired during the project was used to develop an optimal segmentation process for routine clinical practice at Velindre Cancer Centre. To achieve this goal, several novel phantom techniques were also developed and evaluated, and additional work was carried out to implement a new PET simulator tool.

The literature review carried out at the start of the project (cf. I. E) revealed a great variability of the segmentation and validation approaches published to date. This was also observed by expert panels such as the IAEA [43], and the American Association of Physicists in Medicine (AAPM), which appointed the Task Group 211 (TG211) on "Classification, Advantages and Limitations of PET segmentation methods"[2]. TG211 issued a first report highlighting the need for a benchmark tool allowing standardised evaluation of PET segmentation methods on a wide range of clinically relevant data [161]. Due to the relevance of such work in the context of POSITIVE, a close collaboration was started with TG211 in 2012 for the development of the benchmark tool, which was named PET-AS suite of evaluation tools (PETASset). The group in Cardiff University led data processing and software development of the PETASset code, which is currently maintained there. The tool is described in a publication currently under review by the AAPM committee, and was presented to the wider community at several occasions ([157], [162]).

The PET-AS methods used throughout this project were inspired by published work, but were developed in house with custom implementation, as fully automatic

[2] http://aapm.org/org/structure/default.asp?committee_code=TG211

algorithms. These were first tested and optimised with the data from the NEMA IEC body phantom used in II. A. 2. b and III. A. 3. To make the optimisation as general and robust as possible, it was carried out using a clinically relevant range of target images with four different techniques:

- NEMA IEC spherical inserts (cf. II. A. 3)
- Raydose thin-wall spherical and non-spherical inserts (cf. II. B)
- Printed subresolution sandwich phantom irregular and heterogeneous tumours (cf. III. A. 2)
- PETSTEP simulated irregular and heterogeneous tumours (cf. III. B)

The use of non-spherical and heterogeneous tumours highlighted strengths and weaknesses of some PET-AS methods, which had not yet been previously identified when using spherical target objects. For example, a lack of sensitivity of RG was observed, as well as large errors of GC and other gradient-based methods for thin-end objects (cf. II. B. 2). The results also demonstrated that binary methods are not adequate for highly heterogeneous tumours (cf. III. A. 3). This shows the drawback of using basic phantoms to develop PET-AS algorithms. However, some other PET-AS methods, such as AT, RG or GCM, which were also included in ATLAAS (cf. III. B), showed robustness to the different target object types, and performed well on all datasets. This shows that the optimisation using spherical inserts did not hamper these algorithms' usefulness. The robustness of the PET-AS methods was investigated using data from Velindre Cancer Centre only. Testing the PET-AS methods with data from other centres would represent an important step in the validation of the segmentation processes, in particular in future developments of ATLAAS, and for participating in large multi-centre trials.

It is important to note, as shown in II. A. 3, that the choice of the initial VOI does have an impact on the final segmentation result, which represents the only operator input in the segmentation process implemented throughout this work. It is not fully clear how much of an impact the operator input has on the performance of the PET segmentation process. Some studies have shown that higher levels of interactivity lead to higher accuracy [161], but any operator input increases the dependence on human judgement, and generates inter-observer variability. This variability can be minimised by using a protocol for selecting the initial VOI, for example with a set window-level. However, clinician judgment will still have a large impact on the structures to include in the initialisation VOI, and should not be ignored. The role of the clinician extends far beyond the PET-based delineation, and the PET-AS contour should always be checked and edited if needed before being used for planning.

Throughout this PhD work, a number of tools were developed for the generation of realistic PET phantom images. The Raydose phantom inserts, manufactured with a vacuum-moulding technique developed at VCC played a key part in the assessment of the effect of cold walls on PET imaging, which is of crucial importance in the field of nuclear medicine imaging. Work carried out throughout this PhD project has contributed to the knowledge of the field by exposing the limitation of commonly used phantoms (cf. II. B. 1). Using the vacuum-moulding technique, non-spherical inserts were also produced, and further used to identify strengths and weaknesses of the PET-AS approaches implemented during the course of this project (cf. II. B. 2). The printed subresolution sandwich phantom developed within this PhD work (cf. III. A. 1) and the PET simulator PETSTEP, which was implemented and tested during this PhD work [157], proved extremely useful tools for the production of heterogeneous PET uptake. This is currently a hot topic in the field of PET imaging. The PETSTEP tool is under continuous development at the memorial Sloane Kettering Cancer Centre in New York (USA), and it is expected to become a useful tool for the scientific community. PETSTEP was implemented in the PETASset framework and it will be publicly distributed together with the CERR software. The subresolution sandwich printed phantom was shown to be very useful in validating PET-AS methods.

More work is needed to ensure that the technique can also be used in quantitative imaging (cf. III. A. 1). Current work is in progress (HiRis-NM: 3-D Printed sources for High-Resolution Molecular Imaging, Velindre grant number SGS/13/01) to eliminate the use of 2 mm PMMA sheets and print 3D activity volumes.

The use of non-spherical, heterogeneous and irregular lesion models was extremely useful in identifying the benefits of different segmentation approaches in a range of clinically relevant situations for H&N cancer. In particular, this work demonstrated that no PET-AS method was systematically the most accurate among those tested, but showed that different approaches performed best in a specific set of conditions (cf. II. B. 2, III. A. 2 and III. A. 3). Following these observations, it was concluded on the need to combine these methods to achieve optimal segmentation. An Advanced decision Tree-based Learning Algorithm for Automatic Segmentation (ATLAAS) was therefore proposed. ATLAAS is an optimised predictive model for automated PET image segmentation based on the statistical approach of *Decision Tree Learning*. The model showed high accuracy across a range of test images including the phantom and simulated data generated in this work (cf. III. B). The model achieved the near-optimal segmentation accuracy targeted and returned values within the best DSC achievable in 77% of the cases for the different phantom datasets.

The results of the clinical study were highly encouraging, as they showed good conformity between the PET-AS contours and the clinicians' contours. Since ATLAAS was still under development at the time of patient RT planning, the contours were generated with a simplified version of the model described in III. B, only using two PET-AS methods (cf. IV. B). The current version of ATLAAS, which is in continuous development, includes 14 PET-AS methods, to be trained on data from an updated version of PETSTEP, with the addition of metrics describing the tumour geometry. Future work will also aim to improve the segmentation by adding more complex information extracted from registered anatomical imaging, such as CT or MRI data. This could be particularly helpful in the case of lymph nodes (cf. IV. D) for which contrast enhanced CT-based information could discriminate between nodal tissue and blood vessels.

As a pilot project, POSITIVE was key in validating the segmentation workflow described in IV. A. The use of PET-AS algorithms was embedded in the clinical environment and no issues were reported. The current workflow relies on three different software platforms, and on expertise with Matlab/CERR, VelocityAI and ProSoma. Work is in progress to modify the current workflow and to have a single clinician operating all segmentation tasks on one single platform. This will involve building a standalone package including ATLAAS and the initialisation tools described in IV.A.4.

The clinical use of PET-AS contours provided by the optimised segmentation workflow was found beneficial for the planning of oropharyngeal cancer patients at Velindre Cancer Centre (cf. IV. C). In the final prospective study, the PET information previously provided via manual PET/CT-based delineation was provided only via the PET-AS process, which reduced the delineation time for the clinicians involved in the planning by about a third. The additional information brought by the PET-AS contours was largely used to grow or shrink the CT/MRI-based GTV to derive the final volume. Future work should aim to further validate the segmentation process, further underpinning the application of the optimised PET-AS methods and clinical workflows to other tracers, and other malignancies. This is expected to improve the ability of the clinical team to provide patients with state-of-the-art treatments and to lead future research and clinical trials in the field.

In this work the accuracy and reliability of the optimised PET-AS method for delineating H&N tumours was extensively validated. An increasing number of publications currently focus on using such outlines for identifying specific regions within the tumour, rather than the whole tumour extent [58]–[60]. The study described in III. A. 3 showed that binary approaches tend to delineate the high intensity areas in highly heterogeneous lesions. This type of method could be used to determine a volume to be boosted in dose escalation studies. In this case, additional work would be required to evaluate the required sensitivity for the PET-AS methods involved in the segmentation process. Additionally, application of PET-AS segmentation algorithms or a predictive model such as ATLAAS, could be tested and validated with different tracers more specific than FDG to cell radioresistance, such as ^{18}F -MISO, or to tumour hypoxia, such as ^{18}F -FAZA.

V. B. Final remarks

This project aimed at investigating PET automatic segmentation (PET-AS) approaches and providing an optimised method for use in H&N RT planning at Velindre Cancer Centre. For this purpose, the following results were achieved:

- Validation, optimisation, and comparison of a selection of eight advanced PET-AS approaches using a total of 204 phantom images, including 144 fillable inserts and 60 printed H&N lesions. Adaptive iterative thresholding (AT) and region-growing (RG) showed high accuracy in most cases, with Dice Similarity Coefficient values higher than 0.7. The Gaussian mixture models-based clustering method GCM proved the most robust clustering PET-AS method to target object characteristics,
- Demonstration of the fact that PET-AS accuracy increases with object size, tumour to background ratio, and decreases with noise level, object geometry complexity and heterogeneity. The lack of sensitivity of the region growing method was highlighted for thin geometrical shapes, as well as the difficulties encountered by gradient-based methods for delineating complex geometries. The accuracy of clustering methods and of the adaptive iterative thresholding method varied the least with these image parameters,

- Development of ATLAAS, an Advanced decision Tree-based Learning Algorithm for Automatic Segmentation which uses a limited set of PET-AS approaches to provide contours of accuracy within 10% of the best achievable Dice Similarity Coefficient in 89% and 77% cases for simulated and phantom data respectively,
- Developement and application to 10 oropharyngeal cancer patients of a protocol and a workflow for the use of PET-AS in H&N RT planning, which is being considered for implementation as standard clinical practice in Velindre Cancer Centre.

Additional achievements included:

- Use of custom-made vacuum-moulding inserts to show that the presence of inactive plastic walls in fillable phantoms can decrease the image activity recovery by 25% and lead to significantly different segmentation results from PET-AS methods [131]. As a consequence, more advanced phantom techniques were used in the rest of this work,
- Demonstration of the fact that using non-spherical thin-wall inserts to evaluate PET-AS methods can highlight strengths and weaknesses of these methods not seen previously [137],
- Development and validation of a printed subresolution sandwich phantom, which enabled modelling a custom FDG uptake map of the H&N including 13 anatomical structures, to which irregular and heterogeneous lesion uptakes could be added with known ground truth,
- Development and validation of ATLAAS, a model for optimal segmentation that could potentially be applied to any tumour site and used at any institution,
- Demonstration of the use of PET-AS in addition to CT and MRI outlining within a clinical protocol for RT planning of oropharyngeal cancer patients

can lead to (a) growth or reduction of the initial GTV and to the identification of additional involved lymph nodes, while reducing the workflow and bringing the contouring time down by about a third.

This project has contributed to the knowledge and understanding of the PET-AS process in the field of nuclear medicine, in particular within the institutions involved. The work presented in this thesis has paved the way for an increased use of PET-AS in clinical practice at Velindre Cancer Centre.

References

[1] Office for National Statistics, "Cancer Registration Statistics, England 2012, Statistical Bulletin No 43," 2014.

[2] J. Tobias and D. Hochauser, *Cancer and its management*, 6th ed. Wiley-Blackwell, 2010.

[3] L. H. Sobin, C. H. Wittekind, and M. Gospodarowicz, "TNM Classification of Malignant Tumours," Wiley-Liss, New York, 2009.

[4] C. Powell, M. Schmidt, M. Borri, D.-M. Koh, M. Partridge, A. Riddell, G. Cook, S. a Bhide, C. M. Nutting, K. J. Harrington, and K. L. Newbold, "Changes in functional imaging parameters following induction chemotherapy have important implications for individualised patient-based treatment regimens for advanced head and neck cancer.," *Radiother. Oncol.*, vol. 106, no. 1, pp. 112–7, Jan. 2013.

[5] K. Wang, D. E. Heron, D. a Clump, J. C. Flickinger, G. J. Kubicek, J.-C. M. Rwigema, R. L. Ferris, J. P. Ohr, A. E. Quinn, C. Ozhasoglu, and B. F. Branstetter, "Target delineation in stereotactic body radiation therapy for recurrent head and neck cancer: A retrospective analysis of the impact of margins and automated PET-CT segmentation.," *Radiother. Oncol.*, vol. 106, no. 1, pp. 90–5, Jan. 2013.

[6] V. J. Lowe, F. R. Dunphy, M. Varvares, H. Kim, M. Wittry, C. H. Dunphy, T. Dunleavy, E. McDonough, J. Minster, J. W. Fletcher, and J. H. Boyd, "Evaluation of chemotherapy response in patients with advanced head and neck cancer using [F-18]fluorodeoxyglucose positron emission tomography.," *Head Neck*, vol. 19, no. 8, pp. 666–74, Dec. 1997.

[7] The Department of Veterans Affairs Laryngeal Cancer Study Group, "Induction Chemotherapy plus Radiation Compared with Surgery plus Radiation in Patients with Advanced Laryngeal Cancer," *N. Engl. J. Med.*, vol. 324, no. 24, pp. 1685–1690, Jun. 1991.

[8] K.-C. Soo, E.-H. Tan, J. Wee, D. Lim, B.-C. Tai, M.-L. Khoo, C. Goh, S.-S. Leong, T. Tan, K.-W. Fong, P. Lu, a See, and D. Machin, “Surgery and adjuvant radiotherapy vs concurrent chemoradiotherapy in stage III/IV nonmetastatic squamous cell head and neck cancer: a randomised comparison.,” *Br. J. Cancer*, vol. 93, no. 3, pp. 279–86, Aug. 2005.

[9] J.-P. Pignon, A. le Maître, E. Maillard, and J. Bourhis, “Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients.,” *Radiother. Oncol.*, vol. 92, no. 1, pp. 4–14, Jul. 2009.

[10] J. B. Vermorken, E. Remenar, C. van Herpen, T. Gorlia, R. Mesia, M. Degardin, J. S. Stewart, S. Jelic, J. Betka, J. H. Preiss, D. van den Weyngaert, A. Awada, D. Cupissol, H. R. Kienzer, A. Rey, I. Desaunois, J. Bernier, and J.-L. Lefebvre, “Cisplatin, fluorouracil, and docetaxel in unresectable head and neck cancer.,” *N. Engl. J. Med.*, vol. 357, no. 17, pp. 1695–704, Oct. 2007.

[11] M. R. Posner, D. M. Hershock, C. R. Blajman, E. Mickiewicz, E. Winquist, V. Gorbounova, S. Tjulandin, D. M. Shin, K. Cullen, T. J. Ervin, B. a Murphy, L. E. Raez, R. B. Cohen, M. Spaulding, R. B. Tishler, B. Roth, R. D. C. Viroglio, V. Venkatesan, I. Romanov, S. Agarwala, K. W. Harter, M. Dugan, A. Cmelak, A. M. Markoe, P. W. Read, L. Steinbrenner, a D. Colevas, C. M. Norris, and R. I. Haddad, “Cisplatin and fluorouracil alone or with docetaxel in head and neck cancer.,” *N. Engl. J. Med.*, vol. 357, no. 17, pp. 1705–15, Oct. 2007.

[12] S. a Bhide, M. Ahmed, Y. Barbachano, K. Newbold, K. J. Harrington, and C. M. Nutting, “Sequential induction chemotherapy followed by radical chemo-radiation in the treatment of locoregionally advanced head-and-neck cancer.,” *Br. J. Cancer*, vol. 99, no. 1, pp. 57–62, Jul. 2008.

[13] J. a Bonner, P. M. Harari, J. Giralt, N. Azarnia, D. M. Shin, R. B. Cohen, C. U. Jones, R. Sur, D. Raben, J. Jassem, R. Ove, M. S. Kies, J. Baselga, H. Youssoufian, N. Amellal, E. K. Rowinsky, and K. K. Ang, “Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck.,” *N. Engl. J. Med.*, vol. 354, no. 6, pp. 567–78, Feb. 2006.

[14] K. Pigott, S. Dische, and M. I. Saunders, “Where exactly does failure occur after radiation in head and neck cancer?,” *Radiother. Oncol.*, vol. 37, no. 1, pp. 17–19, Oct. 2014.

[15] J. Overgaard, “Hypoxic modification of radiotherapy in squamous cell carcinoma of the head and neck – A systematic review and meta-analysis,” *Radiother. Oncol.*, vol. 100, no. 1, pp. 22–32, 2011.

[16] C. Kelly, “Deterioration in quality of life and depressive symptoms during radiation therapy for head and neck cancer,” *Otolaryngol. - Head Neck Surg.*, vol. 136, pp. 108–111, 2007.

[17] C. M. Nutting, J. P. Morden, K. J. Harrington, T. Guerrero Urbano, S. A. Bhide, C. Clark, E. A. Miles, A. B. Miah, K. Newbold, M. Tanay, F. Adab, S. J. Jefferies, C. Scrase, B. K. Yap, R. P. A’Hern, M. A. Sydenham, M. Emson, E. Hall, and and on behalf of the P. trial management Group, “Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial,” *Lancet Oncol.*, vol. 12, no. 3, pp. 127–136, 2011.

[18] I. C. R. U. International Commission on Radiation Units and Measurements, “Prescribing, Recording, and Reporting Photon Beam Therapy (Report 50),” Bethesda, MD, USA, 1993.

[19] C. F. Nieh, “Tumor delineation: the weakest link in the search for accuracy in Radiotherapy,” *J. Med. Phys.*, vol. 33, no. 4, pp. 136–140, 2008.

[20] R. Hermans, M. Feron, E. Bellon, P. Dupont, W. Van den Bogaert, and A. L. Baert, “Laryngeal tumor volume measurements determined with CT: A study on intra- and interobserver variability,” *Int. J. Radiat. Oncol.*, vol. 40, no. 3, pp. 553–557, 1998.

[21] P. Tai, J. Van Dyk, E. Yu, J. Battista, L. Stitt, and T. Coad, “Variability of target volume delineation in cervical esophageal cancer,” *Int. J. Radiat. Oncol.*, vol. 42, no. 2, pp. 277–288, 1998.

[22] M. Ahmed, M. Schmidt, A. Sohaib, C. Kong, K. Burke, C. Richardson, M. Usher, S. Brennan, A. Riddell, M. Davies, K. Newbold, K. J. Harrington, and C. M. Nutting, "The value of magnetic resonance imaging in target volume delineation of base of tongue tumours - a study using flexible surface coils.," *Radiother. Oncol.*, vol. 94, no. 2, pp. 161–7, Feb. 2010.

[23] D.-L. Hou, G.-F. Shi, X.-S. Gao, J. Asaumi, X.-Y. Li, H. Liu, C. Yao, and J. Y. Chang, "Improved longitudinal length accuracy of gross tumor volume delineation with diffusion weighted magnetic resonance imaging for esophageal squamous cell carcinoma.," *Radiat. Oncol.*, vol. 8, no. 1, p. 169, Jul. 2013.

[24] J.-F. Daisne, T. Duprez, B. Weynand, M. Lonneux, M. Hamoir, H. Reychler, and V. Grégoire, "Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen.," *Radiology*, vol. 233, no. 1, pp. 93–100, Oct. 2004.

[25] X. Geets, J.-F. Daisne, S. Arcangeli, E. Coche, M. De Poel, T. Duprez, G. Nardella, and V. Grégoire, "Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between CT-scan and MRI.," *Radiother. Oncol.*, vol. 77, no. 1, pp. 25–31, Oct. 2005.

[26] W. H. Sweet and G. L. Bronwell, "Localization of brain tumors with positron emittors," *Nucleonics*, vol. 11, pp. 40–45, 1953.

[27] P. F. Sharp, H. G. Gemmel, and A. D. Murray, Eds., *Practical Nuclear Medicine*, Springer V. London: Oxford University Press, 2005.

[28] S. Kitson, V. Cuccurullo, A. Ciarmiello, D. Salvo, and L. Mansi, "Clinical Applications of Positron Emission Tomography (PET) Imaging in Medicine: Oncology, Brain Diseases and Cardiology," *Curr. Radiopharm.*, vol. 2, no. 4, pp. 224–253, Oct. 2009.

[29] O. Warburg, F. Wind, and E. Neglers, "The metabolism of tumors in the body," *J Gen Physiol*, vol. 8, no. 6, pp. 519–530, 1927.

[30] K. Ito, J. Yokoyama, K. Kubota, and M. Morooka, "Comparison of 18F-FDG and 11C-choline PET/CT for detecting recurrences in patients with non squamous cell head and neck malignancies," *Nucl. Med. Commun.*, vol. 31, no. 11, pp. 931–937, 2010.

[31] A. Chalkidou, D. B. Landau, E. W. Odell, V. R. Cornelius, M. J. O'Doherty, and P. K. Marsden, "Correlation between Ki-67 immunohistochemistry and 18F-fluorothymidine uptake in patients with cancer: A systematic review and meta-analysis.," *Eur. J. Cancer*, vol. 48, no. 18, pp. 3499–513, Dec. 2012.

[32] E. G. C. Troost, J. Bussink, A. L. Hoffmann, O. C. Boerman, W. J. G. Oyen, and J. H. a M. Kaanders, "18F-FLT PET/CT for early response monitoring and dose escalation in oropharyngeal tumors.," *J. Nucl. Med.*, vol. 51, no. 6, pp. 866–74, Jun. 2010.

[33] J. Wedman, J. Pruim, and J. L. N. Roodenburg, "Alternative PET tracers in head and neck cancer. A review," *Eur. Arch. Otorhinolaryngol.*, vol. 270, pp. 2595–2601, 2013.

[34] D. M. Brizel, G. S. Sibley, L. R. Prosnitz, R. L. Scher, M. W. Dewhirst, and D. Ph, "Tumor hypoxia adversely affects the prognosis of carcinoma of the Head and Neck," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 38, no. 2, pp. 285–289, 1997.

[35] W.-J. Koh, J. S. Rasey, J. R. Evans, Margaret L. Grierson, K. A. Lewellen, Thomas K. Graham, Michael M. Krohn, and T. W. Griffin, "Imaging of hypoxia in human tumors with [F-18]fluoromisonidazole," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 22, no. 1, pp. 199–212, 1992.

[36] M. Souvatzoglou, A. L. Grosu, B. Röper, B. J. Krause, R. Beck, G. Reischl, M. Picchio, H.-J. Machulla, H.-J. Wester, and M. Piert, "Tumour hypoxia imaging with [18F]FAZA PET in head and neck cancer patients: a pilot study," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 34, no. 10, pp. 1566–1575, Apr. 2007.

[37] L. Chen, Z. Zhang, and Y. Guan, "18F-HX4 hypoxia imaging with PET/CT in head and neck cancer: a comparison with 18F-FMISO," *Nucl. Med. Commun.*, vol. 33, no. 10, pp. 1096–1102, 2012.

[38] Y. Fujibayashi, H. Taniuchi, Y. Yonekura, H. Ohtani, J. Konishi, and A. Yokoyama, “Copper-62-ATSM : A New Hypoxia Imaging Agent with High Membrane Permeability and,” *J. Nucl. Med.*, vol. 38, pp. 1155–1160, 1997.

[39] N. J. Roland and V. (eds) Paleri, “Head and Neck Cancer : Multidisciplinary Management Guidelines, 4th Edition,” London, 2011.

[40] E. M. Rohren, T. G. Turkington, and R. E. Coleman, “Clinical Applications of PET in Oncology,” *Radiology*, vol. 231, pp. 305–332, 2004.

[41] H. Zaidi and I. El Naqa, “PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques.,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 37, no. 11, pp. 2165–87, Nov. 2010.

[42] I. Ciernik, E. Dizendorf, B. Baumert, B. Reiner, C. Burger, J. Davis, U. Lutolf, H. Steinert, and G. Vonschulthess, “Radiation treatment planning with an integrated positron emission and computer tomography (PET/CT): a feasibility study,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 57, no. 3, pp. 853–863, Nov. 2003.

[43] M. MacManus, U. Nestle, K. E. Rosenzweig, I. Carrio, C. Messa, O. Belohlavek, M. Danna, T. Inoue, E. Deniaud-Alexandre, S. Schipani, N. Watanabe, M. Dondi, and B. Jeremic, “Use of PET and PET/CT for radiation therapy planning: IAEA expert report 2006-2007.,” *Radiother. Oncol.*, vol. 91, no. 1, pp. 85–94, Apr. 2009.

[44] T. Leong, C. Everitt, K. Yuen, S. Condron, A. Hui, S. Y. K. Ngan, A. Pitman, E. W. F. Lau, M. MacManus, D. Binns, T. Ackerly, and R. J. Hicks, “A prospective study to evaluate the impact of FDG-PET on CT-based radiotherapy treatment planning for oesophageal cancer.,” *Radiother. Oncol.*, vol. 78, no. 3, pp. 254–61, Mar. 2006.

[45] I. El Naqa, P. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, W. Thorstad, and J. O. Deasy, “Exploring feature-based approaches in PET images for predicting cancer treatment outcomes.,” *Pattern Recognit.*, vol. 42, no. 6, pp. 1162–1171, Jun. 2009.

[46] F. Tixier, M. Hatt, C. Valla, V. Fleury, C. Lamour, S. Ezzouhri, P. Ingrand, R. Perdrisot, D. Visvikis, and C. C. Le Rest, “Visual Versus Quantitative Assessment of Intratumor 18F-FDG PET Uptake Heterogeneity: Prognostic Value in Non-Small Cell Lung Cancer.,” *J. Nucl. Med.*, vol. 55, no. 8, pp. 1235–1241, Jun. 2014.

[47] T. Nishioka, T. Shiga, H. Shirato, E. Tsukamoto, K. Tsuchiya, T. Kato, K. Ohmori, A. Yamazaki, H. Aoyama, S. Hashimoto, T.-C. Chang, and K. Miyasaka, “Image Fusion Between 18 FDG-PETAnd MRI/CT For Radiotherapy Planning Of Oropharyngeal And Nasopharyngeal Carcinomas,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 53, no. 4, pp. 1051–1057, 2002.

[48] D. Wang, C. Schultz, and P. Jursinic, “initial experience of FDG-PET/CT guided IMRT of Head and Neck carcinoma,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 65, pp. 143–151, 2006.

[49] A. Guido, L. Fuccio, B. Rombi, P. Castellucci, A. Cecconi, F. Bunkheila, C. Fuccio, E. Spezi, A. L. Angelini, and E. Barbieri, “Combined 18F-FDG-PET/CT Imaging in Radiotherapy Target Delineation for Head-and-Neck Cancer,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 73, no. 3, pp. 759–763, Mar. 2009.

[50] R. J. H. M. et al. Steenbakkers, “Reduction of observer variation using matched CT-PET for lung cancer delineation: A three-dimensional analysis,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 64, no. 2, pp. 435–448, Feb. 2006.

[51] D. A. X. Schinagl, W. V Vogel, A. L. Hoffmann, J. A. van Dalen, W. J. Oyen, and J. H. A. M. Kaanders, “Comparison of Five Segmentation Tools for 18F-Fluoro-Deoxy-Glucose–Positron Emission Tomography–Based Target Volume Definition in Head and Neck Cancer,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 69, no. 4, pp. 1282–1289, Nov. 2007.

[52] C. B. Caldwell, K. Mah, Y. C. Ung, C. E. Danjoux, J. M. Balogh, S. N. Ganguli, and L. E. Ehrlich, “Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: the impact of 18FDG-hybrid PET fusion.,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 51, no. 4, pp. 923–31, Nov. 2001.

[53] A. C. Riegel, A. M. Berson, S. Destian, T. Ng, L. B. Tena, R. J. Mitnick, and P. S. Wong, “Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion,” *Int. J. Radiat. Oncol.*, vol. 65, no. 3, pp. 726–732, Jul. 2006.

[54] J. B. Davis, B. Reiner, M. Huser, C. Burger, G. Székely, and I. F. Ciernik, “Assessment of 18F PET signals for automatic target volume definition in radiotherapy treatment planning.,” *Radiother. Oncol.*, vol. 80, no. 1, pp. 43–50, Jul. 2006.

[55] L. Deantonio, D. Beldì, G. Gambaro, G. Loi, M. Brambilla, E. Inglese, and M. Krengli, “FDG-PET/CT imaging for staging and radiotherapy treatment planning of head and neck carcinoma.,” *Radiat. Oncol.*, vol. 3, p. 29, Jan. 2008.

[56] J. L. Barker, A. S. Garden, K. K. Ang, J. C. O’Daniel, H. Wang, L. E. Court, W. H. Morrison, D. I. Rosenthal, K. S. C. Chao, S. L. Tucker, R. Mohan, and L. Dong, “Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system.,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 59, no. 4, pp. 960–70, Jul. 2004.

[57] M. R. Vernon, C. J. Maheshwari, Mohit Schultz, M. A. Michel, B. H. Wong, Stuart J. , Campbell, B. L. Massey, F. Wilson, and D. Wang, “Clinical Outcomes of Patients Receiving Integrated PET/CT-Guided Radiotherapy for Head and Neck Carcinoma,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 70, no. 3, pp. 678–684, 2008.

[58] D. E. Heron, R. S. Andrade, J. Flickinger, J. Johnson, S. S. Agarwala, A. Wu, S. Kalnicki, and N. Avril, “Hybrid PET-CT simulation for radiation treatment planning in head-and-neck cancers: a brief technical report.,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 60, no. 5, pp. 1419–24, Dec. 2004.

[59] K. L. Newbold, M. Partridge, G. Cook, B. Sharma, P. Rhys-Evans, K. J. Harrington, and C. M. Nutting, “Evaluation of the role of 18FDG-PET/CT in radiotherapy target definition in patients with head and neck cancer.,” *Acta Oncol.*, vol. 47, no. 7, pp. 1229–36, Jan. 2008.

[60] D. Thorwarth, X. Geets, and M. Paiusco, "Physical radiotherapy treatment planning based on functional PET/CT data.," *Radiother. Oncol.*, vol. 96, no. 3, pp. 317–24, Sep. 2010.

[61] S. S. Gambhir, J. Czernin, J. Schwimmer, D. H. Silverman, R. E. Coleman, and M. E. Phelps, "A tabulated summary of the FDG PET literature.," *J. Nucl. Med.*, vol. 42, no. 5 Suppl, p. 1S–93S, May 2001.

[62] C. Laubenbacher, D. Saumweber, C. Wagner-Manslau, R. J. Kau, M. Herz, N. Avril, S. Ziegler, C. Kruschke, W. Arnold, and M. Schwaiger, "Comparison of fluorine-18-fluorodeoxyglucose PET, MRI and endoscopy for staging head and neck squamous-cell carcinomas.," *J. Nucl. Med.*, vol. 36, no. 10, pp. 1747–57, Oct. 1995.

[63] K. A. Miles, M. R. Griffiths, and C. J. Keith, "Blood flow-metabolic relationships are dependent on tumour size in non-small cell lung cancer: a study using quantitative contrast-enhanced computer tomography and positron emission tomography.," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 33, no. 1, pp. 22–8, Jan. 2006.

[64] E. M. Marom, T. a Aloia, M. B. Moore, M. Hara, J. E. Herndon, D. H. Harpole, P. C. Goodman, and E. F. Patz, "Correlation of FDG-PET imaging with Glut-1 and Glut-3 expression in early-stage non-small cell lung cancer.," *Lung Cancer*, vol. 33, no. 2–3, pp. 99–107, 2001.

[65] K. Kubota, "From tumor biology to clinical Pet: a review of positron emission tomography (PET) in oncology.," *Ann. Nucl. Med.*, vol. 15, no. 6, pp. 471–86, Dec. 2001.

[66] T. M. Blodgett, M. B. Fukui, C. H. Snyderman, B. F. Branstetter, B. M. McCook, W. Dave, and C. C. Meltzer, "Combined PET-CT in the Head and Neck. Part1. Physiologic, Altered Physiologic and Artifactual FDG Uptake," *RadioGraphics*, vol. 25, pp. 897–912, 2005.

[67] J.-F. Daisne, M. Sibomana, A. Bol, T. Doumont, M. Lonneux, and V. Grégoire, "Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms," *Radiother. Oncol.*, vol. 69, no. 3, pp. 247–250, Dec. 2003.

[68] J. Schaefferkoetter, M. Casey, D. Townsend, and G. El Fakhri, “Clinical impact of time-of-flight and point response modeling in PET reconstructions: a lesion detection study.,” *Phys. Med. Biol.*, vol. 58, no. 5, pp. 1465–78, Mar. 2013.

[69] M. Soret, S. L. Bacharach, and I. Buvat, “Partial-volume effect in PET tumor imaging.,” *J. Nucl. Med.*, vol. 48, no. 6, pp. 932–45, Jun. 2007.

[70] F. Hofheinz, C. Potzsch, L. Oehme, B. Beuthien-Baumann, J. Steinbach, J. Kotzerke, and J. van den Hoff, “Automatic volume delineation in oncological PET. Evaluation of a dedicated software tool and comparison with manual delineation in clinical data sets,” *Nuklearmedizin*, vol. 51, no. 1, pp. 9–16, 2012.

[71] J. A. Lee, “Segmentation of positron emission tomography images: Some recommendations for target delineation in radiation oncology.,” *Radiother. Oncol.*, vol. 96, no. 3, pp. 302–307, Sep. 2010.

[72] S. Belhassen and H. Zaidi, “A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET,” *Med. Phys.*, vol. 37, no. 3, pp. 1309–1324, 2010.

[73] M. Hatt, J. Lee, C. R. Schmidtlein, I. El Naqa, C. Caldwell, E. De Bernardi, W. Lu, S. Das, X. Geets, V. Gregoire, R. Jeraj, M. MacManus, O. Mawlawi, U. Nestle, A. Pugachev, H. Schöder, T. Shepherd, E. Spezi, D. Visvikis, H. Zaidi, and A. S. Kirov, “Report of AAPM TG211: Classification and evaluation strategies of auto-segmentation approaches for PET,” *under Rev. with Med. Phys.*, 2014.

[74] U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welsch, D. Hellwig, and C. Ru, “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.,” *J. Nucl. Med.*, vol. 46, no. 8, pp. 1342–1348, 2005.

[75] W. Jentzen, L. Freudenberg, E. G. Eising, M. Heinze, W. Brandau, and A. Bockisch, "Segmentation of PET volumes by iterative image thresholding.," *J. Nucl. Med.*, vol. 48, no. 1, pp. 108–14, Jan. 2007.

[76] V. Gregoire, J.-F. Daisne, and X. Geets, "Comparison of CT- and FDG-PET-defined GT: in regard to Paulino et al. (Int J Radiat Oncol Biol Phys 2005;61:1385-1392).," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 63, no. 1, pp. 308–9; author reply 309, Sep. 2005.

[77] E. C. Ford, P. E. Kinahan, L. Hanlon, A. Alessio, J. Rajendran, D. L. Schwartz, and M. Phillips, "Tumor delineation using PET in head and neck cancers: Threshold contouring and lesion volumes," *Med. Phys.*, vol. 33, no. 11, pp. 4280–4288, 2006.

[78] M. A. Lodge, M. A. Chaudhry, and R. L. Wahl, "Noise considerations for PET quantification using maximum and peak standardized uptake value.," *J. Nucl. Med.*, vol. 53, no. 7, pp. 1041–7, Jul. 2012.

[79] M. Hatt, C. Cheze-Le Rest, E. O. Aboagye, L. M. Kenny, L. Rosso, F. E. Turkheimer, N. M. Albarghach, J.-P. Metges, O. Pradier, and D. Visvikis, "Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements.," *J. Nucl. Med.*, vol. 51, no. 9, pp. 1368–76, Sep. 2010.

[80] L. Drever, W. Roa, A. McEwan, and D. Robinson, "Iterative threshold segmentation for PET target volume delineation," *Med. Phys.*, vol. 34, no. 4, pp. 1253–1265, 2007.

[81] Q. C. Black, I. S. Grills, L. L. Kestin, C.-Y. O. Wong, J. W. Wong, A. a Martinez, and D. Yan, "Defining a radiotherapy target with positron emission tomography.," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 60, no. 4, pp. 1272–82, Nov. 2004.

[82] U. Nestle, S. Kremp, and A.-L. Grosu, "Practical integration of [18F]-FDG-PET and PET-CT in the planning of radiotherapy for non-small cell lung cancer (NSCLC): the technical basis, ICRU-target volumes, problems, perspectives.," *Radiother. Oncol.*, vol. 81, no. 2, pp. 209–25, Nov. 2006.

[83] A. Schaefer, S. Kremp, D. Hellwig, C. Rübe, C.-M. Kirsch, and U. Nestle, "A contrast-oriented algorithm for FDG-PET-based delineation of tumour volumes for the radiotherapy of lung cancer: derivation from phantom measurements and validation in patient data.," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 35, no. 11, pp. 1989–99, Nov. 2008.

[84] F. Hofheinz, S. Dittrich, C. Potzsch, and J. van den Hoff, "Effects of cold sphere walls in PET phantom measurements on the volume reproducing threshold," *Phys. Med. Biol.*, vol. 55, no. 4, pp. 1099–1113, 2010.

[85] J. A. van Dalen, A. L. Hoffmann, V. Dicken, W. V Vogel, B. Wiering, T. J. Ruers, N. Karssemeijer, and W. J. G. Oyen, "A novel iterative method for lesion delineation and volumetric quantification with FDG PET," *Nucl. Med. Commun.*, vol. 28, no. 6, pp. 485–493, 2007.

[86] M. Ollers, G. Bosmans, A. van Baardwijk, A. Dekker, P. Lambin, J. Teule, W. Thimister, A. Rhamy, and D. De Ruysscher, "The integration of PET-CT scans from different hospitals into radiotherapy treatment planning.," *Radiother. Oncol.*, vol. 87, no. 1, pp. 142–6, Apr. 2008.

[87] S. A. Nehmeh, H. El-Zeftawy, C. Greco, J. Schwartz, Y. E. Erdi, A. Kirov, C. R. Schmidtlein, A. B. Gyau, S. M. Larson, and J. L. Humm, "An iterative technique to segment PET lesions using a Monte Carlo based mathematical model," *Med. Phys.*, vol. 36, no. 10, p. 4803, 2009.

[88] H. Li, W. L. Thorstad, K. J. Biehl, R. Laforest, Y. Su, K. I. Shoghi, E. D. Donnelly, D. a. Low, and W. Lu, "A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours," *Med. Phys.*, vol. 35, no. 8, p. 3711, 2008.

[89] E. Day, J. Betler, D. Parda, B. Reitz, A. Kirichenko, S. Mohammadi, and M. Miften, "A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients," *Med. Phys.*, vol. 36, no. 10, pp. 4349–4358, 2009.

[90] D. W. G. Montgomery, A. Amira, and H. Zaidi, “Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model,” *Med. Phys.*, vol. 34, no. 2, pp. 722–736, 2007.

[91] M. Hatt, C. Cheze, A. Turzo, C. Roux, and I. T. Oncology, “A Fuzzy Locally Advanced Bayesian Segmentation Approach for Volume Determination in PET,” *IEEE Trans. Med. Imaging*, vol. 28, no. 6, pp. 881–893, 2009.

[92] M. Aristophanous, B. C. Penney, M. K. Martel, and C. A. Pelizzari, “A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography,” *Med. Phys.*, vol. 34, no. 11, p. 4223, 2007.

[93] E. De Bernardi, C. Soffientini, F. Zito, and G. Baselli, “Joint segmentation and quantification of oncological lesions in PET/CT: Preliminary evaluation on a zeolite phantom,” in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, 2012, no. 1, pp. 3306–3310.

[94] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[95] H. Yu, C. Caldwell, K. Mah, and D. Mozeg, “Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning.,” *IEEE Trans. Med. Imaging*, vol. 28, no. 3, pp. 374–83, Mar. 2009.

[96] H. Yu, C. Caldwell, K. Mah, I. Poon, J. Balogh, R. MacKenzie, N. Khaouam, and R. Tirona, “Automated Radiation Targeting in Head-and-Neck Cancer Using Region-Based Texture Analysis of PET and CT Images,” *Int. J. Radiat. Oncol.*, vol. 75, no. 2, pp. 618–625, 2009.

[97] D. Markel, C. Caldwell, H. Alasti, H. Soliman, Y. Ung, J. Lee, and A. Sun, “Clinical Study Automatic Segmentation of Lung Carcinoma Using 3D Texture Features in 18-FDG PET / CT,” *Int. J. Mol. Imaging*, vol. 980769, 2013.

[98] X. Geets, J. a Lee, A. Bol, M. Lonneux, and V. Grégoire, “A gradient-based method for segmenting FDG-PET images: methodology and validation.,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 34, no. 9, pp. 1427–38, Sep. 2007.

[99] Y. Lee, S. Song, J. Lee, and M. Kim, “Tumor Segmentation from SmaII Animal PET Using Region Growing based on Gradient Magnitude,” in *Proceedings of 7th International Workshop on Enterprise networking and Computing in Healthcare Industry 2005. HEALTHCOM. IEEE*, 2005, pp. 243–247.

[100] P. Tylski, G. Bonniaud, E. Decencière, J. Stawiaski, J. Coulot, and D. Lefkopoulos, “F-FDG PET images segmentation using morphological watershed : a phantom study,” *Nucl. Sci. Symp. Conf. Rec. 2006, IEEE*, vol. 4. IEEE, pp. 2063–2067, 2006.

[101] S. Ray, R. Hagge, M. Gillen, M. Cerejo, S. Shakeri, L. Beckett, T. Greasby, and R. D. Badawi, “Comparison of two-dimensional and three-dimensional iterative watershed segmentation methods in hepatic tumor volumetrics,” *Med. Phys.*, vol. 35, no. 12, pp. 5869–5881, 2008.

[102] C.-Y. Hsu, C.-Y. Liu, and C.-M. Chen, “Automatic segmentation of liver PET images.,” *Comput. Med. Imaging Graph.*, vol. 32, no. 7, pp. 601–10, Oct. 2008.

[103] A. Kanakatte, J. Gubbi, B. Srinivasan, N. Mani, T. Kron, D. Binns, and M. Palaniswami, “Pulmonary tumor volume delineation in PET images using deformable models.,” *CER*

[104] M. S. Sharif, M. Abbod, A. Amira, and H. Zaidi, “Artificial Neural Network-Based System for PET Volume Segmentation.,” *Int. J. Biomed. Imaging*, vol. 2010, Jan. 2010.

[105] J. Fiot, L. D. Cohen, P. Raniga, J. Fripp, U. M. R. Cnrs, and U. Paris, “Efficient brain lesion segmentation using multi-modality tissue-based feature selection and support vector machines,” *Int. j. numer. method. biomed. eng.*, vol. 29, no. January, pp. 905–915, 2013.

[106] A. Kerhet, C. Small, H. Quon, T. Riauka, L. Schrader, R. Greiner, D. Yee, A. McEwan, and W. Roa, “Application of machine learning methodology for PET-based definition of lung cancer.,” *Curr. Oncol.*, vol. 17, no. 1, pp. 41–7, Feb. 2010.

[107] S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.,” *IEEE Trans. Med. Imaging*, vol. 23, no. 7, pp. 903–21, Jul. 2004.

[108] R. J. McGurk, J. Bowsher, J. a Lee, and S. K. Das, “Combining multiple FDG-PET radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods.,” *Med. Phys.*, vol. 40, no. 4, pp. 1–9, Apr. 2013.

[109] A.-S. Dewalle-Vignon, N. Betrouni, R. Lopes, D. Huglo, S. Stute, and D. Vermandel, “A new method for volume segmentation of PET images, based on possibility theory,” *IEEE Trans. Med. Imaging*, vol. 30, pp. 409–423, 2011.

[110] J. O. Deasy, A. I. Blanco, and V. H. Clark, “CERR: A computational environment for radiotherapy research,” *Med. Phys.*, vol. 30, no. 5, pp. 979–985, 2003.

[111] I. Bankman, *Handbook of Medical Image Processing and Analysis. Part II. Segmentation*. Baltimore, MD, USA: Academic Press Inc., 2000, pp. 71–258.

[112] P. E. Danielsson and O. , Seger, “Generalized and Separable Sobel Operators,” in *"Machine vision for three-dimensional scenes*, Academic Press, H. Freeman, Ed. San Diego, 1990, pp. 347 – 380.

[113] C. Ballangan, X. Wang, M. Fulham, S. Eberl, Y. Yin, and D. Feng, “Automated delineation of lung tumors in PET images based on monotonicity and a tumor-customized criterion.,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 5, pp. 691–702, Sep. 2011.

[114] M. Sussman, P. Smereka, and S. Osher, “A Level Set Approach for Computing Solutions to Incompressible Two-Phase Flow,” *J. Comput. Phys.*, vol. 114, no. 1, pp. 146–159, 1994.

[115] T. F. Chan and L. a Vese, “Active contours without edges.,” *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–77, Jan. 2001.

[116] S. Lankton and A. Tannenbaum, “Localizing region-based active contours.,” *IEEE Trans. Image Process.*, vol. 17, no. 11, pp. 2029–39, Nov. 2008.

[117] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. C. Tempany, M. R. Kaus, S. J. Haker, W. M. W. Iii, and F. A. Jolesz, “Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index,” *Acad. Radiol.*, vol. 11, no. 2, pp. 178–189, 2004.

[118] M.-P. Dubuisson, A. K. Jain, E. Lansing, and A. B. B, “A Modified Hausdorff Distance for Object Matching,” *Pattern Recognition, IEEE*, vol. 1 -Conf. A, pp. 566–568, 1994.

[119] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, “Comparing images using the Hausdorff Distance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–862, 1993.

[120] L. a Drever, W. Roa, A. McEwan, and D. Robinson, “Comparison of three image segmentation techniques for target volume delineation in positron emission tomography.,” *J. Appl. Clin. Med. Phys.*, vol. 8, no. 2, pp. 93–109, Jan. 2007.

[121] I. El Naqa, D. Yang, A. Apte, D. Khullar, S. Mutic, J. Zheng, J. D. Bradley, P. Grigsby, and J. O. Deasy, “Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning,” *Med. Phys.*, vol. 34, no. 12, p. 4738, 2007.

[122] O. Rousset, A. Rahmim, A. Alavi, and H. Zaidi, “Partial Volume Correction Strategies in PET,” *Positron Emiss. Tomogr. Clin.*, vol. 2, no. 2, pp. 235–249, 2007.

[123] B.-K. Teo, Y. Seo, S. L. Bacharach, J. a Carrasquillo, S. K. Libutti, H. Shukla, B. H. Hasegawa, R. a Hawkins, and B. L. Franc, “Partial-volume correction in PET: validation of an iterative postreconstruction method with phantom and patient data.,” *J. Nucl. Med.*, vol. 48, no. 5, pp. 802–10, May 2007.

[124] E. De Bernardi, E. Faggiano, F. Zito, P. Gerundini, and G. Baselli, “Lesion quantification in oncological positron emission tomography: A maximum likelihood partial volume correction strategy,” *Med. Phys.*, vol. 36, no. 7, p. 3040, 2009.

[125] J. K. Udupa, V. R. Leblanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn, “A framework for evaluating image segmentation algorithms.,” *Comput. Med. Imaging Graph.*, vol. 30, no. 2, pp. 75–87, Mar. 2006.

[126] S. L. Breen, J. Publicover, S. De Silva, G. Pond, K. Brock, B. O’Sullivan, B. Cummings, L. Dawson, A. Keller, J. Kim, J. Ringash, E. Yu, A. Hendler, and J. Waldron, “Intraobserver and interobserver variability in GTV delineation on FDG-PET-CT images of head and neck cancers,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 68, no. 3, pp. 763–70, 2007.

[127] P. Tylski, S. Stute, N. Grotus, K. Doyeux, S. Hapdey, I. Gardin, B. Vanderlinden, and I. Buvat, “Comparative assessment of methods for estimating tumor volume and standardized uptake value in (18)F-FDG PET.,” *J. Nucl. Med.*, vol. 51, no. 2, pp. 268–76, Feb. 2010.

[128] E. Prieto, P. Lecumberri, M. Pagola, M. Gómez, I. Bilbao, M. Ecay, I. Peñuelas, and J. M. Martí-Climent, “Twelve automated thresholding methods for segmentation of PET images: a phantom study.,” *Phys. Med. Biol.*, vol. 57, no. 12, pp. 3963–3980, May 2012.

[129] J. J. Hamill, M. E. Casey, V. Rappoport, and C. C. Watson, “Characteristics of PET/CT images of the ACR PET phantom,” in *IEEE Nuclear Science Symposium Conference Record*, 2004, vol. 5, pp. 3347–3351.

[130] M. Bazañez-Borgert, R. a. Bundschuh, M. Herz, M.-J. Martínez, M. Schwaiger, and S. I. Ziegler, “Radioactive spheres without inactive wall for lesion simulation in PET,” *Z. Med. Phys.*, vol. 18, no. 1, pp. 37–42, Mar. 2008.

[131] B. Berthon, C. Marshall, A. Edwards, M. Evans, and E. Spezi, “Influence of cold walls on PET image quantification and volume segmentation.,” *Med. Phys.*, vol. 40, no. 8, pp. 1–13, 2013.

[132] L. Delestienne, “Discovery TM PET-CT 690 NEMA Test Procedures. Report Direction 5330685-1EN,” *GE Healthcare*. General Electric Healthcare, Waukesha, USA, pp. 1–49, 2008.

[133] P. J. Julyan, J. H. Taylor, D. L. Hastings, H. A. Williams, and J. Zweit, “SUVpeak: a new parameter for quantification of uptake in FDG PET,” *Nucl. Med. Commun.*, vol. 25, no. 4, 2004.

[134] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, “From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors,” *J. Nucl. Med.*, vol. 50, no. Suppl 1, pp. 1–50, 2009.

[135] T. G. Turkington, T. R. Degrado, and W. H. Sampson, “Small spheres for lesion detection phantoms,” in *IEEE Nucl. Sci. Symp. Conf. Rec.*, 2001, vol. 4, pp. 2234–2237.

[136] F. Zito, E. De Bernardi, C. Soffientini, C. Canzi, R. Casati, P. Gerundini, and G. Baselli, “The use of zeolites to generate PET phantoms for the validation of quantification strategies in oncology.,” *Med. Phys.*, vol. 39, no. 9, pp. 5353–61, Sep. 2012.

[137] B. Berthon, C. Marshall, M. Evans, and E. Spezi, “Evaluation of advanced automatic PET segmentation methods using non-spherical thin-wall inserts,” *Med. Phys.*, vol. 41, no. 2, p. 022502, 2014.

[138] M. Hatt, C. Cheze le Rest, P. Descourt, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, O. Pradier, and D. Visvikis, “Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications.,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 77, no. 1, pp. 301–8, May 2010.

[139] X. Xie, “Active contouring based on gradient vector interaction and constrained level set diffusion.,” *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 154–64, Jan. 2010.

[140] S. A. Larsson, C. Jonsson, M. Pagani, L. Johansson, and H. Jacobsson, “A novel phantom design for emission tomography enabling scatter- and attenuation-‘free’ single-photon emission tomography imaging,” *Eur. J. Nucl. Med.*, vol. 27, no. 2, pp. 131–139, 2000.

[141] K. J. Van Laere, J. Versijpt, M. Koole, S. Vandenberghe, P. Lahorte, I. Lemahieu, and R. a Dierckx, “Experimental performance assessment of SPM for SPECT neuroactivation studies using a subresolution sandwich phantom design.,” *Neuroimage*, vol. 16, no. 1, pp. 200–16, May 2002.

[142] H. El-Ali, M. Ljungberg, S.-E. Strand, J. Palmer, L. Malmgren, and J. Nilsson, “Calibration of a radioactive ink-based stack phantom and its applications in nuclear medicine.,” *Cancer Biother. Radiopharm.*, vol. 18, no. 2, pp. 201–7, Apr. 2003.

[143] J. a van Staden, H. du Raan, M. G. Lötter, a van Aswegen, and C. P. Herbst, “Production of radioactive quality assurance phantoms using a standard inkjet printer.,” *Phys. Med. Biol.*, vol. 52, no. 15, pp. N329–37, Aug. 2007.

[144] V. Sossi, K. R. Buckley, P. Piccioni, A. Rahmim, S. Member, M. Camborde, E. Strome, S. Lapi, and T. J. Ruth, “Printed Sources for Positron Emission Tomography,” *IEEE Trans. Nucl. Sci.*, vol. 52, no. 1, pp. 114–118, 2005.

[145] P. J. Markiewicz, G. I. Angelis, F. Kotasidis, M. Green, W. R. Lionheart, a J. Reader, and J. C. Matthews, “A custom-built PET phantom design for quantitative imaging of printed distributions.,” *Phys. Med. Biol.*, vol. 56, no. 21, pp. N247–61, Nov. 2011.

[146] R. B. Holmes, S. M. a Hoffman, and P. M. Kemp, “Generation of realistic HMPAO SPECT images using a subresolution sandwich phantom.,” *Neuroimage*, vol. 81, pp. 8–14, Nov. 2013.

[147] M. a. Miller and G. D. Hutchins, “Development of anatomically realistic PET and PET/CT phantoms with rapid prototyping technology,” *2007 IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 4252–4257, 2007.

[148] A. Le Maitre, M. Hatt, O. Pradier, C. Cheze-le Rest, and D. Visvikis, “Impact of the accuracy of automatic tumour functional volume delineation on radiotherapy treatment planning.,” *Phys. Med. Biol.*, vol. 57, no. 17, pp. 5381–97, Sep. 2012.

[149] S. S. Chen and P. S. Gopalakrishnan, “Clustering via the Bayesian information criterion with applications in speech recognition,” *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. IEEE*, vol. 2, pp. 645–648, 1998.

[150] J. de Bresser, K. L. Vincken, A. J. Kaspers, G. J. E. Rinkel, M. a Viergever, and G. J. Biessels, “Quantification of cerebral volumes on MRI 6 months after aneurysmal subarachnoid hemorrhage.,” *Stroke.*, vol. 43, no. 10, pp. 2782–4, Oct. 2012.

[151] M. Lyksborg, R. Larsen, M. Soelberg Sørensen, Per Blinkenberg, E. Garde, H. R. Siebner, and T. B. Dyrby, “Segmenting Multiple Sclerosis Lesions Using a Spatially Constrained K-Nearest Neighbour Approach,” *Image Anal. Recognit. Lect. Notes Comput. Sci.*, vol. 7325, pp. 156–163, 2012.

[152] G. Iordanescu, P. N. Venkatasubramanian, and A. M. Wyrwicz, “Automatic segmentation of amyloid plaques in MR images using unsupervised support vector machines.,” *Magn. Reson. Med.*, vol. 67, no. 6, pp. 1794–802, Jun. 2012.

[153] A. Jayachandran and R. Dhanasekaran, “Brain Tumor Detection and Classification of MR Images Using Texture Features and Fuzzy SVM Classifier,” *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 12, pp. 2264–2269, 2013.

[154] C. C. Reyes-aldasoro, “Image Segmentation with Kohonen Neural Network,” in *International Conference on Telecommunications*, 2000.

[155] M. Tabakov and P. Kozak, “Segmentation of histopathology HER2/neu images with fuzzy decision tree and Takagi-Sugeno reasoning.,” *Comput. Biol. Med.*, vol. 49, pp. 19–29, Jun. 2014.

[156] A. E. Hassanien and T. Kim, “Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks,” *J. Appl. Log.*, vol. 10, no. 4, pp. 277–284, Dec. 2012.

[157] T. Shepherd, B. Berthon, P. Galavis, E. Spezi, A. Apte, J. A. Lee, D. Visvikis, M. Hatt, E. De Bernardi, S. Das, I. El Naqa, and U. Nestle, “Design of a Benchmark Platform for Evaluating PET-based Contouring Accuracy in Oncology Applications,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 39, p. S264, 2012.

[158] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Advanced B. Monterey, CA: Wadsworth & Brooks/Cole, 1984.

[159] F. Tixier, M. Hatt, C. Valla, D. Visvikis, and C. Cheze le Rest, “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*, vol. 40, no. 2, p. S210, 2013.

[160] L. Dice, “Measures of the amount of ecologic association between species.,” *Ecology*, vol. 26, pp. 297–302, 1945.

[161] T. Shepherd, M. Teras, R. Beichel, R. Boellaard, M. Bruynooghe, V. Dicken, M. Gooding, P. Julyan, J. Lee, and S. Lefevre, “Comparative Study with New Accuracy Metrics for Target Volume Contouring in PET Image Guided Radiation Therapy.,” *IEEE Trans. Med. Imaging*, vol. 31, no. 11, pp. 2006–2024, Jun. 2012.

[162] B. Berthon, E. Spezi, C. R. Schmidtlein, A. Apte, P. Galavis, H. Zaidi, E. De Bernardi, J. A. Lee, A. Kirov, and inc. Other contributors, “Development of a software platform for evaluating automatic PET segmentation methods: the PET-ASset,” *Radiother. Oncol.*, vol. 111, no. Suppl.1, p. 177, 2014.

[163] B. Berthon, I. Häggström, A. Apte, B. J. Beattie, A. S. Kirov, J. L. Humm, C. Marshall, E. Spezi, A. Larsson, and C. R. Schmidtlein, “PETSTEP: Generation of Synthetic PET Lesions for Fast Evaluation of Segmentation Methods,” *Phys. Med. Biol.*

Glossary

AAPM:	American Association of Physicists in Medicine
AC:	Active Contouring segmentation method
AD:	Apparent Diameter
AR:	Aspect ratio: ratio between the largest diameter and smallest perpendicular diameter
AT:	Automatic Thresholding segmentation method
ATLAAS:	Advanced decision Tree-based Learning Algorithm for Automatic Segmentation
ANN:	Artificial Neural Networks
BTV:	Biological Tumour Volume
CECT:	Contrast Enhanced Computed Tomography
CERR:	open source software: Computational Environment for Radiotherapy Research
CT:	Computed Tomography
COV:	Coefficient Of Variation
CRT:	Concomitant Radio-Chemotherapy
CTV:	Clinical Treatment Volume
EGFR:	Epidermal Growth Factor Receptor
FCM:	Fuzzy Clustering Method
FDG:	2-deoxy-2-[^{18}F]Fluoro-D-Glucose
FLAB:	Fuzzy Locally Adaptive Bayesian segmentation method
FN:	False Negatives
FP:	False Positives
GC:	gradient-based contouring method
GCM:	Gaussian mixture models clustering method
GSO:	Gadolinium orthosilicate
GTV:	Gross Target Volume
GUI:	Graphical User Interface
HMPAO:	Technetium (^{99m}Tc) exametazime or Hexamethylpropyleneamine, radiotracer used for brain perfusion SPECT imaging
H&N:	Head And Neck
HU:	Hounsfield Unit
IAEA:	International Atomic Energy Agency
IEC:	International Electrotechnical Commission
IC:	Induction Chemotherapy
IGRT:	Image Guided Radiotherapy Treatment
IMRT:	Intensity Modulated Radiotherapy Treatment
KM:	K-nearest neighbour clustering Method
LSO:	Lutetium Orthosilicate
ML OSEM:	Maximum Likelihood Ordered Subset Estimation Maximisation
MRI:	Magnetic Resonance Imaging
NEMA:	National Electrical Manufacturers Association
NHS:	National Health Services
OAR:	Organ At Risk
PACS:	Pictures Archiving and Communications System
PET:	Positron Emission Tomography
PET-AS:	PET Auto-Segmentation method
PETIC:	Wales research and diagnostic PET Imaging Centre

PETSTEP:	Positron Emission Tomography Simulator of Tracers via Emission Projection
PMMA:	Polymethyl Methacrylate
PMT:	Photomultiplier Tube
PSF:	Point Spread Function
PTV:	Planning Target Volume
RC:	Recovery Coefficient
RG:	Region–Growing Method
ROI:	Region Of Interest
RT:	Radiotherapy Treatment
RTP:	Radiotherapy Treatment Planning
SCC:	Squamous Cell Carcinoma
SD:	Standard Deviation
SPECT:	Single Photon Emission Computed Tomography
SS:	Sub-resolution Sandwich (phantom)
STAPLE:	Simultaneous Truth And Performance Level Estimation
SUV:	Standardised Uptake Value
TBR:	Tumour-to-Background Ratio
TG211:	Task Group 211
TLC:	Thin Layer Chromatography
TNM:	Tumour, Nodes, Metastases (cancer stage classification system)
TOF:	Time-Of-Flight
TP:	True Positives
UK:	United Kingdom
VCC:	Velindre Cancer Centre
VOI:	Volume Of Interest
WT:	Watershed Transform or Watershed Transform-based segmentation method
3D:	Three-dimensional
2D:	Two-dimensional

Appendix. Published work

Research articles

- B. Berthon, R. Holmes, C. Marshall, E. Spezi, "Performance of PET auto-segmentation for delineating heterogeneous lesions ", In preparation

- B. Berthon, I. Häggström, A. Apte, B. J. Beattie, A. S. Kirov, J. L. Humm, C. Marshall, E. Spezi, A. Larsson and C R. Schmidtlein, "PETSTEP: Generation of Synthetic PET Lesions for Fast Evaluation of Segmentation Methods", In preparation

- B. Berthon, E. Spezi, Galavis, A. T. Shepherd, P. Apte, M. Hatt, H. Fayad, D. Visvikis, E. De Bernardi, C. Soffientini, C. R. Schmidtlein, I. El Naqa, R. Jeraj, W. Lu, S. Das, H. Zaidi, O. Mawlawi, D. Visvikis, J. A. Lee, and A. S. Kirov, " Report of American Association of Physicists in Medicine (AAPM) task group 211 (TG211): Classification and Evaluation Strategies of the Auto- Segmentation Approaches for PET Part II: Design and Implementation of PETASset: Benchmark Evaluation Software for PET Auto-Segmentation Methods", Submitted to Medical Physics .

- B. Berthon, C. Marshall, A. Edwards, E. Spezi "Influence of cold walls on PET image quantification and volume segmentation", Med Phys. 2013, 40(8) pp 1-13

- B. Berthon, C. Marshall, M. Evans, E. Spezi, "Evaluation of advanced automatic PET segmentation methods using non-spherical thin-wall inserts", Med Phys. 2014, 41(2):022502

Oral presentations

- 26 Jun 2014, Medical Physics and Clinical Engineering in Wales Summer Meeting 2014, Cardiff All Nations Center: "Comparison of PET auto-segmentation of GTV with manual PET-CT and CT/MRI outlining in oropharyngeal cancer patients", B. Berthon, T. Rackley, C. Marshall, M. Evans, N. Cole, N. Palaniappan, V. Jayaprakasam E. Spezi

- 4-8 Apr 2014, European Society for Radiotherapy and Oncology (ESTRO) 33, Vienna: "Comparison of PET auto-segmentation of GTV with manual PET-CT and CT/MRI outlining in oropharyngeal cancer patients", B. Berthon, T. Rackley, C. Marshall, M. Evans, N. Cole, N. Palaniappan, V. Jayaprakasam E. Spezi

- 6 Jan 2014, Nuclear Medicine Seminars, Bristol Royal Infirmary, "Validation and comparison of PET auto-segmentation tools for H&N RT planning", B. Berthon, C. Marshall, M. Evans, E. Spezi

- 16 Nov 2013, Postgraduate Research Day Cardiff University Heath Park, "Performance of 18F-FDG automated segmentation methods for non-spherical objects", B. Berthon, C. Marshall, M. Evans, A. Edwards, E. Spezi

- 22 Oct 2013, European Association of Nuclear Medicine (EANM) meeting 2013, Lyon, "Performance of 18F-FDG PET automated segmentation methods for non-spherical objects", B. Berthon, C. Marshall, M. Evans, A. Edwards, E. Spezi

- 22 Oct 2013, EANM 2013, Lyon, “Evaluation of several automatic PET-segmentation algorithms for radiotherapy treatment planning in H&N using a printed subresolution sandwich phantom”, B. Berthon, R. B. Holmes, C. Marshall, S. Jayaprakasam, M. Evans, E. Spezi

- 19 Feb 2013, University College London Hospital, Institute of Nuclear Medicine Seminar, London, “Optimised automatic target volume delineation of 18F-FDG PET images”, B. Berthon

- "Optimizing PET auto-segmentation for H&N radiotherapy treatment planning", presented as part of the IPEM Travel award trip:
    - 26 Sept 2013, Velocity Medical Solutions, Atlanta, USA
    - 1 Oct 2014, John Hopkins University, Baltimore, USA
    - 4 Oct 2013, Wisconsin Institutes for Medical Reseasch Madison, USA
    - 12 Oct 2013, Memorial Sloane Kettering Cancer Center, New York, USA

- 27-31 Oct 2012, EANM 2012, Milan, “Design of a Benchmark Platform for Evaluating PET-based Contouring Accuracy in Oncology Applications”, On behalf of the AAPM TG211

- 27-31 Oct 2012, EANM 2012, Milan, “Comparison of automatic segmentation methods of 18F-FDG PET images for radiation therapy planning in H&N: a phantom study”, B. Berthon, C. Marshall, M. Evans, E. Spezi

- 11 Sept 2012, Institute of Physics and Engineering in Medicine, Biennial Radiotherapy Conference, Oxford, “Influence of cold walls on commonly used PET volume segmentation algorithms: a phantom study”, B. Berthon, C. Marshall, M. Evans, E. Spezi

- 15 June 2012, All Wales Medical Physics and Clinical Engineering Meeting, Cardiff VCC, “ A phantom validation of automatic segmentation methods of 18F-FDG PET H&N scans”, B. Berthon, C. Marshall, M. Evans, E. Spezi

Posters

- 4-8 Apr 2014, ESTRO 33, Vienna: “Development of a software platform for evaluating automatic PET segmentation methods: the PETASset ”, On behalf of the AAPM TG211

- 20-21 Oct 2013, EANM 2013, Lyon, “Implementation and optimisation of automatic 18F-FDG PET segmentation methods”**,** B. Berthon, C. Marshall, M. Evans, E. Spezi

- 6 Nov 2013, Postgraduate Research Day, Cardiff University Heath Park, “Evaluation of automatic PET segmentation algorithms for radiotherapy treatment planning in H&N using a printed subresolution sandwich phantom”, B. Berthon, R. B. Holmes, C. Marshall, E. Spezi

- 23 Sept 2013, International Conference on Medical Physics 2013, Brighton, “Validating 18F-FDG PET automated segmentation methods for clinical use in H&N”, B. Berthon, R. B. Holmes, C. Marshall, M. Evans, A. Edwards, E. Spezi

- 27-31 Oct 2012, EANM 2012, Milan, “Use of a printed subresolution sandwich phantom for simulation of FDG PET images”, B. Berthon, R. B. Holmes, C. Marshall, E. Spezi. The Alzheimer’s Disease Neuroimaging Initiative

- 27-31 Oct 2012, EANM 2012, Milan, “Quantitative effect of cold plastic walls on 18F-FDG PET images”, B. Berthon, C. Marshall, A. Edwards, E. Spezi

- 25 Oct 2012, Velindre NHS Trust 6th Annual Research and Development Conference, Cardiff, “Comparison of automatic segmentation methods of 18F-FDG PET images for radiation therapy planning in H&N: a research study”, B. Berthon, C. Marshall, M.Evans, E.Spezi