Literature Review

The reliability of quantitative thresholding methods for PET aided delineation of GTVs in Head and Neck tumours

S. Barrett[1], R. Appleyard[2]

[1]*Radiation Therapy Department, Galway Clinic, Doughiska, Galway, Ireland,* [2]*Sheffield Hallam University, Sheffield, UK*

Abstract

Introduction: PET−CT scans are commonly used for the purpose of gross tumour volume (GTV) delineation in head and neck cancers. Qualitative visual methods (QVM) are currently employed in most radiotherapy departments but these are subject to inter- and intra-observer variability. Quantitative thresholding methods which appear in the published literature are evaluated with respect to their reliability for delineation of GTVs in head and neck cancers.

Discussion: Image segmentation involves the application of a distinct value to all pixels or voxels in an image dataset. This is a complex process affected by numerous variables. Some of the following segmentation thresholds may be applied to automatically delineate specified regions. Standardised uptake value (SUV) is commonly used to apply a threshold for GTV delineation, however this leads to inappropriately large GTVs. A further common quantitative threshold is based on the maximum signal on the PET image relative to the background uptake, known as signal to background ratio (SBR). This method generates GTVs that correlate well with surgically removed tumour volumes. Applying a fixed threshold of a percentage of the maximal intensity uptake is also documented in the literature but was found to be unsuitable for the purpose of head and neck GTV contouring. Systems based on the physical features of the PET-CT images are also discussed and are found to produce very promising results.

Conclusion: A number of quantitative techniques are evaluated and currently the most suitable is found to be SBR, however even this method was not found to be entirely reliable. More promising techniques need further evaluation before they could be implemented clinically and a Radiation Oncologist or Nuclear Medicine Radiologist must still validate all GTVs produced by quantitative methods.

Correspondence to: Sarah Barrett, Radiotherapy Department, Galway Clinic, Doughishka, Galway, Ireland. Tel: 00353 86 1712277/00353 91 785646, Fax: 00353 91 785573. Email: sarahbarrett43@gmail.com

INTRODUCTION

Positron Emission Tomography (PET) combined with Computed Tomography (CT) has become an increasingly common tool in the

evaluation of head and neck cancers. Accurate delineation of the gross tumour volume (GTV) is a vital part of the radiotherapy planning process. Although its usefulness in diagnostic workup has been deemed limited by Fetcher *et al.* in 2008[1] it has been shown to have a purpose in the radiotherapy planning setting.

A 2010 review by Troost *et al.*[2] showed that the addition of PET images increased the accuracy of delineation of the GTV. They summarised a number of recent studies and concluded PET-CT to have a sensitivity of 93−100% and a specificity of 94−98% in the case of SCC nasopharyngeal tumours. They noted that for the purpose of GTV delineation there are several potential benefits of utilising PET-CT images as it can aid reduction in target volumes or identify tumour extension not visible on conventional imaging modalities but the need for an automated delineation process in order to reduce inter- and intra-observer variability was emphasised.[2]

Hojgaard and Specht (2007)[3] looked at the overall use of PET-CT in head and neck cancers and concluded that PET-CT is emerging as the imaging modality of choice in the staging of head and neck tumours. They also made the assumption based on this that it is the choice modality for GTV delineation but no mention is made of how these volumes should be generated.

PET-CT scans are performed with the patient immobilised in the radiotherapy treatment position to aid GTV delineation at the treatment planning stage. Often a qualitative visual method (QVM) of GTV delineation is employed based on the experience of Nuclear Medicine Radiologists, Radiation Oncologists and Dosimetrists. This method is highly reliant on window levels set by the observer and inter-observer variability is significant as shown by Breen *et al.* in 2007.[4] Riegel *et al.* (2006)[5] also assessed the variability of GTVs contoured in head and neck cancers. Four physicians were evaluated contouring 16 patients. They concluded that significant differences in GTV were found between observers using a QVM.

The quantitative methods described in the literature are reviewed in order to evaluate the feasibility of implementing a standardised automated process to reduce this inconsistency. These methods include the application of thresholds based on standardised uptake values[9], signal to background ratios[13] and percentage of maximal intensity[17] uptake as well as a number of promising evolving techniques more recently documented in the literature.[20,21]

DISCUSSION

Image segmentation

Before automated delineation it is first necessary to perform image segmentation. This process associates to each pixel or voxel in an image a value, thus facilitating setting a threshold for automated GTV delineation.

These values are distinct and do not overlap as discussed by Lee, 2010.[6] Hard thresholding or segmentation is the process of each pixel being assigned one definitive value. Alternatively, soft thresholding employs probability modelling methods to assign each pixel or voxel a value based on numerous calculations (Lee, 2010)[6]. Clustering is a further option, which groups similar pixels/voxels together, and this is beneficial in pixels that contain multiple values. In addition, this process can be further defined by hard or soft thresholds (Lee, 2010).[6] More complex segmentation tools look for steep gradients and spatial relationships within the image. Segmentation can be further enhanced by reducing blur, which is an issue in PET images with poor spatial resolution. Segmentation tools that include algorithms to reduce blur far outperform those without such facilities (Lee, 2010).[6] The success of this process relies heavily on the image quality retrieved and higher resolution images can enhance the process. One must also consider motion in this setting as pixels may be derived from areas of motion within the body. Implementing a 4D gated scanning technique to account for some motion may well help reduce the effect of these pixels. In the case of head and neck radiotherapy planning PET-CT scans, this could prove difficult due to the necessary increase in

188

scan time to facilitate 4D scanning. Head and neck immobilisation devices may become intolerable to the patient for the duration of the scan depending on the individuals' performance status. A 4D gated PET-CT lasts approximately one hour duration.[7]

Zaidi and El Naqa (2010)[8] consider complex segmentation techniques and their suitability in delineating a biological target volume (BTV). Variational approaches are discussed which use mathematical functions to identify steep gradients along the whole image in both 2D and 3D. Stochastic models then evaluate the difference in intensity of tumour uptake as opposed to normal surrounding tissue.

All methods discussed in the literature involve complex mathematical functions and those choosing to implement any such method would be advised to exercise caution and ensure that they are fully aware of the positive and negative attributes of each method. The choice must be based on detailed knowledge of the scan acquisition and reconstruction, as well as the segmentation process itself.

Standardised uptake values

One of the most common segmentation thresholds used results in the division of the image into standardised uptake values (SUV). This value is generated by a relatively simple formula utilising the activity of the tracer, the injected dose and the patient's body weight. Historically an SUV value of greater than or equal to 2.5 is considered significant for malignancy on a PET scan. A relatively simple thresholding method involves auto-contouring uptake of greater than 2.5 SUV. There is much data available on SUV and the prevailing tone of these advocates caution in the use of an SUV threshold alone.

In 2004 Thie[9] discussed the variability of SUV and the significant factors that influence this value. These factors include, among others, the shape of the region of interest (ROI), the body shape of the patient, the partial volume effect and image reconstruction methods. All

of these can be accounted for, however complex formulae are then required to calculate SUV. It was also demonstrated that SUV varies between centres, implying perhaps that different factors are accounted for in different institutions. The average SUV of squamous cell head and neck carcinomas was found to vary between 3.2 in one centre to 9.4 in another. One must assume that lower SUV would also be subject to such variability, implying that what may be auto-contoured in one centre could be missed in another. If implementing an SUV threshold method some independent evaluation of the generated SUVs would be essential in validating a centre's results.

Although not as recent, the unreliability of SUV is also discussed by Keyes in 1995.[10] The significance of the uptake time is discussed in detail and it is shown that SUV within a lesion can vary by 40% over 30 minutes. This issue must be considered when evaluating SUVs on head and neck radiotherapy planning scans where delays could occur in acquiring the images if patient compliance with immobilisation and positioning is poor. In addition this shows that something as simple as scan acquisition duration alone could influence the SUV generated.

Plasma glucose levels are one of the better known factors known to affect SUVs. High plasma glucose levels can significantly decrease the SUV values generated as shown by Lindholm et al. in 1993.[11] Glucose levels are generally checked prior to PET scanning and if unacceptably high then PET scanning cannot proceed. If implementing an SUV threshold auto-delineation it would be even necessary to consider the influence of glucose levels within a 'normal' range so as to appraise its effect on SUV. This could lead to more stringent preparation for patients than simply fasting prior to PET-CT as suggested by Lindholm et al.[11]

If an SUV threshold contour is generated then normal structures with a high uptake must be excluded. This includes metabolically active tissue such as the brain as well as those organs excreting the tracer (i.e. kidneys

and bladder). Within the head and neck region uptake can be high in musculature and inflammatory tissue. Patient movement or talking is discouraged after injection of the tracer in order to reduce this uptake. Distinguishing inflammatory from malignant tissue can become difficult when malignant lesions are abutting these tissues, as would often be the case in either a head and neck primary tumour or lymph node metastasis. One suggested method by Hustinx in 1999[12] to overcome this problem is dual time point scanning in which two PET scans are acquired with the scans separated by a twenty-eight minute interval. It was found that malignant tissue showed increased uptake in the second scan but the uptake remained relatively constant in inflammatory tissue. This would help eliminate any ambiguity relating to SUVs. However, this process could not be implemented without giving serious consideration to time constraints on the PET scanner in a busy nuclear medicine department as well as the potential discomfort for the patient enduring a further scan in the immobilised radiotherapy position.

Signal to background ratio

Image segmentation based on the signal to background ratio (SBR) was proposed by Daisne et al. in 2003.[13] The threshold for applying a contour varied depending on the ratio between the maximum signal and the background noise. Auto contouring on a phantom filled with varying sized spheres was carried out. Several thresholds based on the maximum intensity within a sphere compared to the background intensity measured away from the spheres were applied. The method was found to be accurate in comparison to the actual known volumes and diameters of the spheres. One must consider that this involves evaluating perfectly spherical objects within a low intensity background, which may not always be the case in head and neck tumours or involved lymph nodes as previously discussed. Van Baardwijk et al. evaluated the use of SBR for delineation in thoracic tumours in 2007.[14] It was found that this threshold correlated well with the surgically resected specimens. However, this study

only looked at lung tumours and not head and neck cancer. It is therefore difficult to draw conclusions on its suitability for delineation in head and neck GTVs. Background uptake in the lung is likely to be less than that in the head and neck region and so a direct comparison is difficult.

Daisne et al. (2004)[15] evaluated SBR segmentation in 29 patients with oropharyngeal, hypopharyngeal and laryngeal squamous cell carcinomas. Volumes were delineated on CT, MRI and PET and comparisons made within the imaging modalities and with the surgically removed specimens in those patients who underwent total laryngectomy. It was observed that PET volumes generated by SBR correlated best with the post surgical tumour volume. The average surgical specimen volume was measured as 12.6 cm^3, the average PET GTV was 16.3 cm^3, CT GTV was 20.8 cm^3 and MRI was 23.8 cm^3. All imaging modalities overestimated the tumour volume, but the PET GTV was significantly smaller than GTVs derived from CT and MRI. However for the manual delineation on MRI and CT only one observer delineated the GTVs. A number of observers would seem to provide a more valid measure of the QVM volumes generated. In addition, the process of measuring the actual tumour volume was a complex one and could lead to inaccuracies in the gold standard volume measured. It is worth noting that although all volumes delineated on the scans were larger than the pathological tumour, no imaging modality fully encompassed the actual tumour volume, as none were sensitive enough to appreciate microscopic extension.

Davis et al. (2006)[16] discussed a variation on a direct ratio of signal to background. Background intensity was subtracted from the image before developing a threshold relative to the FDG avid areas only. It was found that a relative threshold of 41%± 2.5% best matched the phantom cylinder diameters. Although this study mainly evaluated a phantom (the limitations of which have already been discussed) they did look at three clinical cases. A head and neck tumour was evaluated using the

software developed by the group. One major downfall of this method was that the metabolically active brain tissue was contoured as GTV. This problem was overcome by manually segmenting this area out of the evaluation. However, clinically this would not be an appealing option in cases where tumour abuts inflammatory or brain tissue as it would lead to operator input, which runs contrary to the aim of employing an automated method.

A further adaptation of the method outlined in Davis *et al.* is image segmentation based on a percentage of the maximal intensity uptake on the scan. Commonly a threshold of 40% to 50% of the maximal intensity threshold is reported in the literature as shown by Wanet *et al.* in 2011[17] and Nestle *et al.* in 2007.[18] Again this data relates to lung tumours and not head and neck cancers. It was shown by Nestle *et al.* in 2005[19] that in lung tumours when comparing various PET delineation methods with a traditional CT-based volume, using a threshold at 40% of the maximal intensity was the least effective method and was the only method in their study not to significantly correlate with CT volumes. As such it is also deemed an inappropriate choice of automated thresholding for head and neck tumours.

Image based techniques

Geets *et al.*[20] proposed an interesting approach to the issue of segmentation in 2007. This complex mathematical method involves algorithms that identify gradient intensity crests within the image. These crests are representative of tissue boundaries. Their proposed method involved first deionising the image, which reduced background noise and then the application of a deblurring filter to better define the edges of the PET avid region. Following these preparatory steps, gradient based segmentation can be implemented. Comparisons were made between gradient based segmentation and SBR segmentation (Geets *et al.* 2007)[20]. A phantom was used to compare the volumes being delineated with actual known volumes of cylinders in the phantom. This analysis did show an underestimation of cylinder size by a statistically significant amount; however these corre-

sponded to very small underestimations of the radii of the cylinders (in the region of 0.5 mm to 1.1 mm). Seven PET scans of patients with T3 or T4 laryngeal tumours were then segmented using SBR and the image gradient based method. The volumes produced were then compared with the surgical specimens of the tumour. Gradient segmentation produced a tumour volume that did not differ significantly from the specimen, whereas SBR segmentation significantly overestimated the tumour volume. However, neither method resulted in a volume that fully encompassed the macroscopic laryngeal tumour.

This method shows promise but before implementation of same, consideration must be given to the method's practicalities. Some discussion by Geets *et al.* (2007)[20] concerns the calibration of the scanner but there is no mention made of the software and hardware required by such a technique. Specialist training would be necessary for those using the technique and the time necessary for such complex calculations to be carried out is also not discussed.

A new technique derived by Yu *et al.* 2009[21] utilises the data from both the PET and a fused CT scan. It is revolutionary as all other methods discussed previously are based solely on PET data and do not take the CT dataset into account. This software appraises the texture features of the images produced including coarseness, homogeneity and contrast among other characteristics. This produces a co-registered multimodality pattern analysis segmentation system (COMPASS) (Yu *et al.* 2009)[21]. COMPASS reviews each voxel within both the PET and CT images and based on the texture features and each is classified as normal or tumour tissue. This technique was applied to ten head and neck patients to produce auto-delineated GTVs that were then compared to physician outlining (deemed the gold standard) and three 'conventional' thresholding methods (SBR, 50% maximal intensity uptake and SUV $\geq$ 2.5). It was found that COMPASS generated the most comparable volume of the four methods in relation to the physician determined GTV. As discussed in Yu *et al.* there are

concerns with intra–observer variability between physicians and it would have been optimal to evaluate the volumes in relation to a surgically removed specimen. In addition it was found that evaluating PET alone was not reliable or reproducible and that it was essential to look at both PET and CT data to gain an accurate contour. This method would seem to show great potential for radiotherapy GTV delineation. It is solely based on image data hence removing variability between centres due to calibration methods, detectors, source activity etc. however, a standardised scan acquisition process may need to be implemented. From a treatment planning perspective it is reassuring to base some of the auto-delineation process on the less noisy and finer resolution CT images. Again there is no information given in this publication regarding the practicalities of implementing COM-PASS. Neither the calculation time required nor any hardware/software requirements are discussed. In addition further evaluation of the technique, namely comparing it to a surgical specimen, is necessary before it could be implemented as standard protocol.

Comparison of techniques

A number of studies evaluate a variety of segmentation methods in order to compare and contrast their effectiveness.

Schinagl, *et al*. in 2007[22] compared GTV delineated by a Radiation Oncologist (QVM) on both CT and PET independently, with GTVs derived from thresholding methods. The thresholds applied were 50% of the maximal intensity, 40% of the maximal intensity, SUV $\geq$ 2.5 and SBR. Seventy-seven head and neck patients were evaluated with various primary tumour sites and staging. It was found that the SUV $\geq$ 2.5 threshold was wholly inappropriate in almost half of the patients and in the remainder produced volumes significantly larger than all other methods. Variation between the remaining methods also existed. GTVs were consistently smaller than those produced by the Radiation Oncologist QVM. No single threshold method was recommended, as they did not correlate well with the QVM drawn GTVs. Their comparison is flawed as it assumes

the QVM GTV to be the most accurate and this has been shown to vary between even the most experienced observers (Riegel *et al*., 2006)[5]. Despite this, their study has definitively shown that SUV $\geq$ 2.5 is unsuitable in delineating head and neck GTVs due to the high SUV values of normal tissue in this region.

Further work by Schinagl *et al*. in 2009[23] evaluated the same segmentation methods and QVM in metastatic neck nodes of seventy-eight squamous cell carcinoma (SCC) head and neck patients. They found no consistency between CT positive nodes (enlarged nodes $\geq$ 10 mm or marginally enlarged 7 mm−10 mm) and PET imaging. They found QVM analysis by an experienced Oncologist identified the most nodes corresponding to the CT results. A substantial proportion of the nodes deemed enlarged on CT were found to be negative on PET and conversely a number of those deemed negative on CT were PET positive. However, the fact that fine needle aspiration cytology was only carried out sporadically and not with a view to validating either the PET or CT data, means it is difficult to consider this study conclusive. What was shown is that a large diversity exists in results achieved by varying the threshold leading to the conclusion that no single threshold provides reliable information on neck nodes in head and neck SCC. In particular they found the application of a threshold of SUV $\geq$ 2.5 to be inappropriate.

CONCLUSION

Much literature is available on the use and suitability of quantitative thresholding methods for GTV delineation on PET scans.

Segmentation of the PET-CT images is a complex process that can be affected by numerous variables. However if these variables are known, then most can be accounted for. For example a 4D gated scanning technique would be desirable to reduce motion artefact and thus aid the segmentation process.

Review of the literature has lead to the conclusion that applying an SUV based

threshold ≥ 2.5 is not suitable in head and neck cancers due to the close proximity of inflammatory tissue in this area. Studies have shown that very large inappropriate GTVs including large areas of normal tissue are generated if auto-delineation of uptake ≥ 2.5 SUV is used. In addition, the method of including all uptake greater than a fixed percentage of the maximal intensity (generally 40−50%) is also shown to be unsuitable. Again this was found to generate inappropriate volumes.

The outcome of using a signal to background ratio was found to be more encouraging with GTVs produced from this thresholding method correlating well with surgical specimens. This method requires detailed phantom measurements by each individual centre that adopts the delineation process, which would lead to increased workload for the physics department, as well as possible issues with access to a busy PET scanner.

More promising emerging methods are based on physical properties of the images produced rather than actual uptake values. These include one approach based on recognising steep gradients within the image, which imply a tissue boundary. The GTVs produced in a study of seven laryngeal patients was more accurate than those produced by SBR. However, this was a small study and further validation would be necessary before employing this technique. A more recent proposal by Yu *et al* (2009)[21] again utilised physical features of the images produced, in this case texture analysis of the data, to identify areas of malignancy. This system (COMPASS) exploited both the PET and fused CT data in order to produce a GTV. Both of these emerging techniques seem very promising and will hopefully play a role in GTV delineation for head and neck tumours after further validation.

Based on the evidence retrieved it is concluded that SBR is currently the most accurate method that has been sufficiently evaluated for use in head and neck cancer. However, the evidence does not suggest this method alone is adequately reliable and an experienced Nuclear Medicine Radiologist or Radiation Oncologist must still verify any volumes produced.

References

1. Fletcher JW, Djulbegovic B, Soares HP *et al*. Recommendations on the use of 18F-FDG PET in oncology. J Nucl Med 2008; 49:480−508.

2. Troost EG, Schinagl DA, Bussink J, Oyen WJ, Kaanders JH. Clinical evidence on PET-CT for radiation therapy planning in head and neck tumours. Radiother Oncol 2010; 96:328−334.

3. Højgaard L, Specht L. PET/CT in head and neck cancer. Eur J Nucl Med Mol Imaging 2007; 34:1329−1333.

4. Breen SL, Publicover J, De Silva S *et al*. Intraobserver and interobserver variability in GTV delineation on FDG-PET-CT images of head and neck cancers. Int J Radiat Oncol Biol Phys 2007; 68:763−770.

5. Riegel AC, Berson AM, Destian S *et al*. Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. Int J Radiat Oncol Biol Phys 2006; 65:726−732.

6. Lee JA. Segmentation of positron emission tomography images: some recommendations for target delineation in radiation oncology. Radiother Oncol 2010; 96:302−307.

7. Rock L., Chief Physicist, UPMC Beacon Hospital, Dublin. January 11, 2012, personal communication.

8. Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging 2010; 37:2165−2187.

9. Thie JA. Understanding the standardized uptake value, its methods, and implications for usage. J Nucl Med 2004; 45:1431−1434.

10. Keyes JW Jr. SUV: standard uptake or silly useless value? J Nucl Med 1995; 36:1836−1839.

11. Lindholm P, Minn H, Leskinen-Kallio S *et al*. Influence of the blood glucose concentration on FDG uptake in cancer−a PET study. J Nucl Med 1993; 34:1−6.

12. Hustinx R, Smith RJ, Benard F *et al*. Dual time point fluorine-18 fluorodeoxyglucose positron emission tomography: a potential method to differentiate malignancy from inflammation and normal tissue in the head and neck. Eur J Nucl Med 1999; 26:1345−1348.

13. Daisne JF, Sibomana M, Bol A *et al*. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Radiother Oncol 2003; 69:247−250.

14. Van Baardwijk A, Bosmans G, Boersma L *et al*. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and

193

involved nodal volumes. Int J Radiat Oncol Biol Phys 2007; 68:771–778.

15. Daisne JF, Duprez T, Weynand B et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology 2004; 233:93–100.

16. Davis JB, Reiner B, Huser M et al. Assessment of 18F PET signals for automatic target volume definition in radiotherapy treatment planning. Radiother Oncol 2006; 80:43–50.

17. Wanet M, Lee JA, Weynand B et al. Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with threshold-based approaches, CT and surgical specimens. Radiother Oncol 2011; 98:117–125.

18. Nestle U, Schaefer-Schuler A, Kremp S et al. Target volume definition for 18F-FDG PET-positive lymph nodes in radiotherapy of patients with non-small cell lung cancer. Eur J Nucl Med Mol Imaging 2007; 34:453–462.

19. Nestle U, Kremp S, Schaefer-Schuler A et al. 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 2005; 46:1342–1348.

20. Geets X, Lee JA, Bol A, Lonneux M, Grégoire V. A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging 2007; 34:1427–1438.

21. Yu H, Caldwell C, Mah K et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. Int J Radiat Oncol Biol Phys 2009; 75:618–625.

22. Schinagl DA, Vogel WV, Hoffmann AL et al. 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 2007; 69:1282–1289.

23. Schinagl DA, Hoffmann AL, Vogel WV et al. Can FDG-PET assist in radiotherapy target volume definition of metastatic lymph nodes in head-and-neck cancer? Radiother Oncol 2009; 91:95–100.