Comparison of Threshold-Based Segmentation Methods on Pre- and Post-Therapy PET Scans

Michael Phillips[1]
Michael.Phillips.1@city.ac.uk
Sally F. Barrington[1]
Derek L.G. Hill[2]
Paul K. Marsden[1]

[1] Division of Imaging Sciences, PET Imaging Centre, King's College London, London, UK.
[2] IXICO Plc, Bioscience Innovation Centre, London, UK

Abstract

The aim of this study was to compare several threshold-based segmentation methods in delineating tumours on pre- and post- therapy PET scans both in terms of the volumes segmented and the effect on predicting survival. On a dataset of 14 patients with mesothelioma, different segmentation methods were found to correlate well with each other for pre- and post- therapy tumour volume (TV) and total lesion glycolysis (TLG) but did not correlated as well for absolute change and % change of TV and TLG between pre- and post- therapy images. This was also found for the effect of the TV and TLG on receiver operator characteristics (ROC) analysis on 6-month progression free survival (PFS) in where the segmentation method had little effect when assessing pre- or post- therapy TV and TLG but showed significant differences when using change of % change as a predictor of PFS.

1 Introduction

There are a number of different threshold-based region growing segmentation methods that have been used for segmenting tumours on PET scans. These range from using a fixed standardised uptake value (SUV) threshold of 2.5, found to distinguish between malignant and benign lesions [1], to methods which take into account the background in an image, such as those defined in the PERCIST criteria [2]. Segmentation can be used to delineate PET scans to obtain tumour volume (TV) and measures combining TV and SUVs, defined as total lesion glycolysis (TLG) [3]. Pre-therapy TV/TLG and the change in TV/TLG between pre- and post- therapy scans have shown promise as useful measures for predicting patient survival in mesothelioma and other cancers [4-8]. However, there is no designated, default method of segmentation and studies have used various threshold based methods such as a fixed 2.5 SUV threshold [4,5], fixed percentage of SUV_{max} [6], PERCIST criteria method [7], and a method using SUV_{mean} (named GRAB) [8]. The aim of this work is to compare these and other methods to see if the segmentation method makes a significant difference to the measurement of TV and TLG and, more importantly,

whether it effects the prediction of response/survival in patients. This work has been completed using 14 pre- and post- therapy PET scans from a dataset investigating the use of Sorafenib as a therapy for mesothelioma [9].

2 Materials and Methods

2.1 PET Image Data

14 pre- and post- therapy PET scans of patients with mesothelioma were used for this study. All scans were acquired at the PET Imaging Centre in St Thomas' Hospital on either GE Discovery ST or GE Discovery VCT PET scanners (Waukesha, WI). Post-therapy scans were performed ~8 weeks after the start of treatment – a first line treatment of pemetrexed plus cisplatin chemotherapy, before a second line Sorafenib chemotherapy - and all patients had measurable disease, as defined by modified RECIST criteria [10]. PET image dimensions were 128 x 128 x 223, 267 or 311 (with voxel sizes of 5.47mm x 5.47mm x 3.27mm for all but one image with voxel sizes of 4.69mm x 4.69mm x 3.27mm). All PET images analysed were attenuation corrected using a smoothed CT dataset. Administered FDG dose ranged from 315 to 380MBq (median, 342MBq). The median time between administration of FDG and the start of the scan was 93min (range, 79-122min). Patients had a median age of 63 (range, 55-77) and 86% were male (12 male, 2 female). The study was approved by the UK National Research Ethics Service.

2.2 Disease Segmentation

All segmentation methods used 3-D region growing using 6-voxel connectivity in a software package called PETTRA (PET Therapy Response Assessor), a software tool designed using commercial software package MATLAB® 2012b (The MathWorks Inc., Natick, MA, 2000). Areas of segmentation were defined by an experienced consultant physician who decided which areas were disease and which were physiological uptake. In instances where the segmentation method clearly included areas of physiological uptake such as the liver, bladder or heart, restrictions were made using a given cubic area in which the segmentation must remain or which could remove unwanted physiological uptake, depending on which was easier. This was the same for all segmentation methods.

A fixed 2.5 SUV region growing segmentation method was used as the default method of segmentation due to its simplicity and objectivity. A fixed threshold based on a percentage of SUV_{max} was also used with various percentages, as has been done in other studies [7]. The PERCIST method, which uses a formula of the mean of the background plus two standard deviations (S.D.), was used as a threshold with the background taken as a large volume of interest (VOI) over the centre of the liver. A large VOI was used to increase the robustness of the algorithm, rather than use a smaller VOI prone to more observer variability as proposed in the PERCIST guidelines [5]. This background measure was also used for other segmentations which included a background mean and/or S.D. Two methods using the SUV_{mean} and background uptake were included, a method by Davis *et al*. (2006) which uses a relative percentage threshold of the SUV_{max} – the background mean (this is then added to the SUV_{mean} of the VOI) [11], and those defined by Nestle *et al*. (2005) and Nestle *et al*. (2007) which use a percentage of the SUV_{mean} and a percentage of

the background added together [12,13]. Finally, the GRAB method, an adaptive method using the SUV_{mean} and the background was used [14]. All the segmentation methods using the SUV_{mean} as a starting point underwent an iterative process to continue recalculating the segmentation until the volume no longer changed or it had exceeded ten iterations with the starting SUV_{mean} taken from a fixed 2.5 SUV threshold segmentation. Each segmentation method was used to segment disease on each of the 28 images (14 pre-therapy and 14 post-therapy). The TV and TLG was then calculated and compared for each segmentation method as was their resulting area under the curve (AUC) for receiver operating characteristics (ROC) analysis for 6-month progression free survival (PFS).

3 Results

3.1 Segmentations

Each pre- and post- therapy image was found to have some disease so all 28 images had a measure for TV and TLG. Segmentation using a threshold based on the percentage of the SUV_{max} has been omitted from the results. A variety of fixed percentages were attempted but over the entire dataset they were found to either have too low thresholds at low percentages, which segmented almost the entire body, or too high thresholds at high percentages, which clearly segmented only a fraction of the disease in the image (Figure 1, (vi)). This was also true of some of the other segmentation methods tested, particularly when using different relative thresholds.

After removing segmentation methods which had problems segmenting many images, five region growing threshold segmentation methods were tested with thresholds based on: (i) a fixed 2.5 SUV [1], (ii) background uptake + 2 S.D., recommended by PERCIST [2], (iii) background mean + 0.10 (SUV_{max} – background mean) where 0.10 is the chosen relative threshold factor, defined by Davis *et al*. (2006) [11], (iv) (0.15 * SUV_{max}) + background mean, defined by Nestle *et al*., (2005) [12], and (v) a threshold based on the adaptive GRAB method [14]. The threshold used in the GRAB method is given as:

$$\text{Threshold} = SUV_{mean} * \text{Threshold Factor}$$
$$\text{Threshold Factor} = 1 - ((SUV_{mean} - MNL) / (SUV_{mean} + MNL))$$
$$MNL = \text{Background Mean} + (\text{Background S.D.} * 3)$$

All background means and S.D.s were taken from a 201ml volume in the middle of the liver placed by an operator. This was true of all but one dataset which had disease present at the top of the liver so a smaller volume of 119ml was used instead to make sure no disease would alter the liver mean and S.D. When using the fixed 2.5 SUV threshold over the 28 images, 21 could be segmented with no restrictions. 4 pre-therapy and 3 post-therapy images had either the segmentation area restricted from physiological uptake or the physiological uptake removed from the segmented area. Physiological uptake included the heart, liver, spleen, bladder and bowel. All other segmentation methods were perceived to visually segment the diseased areas and had reasonable restrictions put in place to stop the segmentation of physiological uptake in comparison with the fixed 2.5 SUV method. A visual comparison of the segmentation methods on datasets 1 and 7 can be seen in Figure 1.

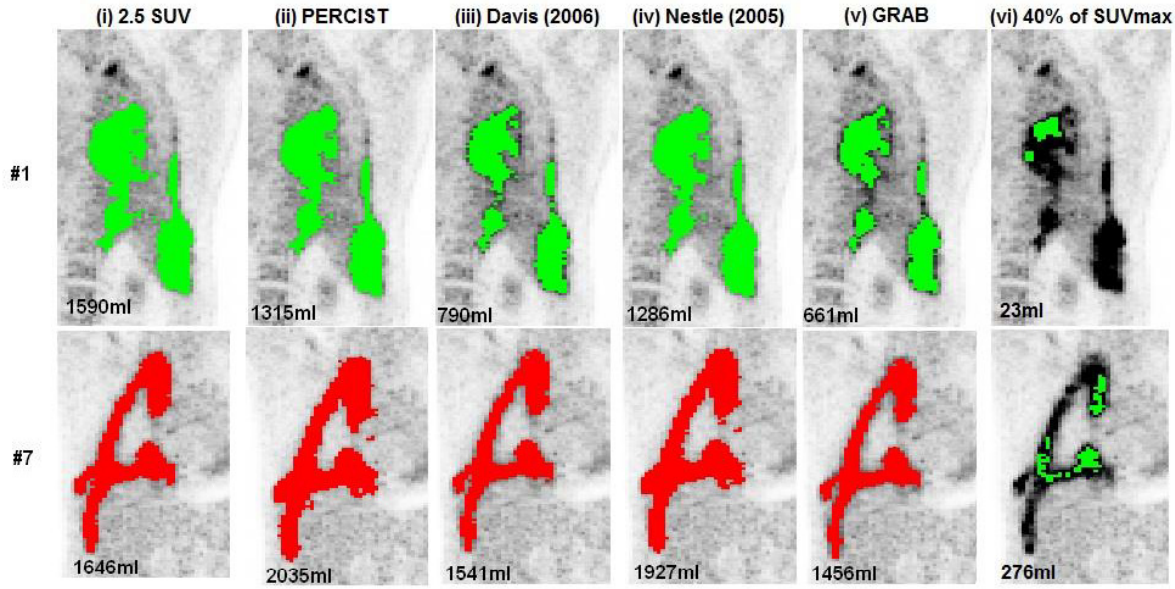

Figure 1: Coronal PET slices of segmentations of two datasets (1 and 7) by the five chosen methods: (i) fixed 2.5 SUV threshold, (ii) background mean + 2 background S.D., as recommended in the PERCIST guidelines, (iii) background mean + 0.10 (SUV_{max} – background mean), defined by Davis *et al*. (2006), (iv) (0.15 * SUV_{max}) + background mean, defined by Nestle *et al*. (2005), and (v) the GRAB method. (vi) A threshold for 40% of the SUV_{max} is also shown, as can be seen to the layman even on this 2-D coronal view, the method has clearly not segmented all of the disease.

3.2 Comparison of Segmentations

As can be seen in Figure 1, all segmentations cover the same areas of disease but differences in the threshold mean the TV can often be very different. A fixed 2.5 SUV threshold had a mean TV of 586ml over the 28 scans, while segmentations using PERCIST and the Nestle method had higher mean TV of 636ml. In comparison, the Davis method and GRAB generally produced lower TVs with means of 459ml and 327ml respectively. Compared to the fixed 2.5 SUV threshold segmentation method, the other four needed less restriction from or removal of physiological uptake in segmentation, however, it is worth noting that the PERCIST and GRAB methods needed to use more user initiated VOIs to do this over the dataset.

Table 1 shows the correlation between the fixed 2.5 SUV segmentation and the other four methods investigated for TV, TLG and the absolute change and absolute % change between them. The results show excellent correlation for all the segmentation methods for TV and TLG across all images but much weaker correlation when comparing the absolute change or % change between pre- and post- therapy images, with the PERCIST and GRAB methods having better correlation to using a fixed 2.5 SUV threshold than other methods.

Segmentation Method	All Images		Change		% Change	
	TV	TLG	TV	TLG	TV	TLG
(ii) PERCIST	0.946	0.984	0.677	0.962	0.943	0.974
(iii) Davis *et al*. (2006)	0.920	0.965	0.879	0.955	0.188	0.369
(iv) Nestle *et al*. (2005)	0.950	0.985	0.416	0.920	0.475	0.654
(v) GRAB	0.927	0.970	0.892	0.982	0.781	0.836

Table 1: Correlation between fixed 2.5 SUV threshold and other segmentation methods: shows the Pearson correlation coefficient (ppc) for TV and TLG over all 28 images and the absolute change and absolute % change over the 14 pre- and post- therapy images. For $p < 0.01$, $pcc > 0.479$ over all images and for $p < 0.01$, $pcc > 0.662$ for changes and % change.

3.3 Segmentation Effect on Predicting Response

A comparison of ROC AUC for 6-month PFS for each segmentation method shows that all segmentation methods produce similar results when using pre- and post-therapy TV and TLG to predict 6-month PFS with all methods producing AUC between 0.8 and 0.9 (with one exception, the Davis method (iii) with a pre-therapy TV AUC of 0.938), potentially caused by just one patient being categorised differently. For pre-therapy TLG all methods have identical ROC curves. However, when using change, or % change, between pre- and post- therapy TV or TLG the AUC ranges from 0.292 to 0.729. This suggests that while the segmentation method to obtain TV or TLG is unlikely to have a great effect when using pre- and post- therapy values to predict PFS, it is likely to have a greater impact when using the change, or % change, between pre- and post- therapy scans.

Method	TV				TLG			
	Pre	Post	Chg	% C	Pre	Post	Chg	% C
(i) 2.5 SUV	0.875	0.812	0.396	0.438	0.875	0.833	0.313	0.396
(ii) PERCIST	0.875	0.875	0.646	0.437	0.875	0.854	0.500	0.396
(iii) Davis	0.938	0.833	0.396	0.417	0.875	0.833	0.313	0.375
(iv) Nestle	0.875	0.875	0.729	0.604	0.875	0.854	0.521	0.458
(v) GRAB	0.875	0.875	0.458	0.354	0.875	0.833	0.375	0.292

Table 2: Area under the ROC curve (AUC) for predicting progression free survival (PFS) at 6 months, using TV and TLG, for all segmentation methods. Chg = Change between pre- and post- therapy images. % C = % change between pre- and post- therapy images.

4 Conclusion

This study compares five threshold-based segmentation methods to delineate disease on pre- and post- therapy PET studies. On 14 patients with mesothelioma, all the other methods were shown to correlate well with the fixed 2.5 SUV threshold method and the segmentation method had little or no effect on ROC analysis when using TV or TLG to predict 6-month PFS. However, when using the change, or % change, between pre- and post- therapy TV and TLG there was a much lower correlation between segmentation methods and a greater impact on ROC analysis where the change, or % change, between TV and TLG produced vastly different AUC depending on the segmentation method.

References

[1] A.C. Paulino and P.A. Johnstone. 2004. FDG-PET in radiotherapy treatment planning: Pandora's box? International Journal of Radiation Oncology Biology Physics, 59, 4-5.

[2] R.L. Wahl, H. Jacene, Y. Kasamon and M.A. Lodge. 2009. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumours. Journal of Nuclear Medicine, 50 (1), 122S-50S.

[3] S.M. Larson, Y. Erdi, T. Akhurst et al. 1999. Tumor Treatment Response Based on Visual and Quantitative Changes in Global Tumor Glycolysis Using PET-FDG Imaging. Clinical Positron Imaging, 2, 159-171.

[4] P. Veit-Haibach, N.G. Schaefer, H.C. Steinert et al. 2010. Combined FDG-PET/CT in response evaluation of malignant pleural mesothelioma. Lung Cancer, 67, 311-317.

[5] M.K. Song, J.S. Chung, H.J. Shin et al. 2012. Prognostic value of metabolic tumor volume on PET/CT in primary gastrointestinal diffuse large B-cell lymphoma. Cancer Science, 103, 477-82.

[6] T.M. Kim, J.C. Paeng, I.K. Chun et al. 2013. Total lesion glycolysis in positron emission tomography is a better predictor of outcome than the International Prognostic Index for patients with diffuse large B-cell lymphoma. Cancer, 119 (6), 1195-202.

[7] H.Y. Lee, S.H. Hyun, K.S. Lee et al. 2010. Volume-Based Parameter of F-18-FDG PET/CT in Malignant Pleural Mesothelioma: Prediction of Therapeutic Response and Prognostic Implications. Annals of Surgical Oncology, 17, 2787-2794.

[8] R.J. Francis, M.J. Byrne, A.A. Van Der Schaaf et al. 2007. Early prediction of response to chemotherapy and survival in malignant pleural mesothelioma using a novel semiautomated 3-dimensional volume-based analysis of serial 18F-FDG PET scans. Journal of Nuclear Medicine, 48, 1449-58.

[9] S. Papa, S. Popat, R.Shah et al. 2013. Phase 2 Study of Sorafenib in Malignant Mesothelioma Previously Treated with Platinum-Containing Chemotherapy. Journal of Thoracic Oncology, 8(6), 783-87.

[10] M.J. Byrne and A.K. Nowak. 2004. Modified RECIST criteria for assessment of response in malignant pleural mesothelioma. Annals of Oncology, 15, 257-260.

[11] J.B. Davis, B. Reiner, M. Huser et al. 2006. Assessment of 18F-PET signals for automatic target volume definition in radiotherapy treatment planning. Radiotherapy & Oncology, 80, 43-50.

[12] U. Nestle, S. Kremp, A. Schaefer-Schuler et al. 2005. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. Journal of Nuclear Medicine, 46, 1342-8.

[13] U. Nestle, A. Schaefer-Schuler, S. Kremp et al. 2007. Target volume definition 18F-FDG PET-positive lymph nodes in radiotherapy of patients with non-small cell lung cancer. European Journal of Nuclear Medicine and Molecular Imaging, 34, 453-62.

[14] J.A. Boucek, R.J. Francis, C.G. Jones et al. 2008. Assessment of tumour response with (18)F-fluorodeoxyglucose positron emission tomography using three-dimensional measures compared to SUVmax – a phantom study. Physics in Medicine and Biology, 53, 4213-30.