Comparison of SUVmax and SUVpeak based segmentation to determine primary lung tumour volume on FDG PET-CT correlated with pathology data

DOI: 10.1016/j.radonc.2018.06.028

Citation for published version (APA):

Published in:
Radiotherapy and Oncology

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Title page

Title: Comparison of SUVmax and SUVpeak based segmentation to determine primary lung tumour volume on FDG PET-CT correlated with pathology data.

Authors: Susan Mercieca\textsuperscript{1,2}, José Belderbos\textsuperscript{3}, Judith van Loon\textsuperscript{4}, Kenneth Gilhuijs\textsuperscript{5}, Peter Julyan\textsuperscript{6}, Marcel van Herk\textsuperscript{7}

\textsuperscript{1}Faculty of Health Science, University of Malta. Msida Malta MSD 2080. email address: susan.mercieca@um.edu.mt, Tel no: 00356 2340 1171 (the corresponding author)
\textsuperscript{2}Faculty of Medicine (AMC) - University of Amsterdam, The Netherlands
\textsuperscript{3}Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands
\textsuperscript{4}Department of Radiation Oncology (Maastro Clinic), Maastricht, The Netherlands,
\textsuperscript{5}Department of Radiology / Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands,
\textsuperscript{6}Nuclear Medicine Department, The Christie NHS Foundation Trust Wilmslow Road, Manchester, United Kingdom.
\textsuperscript{7}University of Manchester, Manchester Academic Health Centre, The Christie NHS Foundation Trust Wilmslow Road, Manchester, United Kingdom.

Running Title:

SUVpeak and SUVmax segmentation for lung cancer

Key Words:

SUVpeak, SUVmax, Segmentation, Lung tumour volume, 18F-FDG PET-CT

Disclosure:

The authors have no conflict of interest to declare. No funding was received to conduct this study.
Highlights

- SUVpeak based segmentation has been proposed as an alternative to the SUVmax since it is less affected by noise.
- The mean optimal percentage threshold was 50% and 62% for the SUVmax and SUVpeak without background correction and 47% and 60% for the SUVmax and SUVpeak respectively with background correction.
- For all four segmentation methods the threshold as a function of pathology tumour volume could be fitted well with a power law function.
- Both SUVmax and SUVpeak can be used to determine the tumour volumes in resected NSCLC tumours.
Abstract

Purpose:

The aim of the study was to compare simple SUVmax and SUVpeak based segmentation methods for calculating the lung tumour volume, compared to a pathology ground truth.

Methods:

Thirty patients diagnosed with early stage Non-Small Cell lung cancer (NSCLC) underwent surgical resection in the Netherlands between 2006 and 2008. FDG PET-CT scans for these patients were acquired within a median of 20 days before surgery. The tumour volume for each percentage SUVmax and SUVpeak threshold, with and without background correction, were calculated for each patient. The percentage threshold that provided the tumour volume that corresponded best with the pathology volume was considered to be the optimal threshold. The optimal thresholds were plotted as a function of tumour volume using a power law function and cross validated using the leave-one-out technique.

Results

The mean optimal percentage threshold was 50% +/-10% and 62% +/-15% for the SUVmax and SUVpeak without background correction respectively and 47% +/-10% and 60 +/-15% for the SUVmax and SUVpeak with background correction respectively. The optimal threshold curves could be fitted well with power law function. After cross validation the correlation between the effective tumour diameter in pathology and autosegmentation was 0.900 and 0.905 for the SUVmax and SUVpeak without background correction respectively and 0.913 and 0.908 for the SUVmax and SUVpeak with background correction respectively.

Conclusion

No benefit was shown on clinical data for the SUVpeak based segmentation method over a SUVmax based one. Both methods can be used to determine the tumour volumes in resected NSCLC tumours.
Introduction

An accurate estimation of the lung tumour volume from [18F]-fluorodeoxyglucose positron emission tomography-computed tomography (FDG PET-CT) is essential for diagnosis [1], radiotherapy planning [2,3] and therapy response assessment [4,5]. The manual definition of lung tumours on FDG PET-CT images is time consuming and prone to observer variation leading to an inaccurate estimation of the tumour volume [6,7]. To overcome this problem various automatic and semi-automatic segmentation algorithms have been proposed using amongst others, source-to-background ratios, standardised uptake values (SUV), pipeline segmentation and full kinetic analysis[6,8]. These methods vary significantly in complexity, computational resources used and their ability to account for tumour heterogeneity[8,9]. The European Association of Nuclear Medicine Research (EARN) highlights the need for standardisation of tumour volume measurements so as to facilitate the comparison of treatment outcomes across multiple institutions [8].

The most common segmentation method used to define the tumour volume is based on simple thresholding of the maximum SUV whereby the tumour border is defined at a percentage threshold of the maximum intensity voxel within a predefined region of interest [9]. Shepherd et al[9], reviewed 30 different segmentation algorithms used in 13 different institutions. The findings of this study indicate that manual segmentation still provides the most accurate delineation. However, simple threshold segmentation algorithms performed well in patient studies when compared to more complex segmentation algorithms such as pipeline methods even though the latter algorithms account better for tumour heterogeneity. Thresholding methods have the additional advantage of being simpler to use and more widely available. Nevertheless, simple thresholding methods have a number of limitations. First, there is no consensus on the optimal threshold to use for segmentation with proposed values ranging from 15-60% with 42% being the most commonly used threshold [10,11]. Van Loon et al,[14] compared the equivalent tumour volume diameter with pathology volume and found a strong correlation when contouring on FDG PET-CT using the 42% threshold of SUVmax with pathology. Some studies argue that a fixed threshold may not be suitable to define all tumour volumes [15,16]. As the SUVmax is measured in a single pixel, it may be affected by noise in the image leading to an inaccurate segmentation [15]. Furthermore, it also ignores the effect of lesions size and background FDG uptake [10]. Although the CT could be used to guide the definition of the GTV, it is important to note that the CT tends to overestimate the lung tumour volume particularly in regions with atelectasis [14].

The SUVpeak has been proposed as an alternative evaluation method and is determined by averaging pixel values over 1.0 cc spherical kernel centred at the high uptake part of the tumour [9]. Since this method does not rely on a single pixel, it is less likely to be affected by noise and scan time particularly for small tumours. It was therefore found to be a more
reliable and reproducible method for tumour segmentation in both phantom [16] and when using manual delineation on CT or FDG PET-CT images as gold standard [13,17-19]. Furthermore, in some studies the SUVpeak was found to correlate better with clinical outcomes[20–23]. However, to our knowledge no studies have been published validating the accuracy of using the SUVpeak as a basis to segment lung tumour volumes with pathology data as ground truth.

Therefore, the aim of the study was to evaluate the suitability of using SUVpeak in the definition of the tumour volume using pathology data as a gold standard. The objectives of the study were to identify:

- The impact of using background correction on the accuracy of the lung tumour volume segmentation;
- Whether the tumour volume has an impact on the optimal threshold;
- An optimal SUVpeak and SUVmax threshold that can be used to define the lung tumour volume;
- Whether the tumour volume defined on FDG PET-CT using SUVpeak provides a better correlation with the pathology lung tumour volume when compared to SUVmax.

Materials and Methods

Patient Data.

Between 2006 and 2008, thirty patients with non-small cell lung cancer (NSCLC) underwent surgical resection in the Netherlands in one of the following hospitals; Netherlands Cancer Institute, The Onze Lieve Vrouwe Gasthuis, or The Maastricht University Medical Centre. The patient and tumour characteristics are available in Table 1. Each patient has one lesion and therefore a total of thirty lesions were resected.

Pathology data analysis

Pathology samples were macroscopically and microscopically analysed and tumour volumes calculated as described by van Loon et al[14]. The lung lobe was first inflated with 10% formaldehyde and then sectioned into 0.5cm thick slices containing the gross tumour. A photograph was then acquired of each separate slice and the tumour was indexed on each glass slide by two experienced pathologists. The CT scan was used to correct for deformations in the tumour following surgical resection as described by Stroom et al[24]. To obtain the volume on each slide the indexed tumour area was calculated on each slide and multiplied by the slice thickness. The gross tumour volumes on each slide were then combined to obtain the total volume.
Acquisition of imaging data

FDG PET-CT scans were acquired between days 1 to 57 (median 20 days) before undergoing surgical resection. The time between the procedures was kept as short as possible so as to reduce this risk of tumour growth as show in table 1. FDG was administered to the patients as an intra-venous bolus. All patients fasted for 6 hours prior the scan. The CT slice thickness ranged from 1 to 5 mm, depending on whether the scan was made in the diagnostic process or for study purposes. A free breathing PET scan was acquired using a voxel size of 5 mm, and a low-dose CT was used to correct for tissue attenuation. The PET scanners were calibrated between participating institutes using a dedicated NEMA phantom as per EARL accreditation standards[25]. The GTV on CT was outlined manually and the tumour volume calculated.

Data Analysis

All FDG PET-CT images were loaded on in-house developed image processing software known as Worldmatch[26] and a region of interest around the tumour was outlined on all images. A spherical region of interest with a 2.5cm diameter was also defined manually in the healthy lung to calculate the mean SUV in the background. The primary tumour volumes at each percentage SUVmax and SUVpeak threshold with and without background correction were generated from the FDG PET-CT of each patient (Fig1). Background correction was performed by subtracting half of the background SUV from the image. The parameters used for the SUVpeak determination were a resolution of 0.5mm, with a spherical uniform cross section filter of 25x25x25 pixels, corresponding to a kernel volume of 1.09cc. The PET-CT data were resampled prior SUVpeak calculation using tri-linearly interpolation, while making sure that the original pixel values were included in the resampled data. This was done in order to generate a smoothing kernel close to 1cc.

The percentage SUVmax and SUVpeak thresholds that provided the same segmented tumour volume as the pathology volume were considered to be the optimal threshold. The data were fitted using the power law function model available on the Statistical Package for the Social Sciences software (SPSS) version 20. Following visual analysis, the resulting curve from the power law function fitted the shape of the data well. However, to further confirm the goodness of fit the correlation of determination $R^2$ and the residual error using the least square criterion were calculated. The least square criterion is defined as the summation of the mean residual errors squared. The lower the least square criterion the better is the fit of the model[27].

To assess the ability of the power law function to make volume prediction, the leave-one-out-cross-validation method was used[28]. This technique involves determining the volume for each case using a power law function that was fitted on all clinical data excluding the corresponding case. The iterative algorithm starts with 50% threshold, segments the tumour...
to estimate its volume, looks up the threshold from the power law function and repeats the latter steps until converged.

The equivalent volume diameter (eqdiam) was then calculated for the pathology and the newly obtained segmented volumes using the formula:

$$eqdiam = 2\left(\frac{3V}{4\pi}\right)^{1/3}.$$  

The eqdiam on pathology and the eqdiam using the four segmentation methods were plotted and the $R^2$ was calculated. The mean percentage difference between the eqdiam using these methods and the pathology was also calculated. Based on the Shapiro-wilk test of normality, the Wilcoxon sign rank test and the paired sample t-test were used accordingly to determine whether there was any statistical difference between the mean eqdiam on pathology and mean eqdiam obtained using the four segmentation methods (i.e. SUVmax and SUVpeak with and without background correction) and also between the means of all segmentation methods.

Finally, Bland-Altman plots were used to evaluate the agreement between the pathology eqdiam and the segmented eqdiam so as to identify any systemic bias or outliers[29]. The lesions were classified according to their location as central and peripheral. In the Bland-Altman plot the differences in the eqdiam between the segmentation methods and pathology were plotted against the averages of the two eqdiam. The limits of agreement were defined as the mean difference plus and minus 1.96 times the standard deviation of the differences are identified. Any value exceeding the limits of agreement was considered an outlier. A linear regression analysis was also performed so as to identify any systemic bias in the data.

For all statistical tests a p-value below 0.05 was considered to be statistically significant.

**Ethical considerations**

The study was approved by the central medical ethics board of the Netherlands Cancer Institute. All patients gave written informed consent to participate in this study and all patient data were anonymised.

**Results**

A total of thirty patients diagnosed with NSCLC were analysed. The majority of these patients (77%) were diagnosed with pathologic stage 1 disease, while 17% were diagnosed with stage 2 and 7% had stage 3A disease (Table 1).

The mean optimal percentage thresholds were 50% +/-10% and 62%+/-15 for the SUVmax and SUVpeak respectively. The addition of background correction lowered the optimal
threshold to 47% +/-10% and 60% +/-15% for the SUVmax and SUVpeak respectively. The optimal threshold could be fitted well with a power law function (Fig.1) with a higher correlation of determination for the optimal percentage threshold for the SUVpeak \( (y=78.785x^{-0.133}, R^2=0.539, p<0.001) \) as opposed to the SUVmax \( (y=57.038x^{-0.076}, R^2=0.232, p=0.185) \). When adding background correction, the correlation was 0.184 \( (y=53.697x^{-0.072}, p=0.018) \) and 0.496 \( (y=76.023x^{-0.135}, p<0.001) \) for the SUVmax and SUVpeak respectively (Fig.1).

Cross validation on the power law function gave a correlation of 0.900 and 0.905 was obtained for the SUVmax and SUVpeak respectively (Fig.2). Use of background correction improved the correlation slightly \( (R^2=0.913\) for SUVmax and R\(^2\)=0.908 for SUV\).

The mean eqdiam with no background correction was 0.52% and 2.40% larger than pathology for the SUVmax and SUVpeak respectively. The mean eqdiam with background correction was smaller by 3.30% and 3.61% for the SUVmax and SUVpeak respectively (Table 2). However, these differences were not statistically significant (Table 3). When comparing the mean eqdiam between the segmentation methods the addition of background correction lead to statistically significantly smaller mean eqdiam when compared with the segmentation methods without background correction (Table 3).

The Bland-Altman plots (Fig.3) showed that all data points for all segmentation methods were within the limits of agreement, (i.e. ±1.96×SD of the average) except for case 28 and 26 with both tumours being located close to the diaphragm. All segmentation methods seemed to overestimate the eqdiam for tumours larger than 3.5cm. A statistically significant slightly positive correlation between the mean segmented eqdiam and tumour diameter was noted for the SUVpeak with background correction \( (p=0.042) \) and without background correction \( (p=0.035) \) but not for the SUVmax as shown in Table2.

**Discussion**

Segmentation based on SUVmax provides a simple pragmatic approach to define the tumour volume on FDG PET-CT in routine clinical practice. However, SUVmax measurements are based on a single pixel that can be affected by noise and may therefore lead to an inaccurate tumour volume. In view of this the SUVpeak has been proposed as an alternative since it is less affected by noise and therefore more likely expected to give an accurate estimate of the lung tumour activity and therewith its volume. The SUVpeak has been found to be a suitable alternative to the SUVmax in numerous studies using phantoms, or manual delineations as a gold standard [18,30]. To our knowledge, no studies have been published correlating the segmented volumes using SUVpeak with pathology for lung cancer. Although, studies have been published correlating the segmented tumour volume using
SUVmax with pathology [14,24–26]. However, in these studies the lung lobe was not always inflated following surgical resection as was performed in our study, potentially leading to an underestimation of the tumour volume [14]. Furthermore, in our study, the tumour volume was measured in three dimensions and not estimated using the pathology tumour diameter leading to a more accurate measure.

Similarly to previous published studies[15,16], our results indicate that a single fixed threshold is not suitable to define the tumour volume. A higher percentage threshold is required to define the volume for smaller tumours[33]. The mean optimal percentage threshold in our study was found to be 50% +/-10% and 62% +/-15% for the SUVmax and SUVpeak respectively. The additional of background correction lowered the optimal threshold to 47% +/-10% and 60% +/-15% for the SUVmax and SUVpeak respectively. In previous published studies, the optimal threshold for the SUVmax ranged from 15% to 60% with 42% being the most commonly used threshold [10,11] while a threshold of 50% is recommended for the SUVpeak [27]. For all four segmentation methods the threshold as a function of pathology tumour volume could be fitted well with a power law function with a higher correlation for the optimal percentage threshold for the SUVpeak ($R^2=0.539$) as opposed to the SUVmax ($R^2=0.232$). The power function fit was slightly worse with background correction for both the SUVpeak ($R^2=0.496$) and the SUVmax ($R^2=0.184$). However, $R^2$ tends to be influenced by outliers and does not represent the goodness of fit of power law function accurately. Analysis of the residual errors using the least squares criterion showed that the residual errors were similar for all models.

Leave-one-out cross validation of the fitted power law function showed a very strong correlation above 0.9 for all segmentation methods and the Wilcoxon test showed no statistically significant difference between the mean segmented volumes. This indicates that all methods can be reliably used to segment lung tumour volumes, and no advantage was seen for the SUVpeak over SUVmax, although using background correction provided a slightly better correlation.

The Bland-Altman plot shows that the SUVmax and the SUVpeak lead to a slight overestimation of the tumour volume while the addition of background correction causes slight underestimation of the tumour volume (Fig.3). When analysing the Bland-Altman plot for systematic bias, a slightly positive correlation between the mean pathology-segmented volume and the mean eqdiam difference between the pathology and the segmentation methods was noted. Almost all tumours with an eqdiam above 3.5cm had a larger diameter on FDG PET when compared with pathology. This bias could have been introduced by the fact that the sample size was small and larger tumours were underrepresented in our model since patients with large tumours tend to be medically inoperable. This could also explain why the optimal threshold in our study was often higher than the 42% SUVmax threshold recommended in some studies[14,15].
Our study has other limitations that have to be acknowledged. Variations in tumour histology, intratumour heterogeneity, shape and location as well as scanning time after injection, can affect the FDG uptake [32–35]. Patient characteristics such as variations in glucose levels, weight and age can also have an impact of the uptake of FDG by the tumour [35]. These variables were not taken into account in our study and may have influenced the accuracy of the result. Considering the number of variables that may influence the final segmented tumour volume it is extremely important that irrespective of the method used the final delineations is checked visually and edited accordingly so as to ensure that it has been segmented correctly as recommended by Konert et al [36].

Furthermore, respiratory tumour motion has been shown to blur the FDG signal leading to an inaccurate estimate of the SUV and an error in the tumour volume estimation [37]. Peripheral lesions and smaller tumours tend to have larger motion amplitudes and therefore are more likely to be affected by errors in volume estimation [37,38].

In our study the location the segmented volume error did not seem to be influenced by tumour location (central versus peripheral) as shown in Fig.3. All the data points were within the limits of agreement with the exception of cases 28 and 26. Both tumours were peripherally located close to the diaphragm an area that is more prone to motion. As reported by Liu et al [37], tumours with high motion tend to have decreased lesion activity leading to an overestimation of the tumour volume. However, in some cases particularly for tumours close to the diaphragm, the SUV can also be overestimated leading to smaller volumes, probably due to the mismatched attenuation correction as observed in case 26. The addition of background correction seems to slightly reduce this error.

A drawback of the 4D PET-CT is that it significantly increases the scanning time, increasing the chance that the patient moves during the procedure. This long procedure may not be suitable for patients suffering from shortness of breath.

For future studies we recommend to use 4D PET-CT to account for this uncertainty, but such scans were not available in this study.

Due to logistic issues the FDG PET-CT scans had to be acquired in different hospitals potentially leading to variations in the SUV measurements. However, these scanners were calibrated using a standardised phantom to reduce this uncertainty [25]. Furthermore, the tumour might have grown between the acquisition of the FDG PET-CT and the surgical resection. Everett et al [39] has shown that there is a 32% probability that the tumour has to be upstaged within 24 days and that the tumour doubling time is estimated to be 66 days. However, we had a relatively short interval (median 20 days) between the two procedures to minimise this risk. No correction was applied for tumour growth since this is very difficult
to accurately estimate for each individual tumour as a number of factors could influence this such as tumour location, size, motion and histology as stated by Everett et al[39]

Segmentation based on SUVmax and SUVpeak with and without background correction provided similarly accurate tumour volumes compared to pathology measurements, indicating that all methods can be used to calculate the primary lung tumour volume as long as the appropriate threshold based on tumour volume is applied.
Figures and Legends

**Table 1:** Patient and tumour characteristics

**Table 2:** The table shows the mean and range difference between the pathology eqdiam of the lung tumour and the segmented methods. The probability of having systematic bias is also displayed together with the linear regression results from the Bland-Altman plot.

**Table 3:** Mean eqdiam (cm) diameter difference between the four segmentation method and pathology using the Wilcoxon sign rank test and paired sample t-test.

**Fig 1:** Correlation between optimal percentage threshold and pathology volume (cc) for SUVmax and SUVpeak without and with background correction.

**Fig.2:** Correlation between the pathology eqdiam and the segmented tumour volume based on the threshold fits of Fig.1 for the SUVpeak and SUVmax (without and with background correction)

**Fig.3:** Bland-Altman plots of the primary lung tumour eqdiam (cm) obtained using SUVmax and SUVpeak with and without background versus those derived from pathology. Dashed lines represent limits of agreement (±1.96×SD of the mean). The pink lines represent the linear regression correlation, showing a slight underestimation for larger tumour volumes. The orange squares represent peripheral tumours while blue circles represent the central tumours.
Fig 1: Correlation between optimal percentage threshold and pathology volume (cc) for SUVmax and SUVpeak without and with background correction.
Fig. 2: Correlation between the pathology eqdiam and the segmented tumour volume based on the threshold fits of Fig. 1 for the SUVpeak and SUVmax (without and with background correction)
Fig.3: Bland-Altman plots of the primary lung tumour eqdiam (cm) obtained using SUVmax and SUVpeak with and without background versus those derived from pathology. Dashed lines represent limits of agreement (±1.96×SD of the mean). The pink lines represent the linear regression correlation, showing a slight underestimation for larger tumour volumes. The orange squares represent peripheral tumours while blue circles represent the central tumours.
Table 1: Patient and tumour characteristics.

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Age</th>
<th>Tumour location</th>
<th>Days between PET-CT and surgery</th>
<th>Histology</th>
<th>Pathology Vol (cc)</th>
<th>Pathology Stage</th>
<th>CT Vol (cc)</th>
<th>CT diameter (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m</td>
<td>78</td>
<td>LUL</td>
<td>5</td>
<td>Squamous cell</td>
<td>5.02</td>
<td>1B</td>
<td>14.47</td>
<td>3.10</td>
</tr>
<tr>
<td>2</td>
<td>m</td>
<td>60</td>
<td>RUL</td>
<td>35</td>
<td>Mixed</td>
<td>5.60</td>
<td>1A</td>
<td>10.41</td>
<td>2.82</td>
</tr>
<tr>
<td>3</td>
<td>m</td>
<td>42</td>
<td>RUL</td>
<td>18</td>
<td>Mixed</td>
<td>11.60</td>
<td>1A</td>
<td>8.16</td>
<td>2.58</td>
</tr>
<tr>
<td>4</td>
<td>m</td>
<td>N/A</td>
<td>RLL</td>
<td>4</td>
<td>Adeno</td>
<td>1.45</td>
<td>1A</td>
<td>2.51</td>
<td>1.76</td>
</tr>
<tr>
<td>5</td>
<td>f</td>
<td>68</td>
<td>RUL</td>
<td>1</td>
<td>Not specified</td>
<td>4.77</td>
<td>1A</td>
<td>7.84</td>
<td>2.44</td>
</tr>
<tr>
<td>6</td>
<td>f</td>
<td>65</td>
<td>LUL</td>
<td>21</td>
<td>Adeno</td>
<td>4.00</td>
<td>1A</td>
<td>4.86</td>
<td>2.19</td>
</tr>
<tr>
<td>7</td>
<td>f</td>
<td>45</td>
<td>RUL</td>
<td>34</td>
<td>Adeno</td>
<td>29.00</td>
<td>1B</td>
<td>49.7</td>
<td>4.53</td>
</tr>
<tr>
<td>8</td>
<td>m</td>
<td>54</td>
<td>LUL</td>
<td>28</td>
<td>Large cell</td>
<td>27.25</td>
<td>2A</td>
<td>49.26</td>
<td>4.66</td>
</tr>
<tr>
<td>9</td>
<td>m</td>
<td>56</td>
<td>LUL</td>
<td>1</td>
<td>Adeno</td>
<td>3.84</td>
<td>1A</td>
<td>10.95</td>
<td>2.73</td>
</tr>
<tr>
<td>10</td>
<td>f</td>
<td>65</td>
<td>RUL</td>
<td>1</td>
<td>Adeno</td>
<td>12.02</td>
<td>1B</td>
<td>21.62</td>
<td>3.54</td>
</tr>
<tr>
<td>11</td>
<td>f</td>
<td>59</td>
<td>LUL</td>
<td>56</td>
<td>Adeno</td>
<td>13.00</td>
<td>1B</td>
<td>17.81</td>
<td>3.29</td>
</tr>
<tr>
<td>12</td>
<td>f</td>
<td>62</td>
<td>RLL</td>
<td>13</td>
<td>Adeno</td>
<td>1.10</td>
<td>1A</td>
<td>3.86</td>
<td>1.96</td>
</tr>
<tr>
<td>13</td>
<td>m</td>
<td>61</td>
<td>LUL</td>
<td>3</td>
<td>Adeno</td>
<td>1.30</td>
<td>1A</td>
<td>5.42</td>
<td>2.30</td>
</tr>
<tr>
<td>14</td>
<td>m</td>
<td>77</td>
<td>LUL</td>
<td>3</td>
<td>Squamous cell</td>
<td>0.63</td>
<td>1A</td>
<td>1.87</td>
<td>1.62</td>
</tr>
<tr>
<td>15</td>
<td>f</td>
<td>46</td>
<td>RUL</td>
<td>1</td>
<td>Adeno</td>
<td>0.72</td>
<td>1A</td>
<td>1.31</td>
<td>1.34</td>
</tr>
<tr>
<td>16</td>
<td>m</td>
<td>47</td>
<td>LUL</td>
<td>41</td>
<td>Large cell</td>
<td>3.09</td>
<td>1A</td>
<td>4.72</td>
<td>2.28</td>
</tr>
<tr>
<td>17</td>
<td>m</td>
<td>42</td>
<td>LUL</td>
<td>18</td>
<td>Large cell</td>
<td>28.82</td>
<td>2B</td>
<td>71.88</td>
<td>5.32</td>
</tr>
<tr>
<td>18</td>
<td>f</td>
<td>56</td>
<td>RUL</td>
<td>49</td>
<td>Adeno</td>
<td>2.75</td>
<td>1A</td>
<td>8.52</td>
<td>2.01</td>
</tr>
<tr>
<td>19</td>
<td>f</td>
<td>66</td>
<td>RUL</td>
<td>20</td>
<td>Adeno</td>
<td>12.30</td>
<td>1B</td>
<td>24.26</td>
<td>3.73</td>
</tr>
<tr>
<td>20</td>
<td>m</td>
<td>46</td>
<td>RUL</td>
<td>2</td>
<td>Mixed</td>
<td>3.98</td>
<td>1B</td>
<td>26.67</td>
<td>3.36</td>
</tr>
<tr>
<td>21</td>
<td>f</td>
<td>72</td>
<td>LUL</td>
<td>37</td>
<td>Adeno</td>
<td>12.59</td>
<td>1B</td>
<td>14.26</td>
<td>3.17</td>
</tr>
<tr>
<td>22</td>
<td>m</td>
<td>80</td>
<td>LLL</td>
<td>52</td>
<td>Squamous cell</td>
<td>42.33</td>
<td>3A</td>
<td>233.83</td>
<td>7.77</td>
</tr>
<tr>
<td>23</td>
<td>f</td>
<td>60</td>
<td>LLL</td>
<td>42</td>
<td>Adeno</td>
<td>14.98</td>
<td>1B</td>
<td>30.63</td>
<td>3.95</td>
</tr>
<tr>
<td>24</td>
<td>f</td>
<td>42</td>
<td>RUL</td>
<td>35</td>
<td>Adeno</td>
<td>81.89</td>
<td>2B</td>
<td>180.5</td>
<td>6.99</td>
</tr>
<tr>
<td>25</td>
<td>m</td>
<td>78</td>
<td>RUL</td>
<td>3</td>
<td>Adeno</td>
<td>1.72</td>
<td>1A</td>
<td>3.82</td>
<td>2.04</td>
</tr>
<tr>
<td>26</td>
<td>m</td>
<td>66</td>
<td>LLL</td>
<td>1</td>
<td>Adeno</td>
<td>75.19</td>
<td>3A</td>
<td>172.67</td>
<td>7.08</td>
</tr>
<tr>
<td>27</td>
<td>f</td>
<td>64</td>
<td>RML</td>
<td>27</td>
<td>Large cell</td>
<td>24.38</td>
<td>2B</td>
<td>76.16</td>
<td>5.44</td>
</tr>
<tr>
<td>28</td>
<td>m</td>
<td>76</td>
<td>RLL</td>
<td>28</td>
<td>Adeno</td>
<td>17.48</td>
<td>2A</td>
<td>32.32</td>
<td>4.11</td>
</tr>
<tr>
<td>29</td>
<td>m</td>
<td>75</td>
<td>LUL</td>
<td>57</td>
<td>Large cell</td>
<td>7.50</td>
<td>1A</td>
<td>9.99</td>
<td>2.66</td>
</tr>
<tr>
<td>30</td>
<td>m</td>
<td>56</td>
<td>LLL</td>
<td>48</td>
<td>Squamous cell</td>
<td>3.51</td>
<td>1A</td>
<td>6.11</td>
<td>2.24</td>
</tr>
</tbody>
</table>

Left Upper Lobe: LUL, Left Lower Lobe: LLL, Right Upper Lobe: RUL, Right Middle Lobe: RML, Right Lower Lobe: RLL
Table 2: The table shows the mean and range difference between the four segmented eqdiam and the pathology eqdiam of the lung tumour. The probability of having systematic bias is also displayed together with the linear regression results from the Bland-Altman plot.

<table>
<thead>
<tr>
<th></th>
<th>Mean Eqdiam (cm)</th>
<th>Mean difference from pathology (cm)</th>
<th>Mean % difference from pathology</th>
<th>Range (cm)</th>
<th>p-Value</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUVmax no background</td>
<td>2.63</td>
<td>+0.01</td>
<td>1.19%</td>
<td>0.82 to -0.80</td>
<td>0.067</td>
<td>0.116</td>
</tr>
<tr>
<td>SUVmax + background</td>
<td>2.53</td>
<td>-0.09</td>
<td>-2.73%</td>
<td>0.65 to -0.82</td>
<td>0.110</td>
<td>0.094</td>
</tr>
<tr>
<td>SUVpeak no background</td>
<td>2.67</td>
<td>+0.06</td>
<td>2.75%</td>
<td>0.87 to -0.75</td>
<td>0.035</td>
<td>0.131</td>
</tr>
<tr>
<td>SUVpeak + background</td>
<td>2.52</td>
<td>-0.09</td>
<td>-3.26%</td>
<td>0.69 to -0.88</td>
<td>0.042</td>
<td>0.125</td>
</tr>
</tbody>
</table>
Table 3: Mean eqdiam (cm) diameter difference between the four segmentation method and pathology using the Wilcoxon sign rank test and paired sample t-test.

<table>
<thead>
<tr>
<th>Eqdiam</th>
<th>Paired Differences</th>
<th>Paired sample t-test p-Value</th>
<th>Wilcoxon test p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Deviation</td>
<td></td>
</tr>
<tr>
<td>Pathology* - SUVmax no background</td>
<td>-0.01</td>
<td>0.41</td>
<td>0.893</td>
</tr>
<tr>
<td>Pathology* - SUVmax with background</td>
<td>0.09</td>
<td>0.37</td>
<td>0.556</td>
</tr>
<tr>
<td>Pathology* - SUVpeak no background</td>
<td>-0.06</td>
<td>0.41</td>
<td>0.464</td>
</tr>
<tr>
<td>Pathology* - SUVpeak with background</td>
<td>0.09</td>
<td>0.40</td>
<td>0.131</td>
</tr>
<tr>
<td>SUVmax no background - SUVmax + background</td>
<td>0.10</td>
<td>0.19</td>
<td>0.007</td>
</tr>
<tr>
<td>SUVmax no background - SUVpeak no background</td>
<td>-0.05</td>
<td>0.22</td>
<td>0.240</td>
</tr>
<tr>
<td>SUVmax no background - SUVpeak + background</td>
<td>0.11</td>
<td>0.32</td>
<td>0.070</td>
</tr>
<tr>
<td>SUVmax + background - SUVpeak no background</td>
<td>-0.15</td>
<td>0.29</td>
<td>0.008</td>
</tr>
<tr>
<td>SUVmax + background - SUVpeak + background</td>
<td>0.01</td>
<td>0.21</td>
<td>0.837</td>
</tr>
<tr>
<td>SUVpeak no background - SUVmax no background</td>
<td>0.05</td>
<td>0.22</td>
<td>0.240</td>
</tr>
<tr>
<td>SUVpeak no background - SUVpeak + background</td>
<td>0.16</td>
<td>0.24</td>
<td>0.001</td>
</tr>
</tbody>
</table>

* Based on the Shapiro-Wilk test the eqdiam of the pathology data was normally distributed (p=0.055) but the eqdiam of all other segmentation methods were not. Therefore both the results of the parametric paired sample t-test and non-parametric equivalent the Wilcoxon signed rank test are provided.


