COMPARISON OF STANDARDIZED UPTAKE VALUE–BASED POSITRON EMISSION TOMOGRAPHY AND COMPUTED TOMOGRAPHY TARGET VOLUMES IN ESOPHAGEAL CANCER PATIENTS UNDERGOING RADIOTHERAPY


*Department of Radiation Oncology, †Department of Preventive Medicine and Epidemiology, and ‡Division of Surgical Oncology, Department of Surgery, Loyola University Medical Center, Maywood, IL; †Department of Radiology, Section of Nuclear Medicine, Orlando VA Medical Center, Orlando, FL; and §Department of Radiation Oncology, Virginia Hospital Center, Arlington, VA

Purpose: To study various standardized uptake value (SUV)-based approaches to ascertain the best strategy for delineating metabolic tumor volumes (MTV).

Methods and Materials: Twenty-two consecutive previously treated esophageal cancer patients with positron emission tomography (PET) imaging and computed tomography (CT)-based radiotherapy plans were studied. At the level of the tumor epicenter, MTVs were delineated at 11 different thresholds: SUV$_{2}$, SUV$_{2.5}$, SUV$_{3}$, SUV$_{3.5}$ (SUV$_{n}$); SUV$_{40\%}$, SUV$_{45\%}$, and SUV$_{50\%}$ of the maximum (SUV$_{n\%}$); and mean liver SUV + 1, 2, 3, and 4 standard deviations (SUV$_{L\sigma}$). The volume ratio and conformity index were determined between MTVs, and the corresponding CT/endoscopic ultrasound-based gross tumor volume (GTV) at the epicenter. Means were analyzed by one-way analysis of variance for repeated measures and further compared using a paired $t$ test for repeated measures.

Results: The mean conformity indices ranged from 0.33 to 0.48, being significantly ($p < 0.05$) closest to 1 at SUV$_{2.5}$ (0.47 ± 0.03) and SUV$_{L4\sigma}$ (0.48 ± 0.03). The mean volume ratios ranged from 0.39 to 2.82, being significantly closest to 1 at SUV$_{2.5}$ (1.18 ± 0.36) and SUV$_{L4\sigma}$ (1.09 ± 0.15). The mean value of the SUVs calculated using the SUV$_{L4\sigma}$ approach was 2.4.

Conclusions: Regardless of the SUV thresholding method used (i.e., absolute or relative to liver mean), a threshold of approximately 2.5 yields the highest conformity index and best approximates the CT-based GTV at the epicenter. These findings may ultimately aid radiation oncologists in the delineation of the entire GTV in esophageal cancer patients. © 2010 Elsevier Inc.

Esophageal cancer, Positron emission tomography, Radiotherapy planning, Target volume, Standardized uptake value.

INTRODUCTION

Locally advanced esophageal cancer continues to be a lethal disease, with an estimated 5-year survival rate of 20% (1). The majority of such patients are treated neoadjuvantly or definitively with a combination of radiotherapy (RT) and chemotherapy (2). The accurate delineation of tumor volume can be difficult in the nonsurgical setting and is vital to maximizing the therapeutic ratio from radiotherapy.

Computed tomography (CT) is the standard imaging modality used for tumor volume delineation during RT planning for esophageal cancer (3, 4). Computed tomography, combined with information from endoscopic ultrasound (EUS), is deemed sufficiently accurate for delineating the radial extent of the tumor (5–7). However, CT is often inadequate in defining the tumor’s longitudinal extent (8). Other modalities, such as esophagogastroduodenoscopy, may aid in longitudinal delineation of tumor but cannot be directly imported in RT planning systems and are therefore limiting.

The value of $^{18}$F-fluorodeoxy-2-D-glucose positron emission tomography ($^{18}$F-FDG PET) in the staging of esophageal cancer has been demonstrated, with reported sensitivities and specificities for detecting metastatic disease in the 80–90% range (5). Furthermore, much interest has been expressed in incorporating metabolic data from $^{18}$F-FDG PET scans to improve the accuracy of delineating tumor for RT planning by identifying a metabolic tumor volume (MTV) (9, 10). Standardized uptake value (SUV)-based approaches that identify a threshold SUV in the region of
the tumor above which all cells are deemed to be neoplastic have received much attention (11–16). The challenges with these approaches have included the interobserver and intraobserver deviations in SUV-based MTV delineation. These deviations are compounded by the variability in the dose of radioactive glucose administered and the time at which it is administered—both parameters that can alter SUV-based MTVs dramatically (17). In the setting of non–small-cell lung cancer, Nestle et al. (18) document a 3.6-fold difference in mean MTV, depending on the SUV-based thresholding mechanism used. Although none of the thresholding approaches have been standardized, the most common ones are broadly categorized as (1) absolute SUV (SUV\textsubscript{Abs}), (2) SUV relative to the maximum tumor SUV (SUV\textsubscript{eq}), and (3) SUV relative to the mean liver SUV (SUV\textsubscript{Lr}) (8, 13, 18, 19). It is unclear which general category to use, and more importantly, what value to use within that category.

To discern the optimal SUV-based thresholding strategy to use for delineating the entire MTV (including longitudinal extent) in esophageal cancer patients, we systematically compared the radial extent of the corresponding MTVs against the radial extent of the least ambiguous portion of the tumor epicenter as established on CT and EUS. In doing so, we assumed that the strategy that would provide the best match of the easily identified radial margins on CT/EUS would also provide the best match of the less easily identified longitudinal margins of the tumor.

METHODS AND MATERIALS

Study population

This study was approved by Loyola University Medical Center’s institutional review board. Twenty-two consecutive patients with locally advanced esophageal cancer treated in the Department of Radiation Oncology were retrospectively reviewed. All patients had \textsuperscript{18}F-FDG PET imaging performed as part of the initial staging workup. Patients were deemed eligible for the study as long as their primary esophageal tumor was metabolically active and distinct from other metabolic activity. A single patient with an inflammatory pneumonitic infiltrate adjacent to the primary tumor was excluded. Hence the final study population comprised 21 patients.

PET acquisition

Patients were scanned with a dedicated whole-body Philips Allegra PET scanner (Philips Medical Systems, Bothell, WA) from the mid-skel to the mid-thigh, after being injected intravenously with FDG. All patients had their height and weight determined at the time of study. Serum resting glucose values were measured and determined to be within acceptable ranges (65–200 mg/dL) before FDG administration. Fluorodeoxyglucose quantity was based on weight and ranged between 10 and 15 mCi (370–555 MBq). With gentle intravenous fluid hydration, FDG was injected intravenously, and isotope was distributed over 90 min in a resting state. The patient was positioned supine, and scan acquisition occurred over approximately 45 min. Cesium-137 transmission was used for attenuation correction. Images were reconstructed using an iterative three-dimensional algorithm (row-action maximum likelihood algorithm) with 4-mm voxels and stored using a Digital Imaging and Communications in Medicine (DICOM) format. All studies were read on the same day of acquisition by an experienced American Board of Nuclear Medicine–certified physician. The maximum SUV in the region of the tumor was calculated and recorded. The mean SUV of the liver was retrospectively obtained for the purposes of this study (Hermes Medical Solutions, Stockholm, Sweden).

Planning CT acquisition

Patients were simulated supine with arms overhead using a Philips AcQSim CT scanner (Philips Medical Systems) acquiring images with a slice thickness and table index of 5 mm (512 × 512 matrix, 0.94-mm pixel size) from the second cervical vertebral body through to the liver. Custom cradle immobilization was used (Smithers Medical Products, North Canton, OH), and 1 tablespoon of oral barium paste and 100 mL of intravenous contrast were administered to each patient immediately before the scan.

PET/CT coregistration and fusion

For the purposes of this study, the diagnostic PET-DICOM data were transferred to a RT planning workstation (Focal v4.31; CMS, St. Louis, MO) and fused to the original CT used for RT planning. A combination of manual and automated coregistration was used. Manual coregistration was carried out by visualizing the PET images using arbitrary window and level values that adequately displayed normal metabolically active organs, such as the kidneys and heart, and fusing these organs with their CT counterparts.

CT/EUS-based target volume delineation

To ultimately compare the metabolic and CT volumes, we defined a reference CT/EUS-based target volume, CT-GTV\textsubscript{E}. The CT-GTV\textsubscript{E} consisted of a 1-cm subsection of the original CT/EUS-based gross tumor volume (used for RT planning), at the level of the tumor epicenter. This was derived by reducing the original GTV superiorly and inferiorly around the axial CT slice containing the greatest tumor burden and the least amount of discrepancy in delineating the radial extent. No alterations were made in the radial direction. The tumor epicenter was determined by a single radiation oncologist (S.N.) and verified by another (F.S.V.). Endoscopic information was also used for verification. Figure 1 depicts the axial and sagittal view of one such CT-GTV\textsubscript{E} in a patient.

PET isocontour delineation

We studied 11 different SUV-based MTVs for each patient. Standardized uptake value thresholds were used to generate these metabolic volumes, each containing voxels with SUVs greater than or equal to the threshold value. Again, the thresholded values fell into three broad categories: SUV\textsubscript{Abs}, SUV\textsubscript{eq}, and SUV\textsubscript{Lr} (see Introduction). Within SUV\textsubscript{Abs} we studied SUV \(\geq 2.0\), \(\geq 2.5\), \(\geq 3.0\), and \(\geq 3.5\) (SUV\textsubscript{2.0}, SUV\textsubscript{2.5}, SUV\textsubscript{3.0}, and SUV\textsubscript{3.5}, respectively). Within SUV\textsubscript{eq} we studied SUV \(\geq 40\%\), \(\geq 45\%\), and \(\geq 50\%\) of maximum (SUV\textsubscript{40\%}, SUV\textsubscript{45\%}, and SUV\textsubscript{50\%}, respectively). Within SUV\textsubscript{Lr} we studied SUV \(\geq \text{mean liver SUV} (L) + 1 \text{ standard deviation (SD)}\), \(\geq L + 2SD\), \(\geq L + 3SD\), and \(\geq L + 4SD (\text{SUV}_{L+2SD}, \text{SUV}_{L+3SD}, \text{and SUV}_{L+4SD}, \text{respectively})\) (13, 18, 20). The window/level values on the RT planning system were adjusted to display the appropriate SUV threshold-based volume using previously published methods (13). The resulting metabolic volume was manually contoured and was modified to exclude any overlap with the heart, but any overlap with the liver or lungs was left unmodified. We anatomically excluded the heart because its muscle is usually at least as FDG avid as tumor (if not more) and would have resulted in clinically insignificant MTVs. The normal liver and lungs were not necessarily
excluded because they typically have only low-level FDG activity. Therefore, any significant FDG activity in the lung/liver was included as tumor because we believed that other causes for increased FDG activity in these organs could not be distinguished reliably. For each patient we generated 12 tumor volumes at the level of the tumor epicenter: 1 CT/EUS based (CT-GTVE) and 11 metabolic based. 

**Figure 2** demonstrates four such MTVs (SUV2.0, SUV2.5, SUV3.0, and SUV40%) in relation to the CT-GTVE, in the same patient on the same axial slice.

**Volume ratio and conformality index**

To compare the metabolic volumes with the CT-GTVE we used two measures: volume ratio (VR) and conformality index (CI). The VR is simply the ratio of two volumes. The CI describes the spatial relationship between two volumes and is calculated using the formula 

\[ CI = \frac{V_1 \cap V_2}{V_1 \cup V_2} \]

where V1 and V2 are the two volumes of interest, \( \cap \) is the symbol for their intersection, and \( \cup \) symbolizes their union. A normalized CI was calculated by identifying the maximum CI for a particular patient and then dividing each of their 11 CIs by that maximum.

**Statistical analysis**

The mean CI, normalized CI, and VR for each of the 11 SUV thresholds were calculated. The overall difference among the means of each of the thresholds was tested for significance with a repeated-measures analysis of variance. Then, the SUV threshold resulting in the mean CI closest to 1 was compared against the mean CIs produced by the other SUV thresholds using a paired \( t \) test for related samples. In general, the mean CIs ranged from 0.325 (SUV50%) to 0.478 (SUVL4s) \((F = 3.28, p = 0.02)\), and the mean normalized CIs ranged from 0.575 (SUV50%) to 0.846 (SUVL4s) \((F = 3.59, p = 0.01)\). The mean VRs ranged from 0.393 (SUV50%) to 2.875 (SUV2.0) \((F = 6.6, p = 0.01)\) (Table 2). The trend in mean CI, as it relates to the thresholding strategies and their calculated SUVs, is depicted in Fig. 3. Similarly, the trend in mean VR is shown in Fig. 4. Both figures indicate that SUVL4s results in mean CIs and VRs closest to 1, with SUVL3s and SUV2.5 not far behind.

**Comparison with SUVL4s**

When SUVL4s (being the threshold strategy that resulted in volumes most similar to our reference CT/EUS-based volume) was compared against each of the other thresholding approaches, significant differences in the mean CI were seen with SUVL3s, SUV3.5, SUV40%, SUV45%, and SUV50% (Table 3). No differences were seen with SUV2.5 and SUV2.0 (Table 2).
With regard to VR, when compared with SUVs significant differences were observed with all SUV thresholds except for SUV2.0, SUV2.5, and SUV40% (Table 3).

Comparison with the hypothetical ideal threshold

When all 11 of the thresholding strategies were compared against some hypothetical ideal threshold (with a mean VR of exactly 1), only SUV2.5, SUVL4s, and SUVL3s were noted to be significantly similar to the ideal, with regard to VR (Table 4). In addition, among the three strategies, SUV2.5 (p = 0.63) and SUVL4s (p = 0.56) were more strongly similar to the ideal threshold than SUVL3s (p = 0.15).

DISCUSSION

This study is unique in that it not only evaluates all three commonly used strategies in the setting of esophageal cancer,

**Table 1. Mean SUVs corresponding to each of the 11 SUV thresholding strategies**

<table>
<thead>
<tr>
<th>SUV threshold</th>
<th>Mean SUV ± SE*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUV2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>SUVL1σ</td>
<td>2.007 ± 0.090</td>
</tr>
<tr>
<td>SUVL2σ</td>
<td>2.138 ± 0.095</td>
</tr>
<tr>
<td>SUVL3σ</td>
<td>2.270 ± 0.101</td>
</tr>
<tr>
<td>SUVL4σ</td>
<td>2.401 ± 0.108</td>
</tr>
<tr>
<td>SUV2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>SUV3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>SUV3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>SUV40%</td>
<td>4.238 ± 0.493</td>
</tr>
<tr>
<td>SUV45%</td>
<td>4.768 ± 0.554</td>
</tr>
<tr>
<td>SUV50%</td>
<td>5.298 ± 0.616</td>
</tr>
</tbody>
</table>

**Abbreviations:** SUV = standardized uptake value; SE = standard error; SUVn = SUV of n; SUVLnσ = mean liver SUV + n standard deviations; SUVn% = n% of maximum tumor SUV.

* Because SUV 2.0, 2.5, 3.0, and 3.5 directly specify their numeric threshold value (instead of being relative to the mean liver SUV or the maximum SUV), they have no variation in their mean values.

**Table 2. Mean CI, NCI, and VR corresponding to each of the 11 SUV thresholding strategies**

<table>
<thead>
<tr>
<th>SUV threshold</th>
<th>Mean CI ± SE*</th>
<th>Mean NCI ± SE*</th>
<th>Mean VR ± SE*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUV2.0</td>
<td>0.454 ± 0.044</td>
<td>0.790 ± 0.064</td>
<td>2.875 ± 0.893</td>
</tr>
<tr>
<td>SUV2.5</td>
<td>0.467 ± 0.034</td>
<td>0.832 ± 0.046</td>
<td>1.176 ± 0.355</td>
</tr>
<tr>
<td>SUV3.0</td>
<td>0.409 ± 0.039</td>
<td>0.723 ± 0.058</td>
<td>0.645 ± 0.094</td>
</tr>
<tr>
<td>SUV3.5</td>
<td>0.362 ± 0.044</td>
<td>0.640 ± 0.069</td>
<td>0.464 ± 0.069</td>
</tr>
<tr>
<td>SUV40%</td>
<td>0.399 ± 0.035</td>
<td>0.703 ± 0.051</td>
<td>0.735 ± 0.130</td>
</tr>
<tr>
<td>SUV45%</td>
<td>0.373 ± 0.038</td>
<td>0.658 ± 0.059</td>
<td>0.514 ± 0.083</td>
</tr>
<tr>
<td>SUV50%</td>
<td>0.326 ± 0.037</td>
<td>0.575 ± 0.059</td>
<td>0.393 ± 0.059</td>
</tr>
<tr>
<td>SUVL1σ</td>
<td>0.416 ± 0.038</td>
<td>0.740 ± 0.056</td>
<td>1.990 ± 0.256</td>
</tr>
<tr>
<td>SUVL2σ</td>
<td>0.457 ± 0.033</td>
<td>0.813 ± 0.043</td>
<td>1.510 ± 0.196</td>
</tr>
<tr>
<td>SUVL3σ</td>
<td>0.474 ± 0.030</td>
<td>0.839 ± 0.032</td>
<td>1.243 ± 0.163</td>
</tr>
<tr>
<td>SUVL4σ</td>
<td>0.478 ± 0.031</td>
<td>0.846 ± 0.036</td>
<td>1.091 ± 0.152</td>
</tr>
</tbody>
</table>

**Abbreviations:** CI = conformity index; NCI = normalized CI; VR = volume ratio. Other abbreviations as in Table 1.

* Values closer to 1 point toward a better thresholding strategy.
but does so by using objective indices of similarity, including CIs and VRs. Although we focused our comparisons on the radial extent of the tumor, our results suggest that using either an absolute SUV threshold of 2.5 or an SUV threshold that is 4 SDs above the mean liver SUV might be accurate enough for contouring MTVs in esophageal cancer. The results of this study do not contradict data presented elsewhere (11, 20, 21).

For our comparison volume, we used a CT/EUS-based volume at the level of the tumor epicenter. Because 95% of our patients had tumor invading the adventitia (Stage III) at the epicenter, and because CT, combined with information from the EUS, has been deemed quite accurate in determining the radial extent of the tumor, we believed this was an acceptable reference volume (3, 4, 6, 7). Other studies have used pathologic tumor measurements as a reference comparison (19, 21). This is not feasible in patients with locally advanced esophageal cancer, who typically undergo neoadjuvant or definitive concurrent chemoradiotherapy. Additionally, the accurate quantification of the GTV and tumor length on the pathology specimen can be quite challenging with the constriction of tissue that takes place after esophagectomy (21).

The utility of a CI to compare PET and CT volumes has been described before (11). A CI captures the spatial relationship between two volumes and describes the overlap between them. Whereas an index of 1 suggests that there is 100% overlap, a CI of <0.01 suggests that the thresholding approach is as good as some hypothetical ideal thresholding strategy that results in a mean VR of exactly 1.
overlapping the two volumes, with no portion of either volume outside the other, a CI of 0.5 reflects an overlap of 67% between two equally sized volumes, leaving 33% of each volume outside the other. In our study, in addition to measuring the CI, we also calculated the normalized CI. We believed that normalizing the indices for each patient would compensate for some of the uncertainties introduced by coregistration and fusion of the PET scan with the CT.

In deciding which thresholding strategies were most desirable, we considered their mean CIs and VRs. With regard to CI, the three best means resulted from (1) SUV_{L,4σ}, (2) SUV_{L,1σ}, and (3) SUV_{2.5}. Although SUV_{L,4σ} was significantly better than the other eight strategies, none of the top three strategies were significantly better than the other, with regard to CI. The same three strategies resulted in the best VRs, but in a slightly different order: (1) SUV_{L,4σ}, (2) SUV_{2.5}, and (3) SUV_{L,1σ}. Although there was no difference of significance between SUV_{L,4σ} and SUV_{2.5}, with regard to VR, SUV_{L,3σ} was noted to be significantly inferior. In addition, when compared with the hypothetical ideal threshold, only SUV_{L,4σ}, SUV_{2.5}, and SUV_{L,3σ} were significantly similar, with regard to VR. However, when comparing the strength of its similarity (to the ideal threshold) with that of SUV_{L,4σ} and SUV_{2.5}, once again SUV_{L,3σ} seemed to be inferior. Thus, we believed that SUV_{L,4σ} and SUV_{2.5} were generally superior to SUV_{L,3σ} and the rest of the thresholding strategies.

It is also interesting to note that as the mean SUV values obtained from the various SUV thresholding strategies moved further away from SUV_{L,4σ} (mean SUV of 2.4) in either direction, the mean VRs moved consistently further away from 1. Thus, even though we did not specifically study SUV_{L,5σ} (and greater), the aforementioned trend in VRs suggests that their study may not have resulted in a better VR.

Our results also indicate that a thresholding strategy that relies on the maximum SUV (SUV_{max}) is significantly inferior to SUV_{Abs} and SUV_{L,4σ} in terms of mean CI. In addition, the SUV_{max} strategy generally results in smaller VRs than the other strategies. Although this is mostly consistent with the findings of Nestle et al. (18), it is important to note that SUV_{40%} results in mean VRs that are not significantly different than SUV_{L,4σ}. Yet we think that the consistently low mean VRs, combined with the significantly poorer mean CIs, suggest that the SUV_{max} strategy is an inferior strategy to use in general when delineating the MTV in esophageal cancer. This is in contrast to the findings in the setting of lung cancer (22), which could be a consequence of methodologic differences.

Because both SUV_{L,4σ} and SUV_{2.5} result in mean VRs and CIs that are closest to the ideal values, and because neither results in values that are significantly different from the other, we conclude that either strategy is equally efficacious. It is interesting to note that by either method the mean SUV threshold resulting in the highest CI and VR was between 2.4 and 2.5. Although there may be theoretical advantages to using the SUV_{L,σ} strategy, because it individualizes the threshold according to background liver uptake, at certain institutions it may be more practical to use an absolute SUV threshold of 2.5 in delineating the MTV. Our findings, along with the findings of Zhong et al. (21), support the use of SUV_{2.5}.

One may note that although the CIs of SUV_{2.5} and SUV_{L,4σ} were among the highest, they were still far from ideal. Certain limitations in our study may account for this. First, we did not have access to an integrated PET/CT scanner. Our coregistrations and fusions may have introduced more uncertainty than if they had been done with an integrated scanner. Yet the mean CIs resulting from SUV_{2.5} (0.454) and SUV_{L,4σ} (0.478) were similar to the mean CI (0.464) obtained by Gondi et al. (11), who used an integrated scanner. This should be interpreted in the context of their thresholding strategy, which was much more subjective. Second, respiratory motion of the tumor during the much longer acquisition time of the PET scan may have decreased accuracy during coregistration. Four-dimensional CT simulation was not used in this study. Whether the comparisons of the volumes should be adjusted for these uncertainties is an area of future investigations.

Finally, because of being unable to identify an accurate reference volume for studying the length of the tumor (CT and esophagogastroduodenoscopy are limited in this regard), it was not possible to determine how well our thresholding strategies truly represented tumor length. Because we focused our comparisons on the radial extent of the tumor in extrapolating our results to include the longitudinal extent of the tumor, one assumes that the least metabolically active neoplastic cells within the tumor epicenter might have a similar metabolic index to the least metabolically active neoplastic cells at the longitudinal edges of the tumor. Although we believe this to be a reasonable assumption, this area warrants further study.

CONCLUSIONS

We conclude that regardless of the SUV thresholding method used (i.e., absolute or relative to liver mean), a threshold of approximately 2.5 yields the highest CI and best approximates the radial extent of the gross esophageal tumor at the level of the epicenter. These findings may ultimately aid radiation oncologists in the delineation of the entire GTV in esophageal cancer patients.

REFERENCES


