

PET-based treatment response studies typically measure the change in the standardized uptake value (SUV) to quantify response. The relative changes of different SUV measures, such as maximum, peak, mean, or total SUVs (SUV\(_{\text{max}}\), SUV\(_{\text{peak}}\), SUV\(_{\text{mean}}\), or SUV\(_{\text{total}}\), respectively), are used across the literature to classify patients into response categories, with quantitative thresholds separating the different categories. We investigated the impact of different SUV measures on the quantification and classification of PET-based treatment response. Methods: Sixteen patients with solid malignancies were treated with a multitargeted receptor tyrosine kinase inhibitor, resulting in a variety of responses. Using the cellular proliferation marker 9\(^{-}\)-deoxy-3\(^{-}\)-18F-fluorothymidine (\(18\)F-FLT), we acquired whole-body PET/CT scans at baseline, during treatment, and after treatment. The highest \(18\)F-FLT uptake lesions were segmented on PET images. Tumor PET response was assessed via the relative change in SUV\(_{\text{max}}\), SUV\(_{\text{peak}}\), SUV\(_{\text{mean}}\), and SUV\(_{\text{total}}\), thereby yielding 4 different responses for each tumor at mid- and posttreatment. For each SUV measure, a population average PET response was determined over all tumors. Standard deviation (SD) and range were used to quantify variation of PET response within individual tumors and population averages.

Results: Different SUV measures resulted in substantial variation of individual tumor PET response assessments (average SD, 20%; average range, 40%). The most extreme variation between 4 PET response measures was 90% in individual tumors. Classification of tumor PET response depended strongly on the SUV measure, because different SUV measures resulted in conflicting categorizations of PET response (ambiguous treatment response assessment) in more than 80% of tumors. Variation of the population average PET response was considerably smaller (average SD, 7%; average range, 16%), and this variation was not statistically significant. Differences in tumor PET response were greatest between SUV\(_{\text{mean}}\) and SUV\(_{\text{total}}\) and smallest between SUV\(_{\text{max}}\) and SUV\(_{\text{peak}}\). Variations of tumor PET response at midtreatment and posttreatment were similar. Conclusion: Quantification and classification of PET-based treatment response in individual patients were strongly affected by the SUV measure used to assess response. This substantial uncertainty in individual patient PET response was present despite the concurrent robustness of the population average PET response. Given the ambiguity of individual patient PET responses, selection of PET-based treatment response measures and their associated thresholds should be carefully optimized.
treatment decisions. Furthermore, the quantitative thresholds governing PET-based response categorization may strongly depend on the SUV measure used to gauge response, with different thresholds applying to different measures. The sensitivity of PET-based response assessment to different SUV measures could have significant clinical implications regarding the use of PET for quantification of treatment response. Consequently, we investigated the impact of different SUV measures on quantification and classification of PET-based treatment response.

MATERIALS AND METHODS

Treatment and Imaging

Sixteen patients with advanced solid malignancies were treated with sunitinib malate (Sutent; Pfizer), a multitargeted receptor tyrosine kinase inhibitor. Sunitinib has been demonstrated to increase objective response rate and progression-free survival (PFS) in patients with renal cell carcinoma (18) and gastrointestinal stromal tumors (19) and has shown significant antitumor activity in patients with metastatic breast cancer (20), non–small cell lung cancer (21), and neuroendocrine tumors (22). Malignancies in this study included a diverse range of tumor types: renal cell carcinoma (n = 7), esophageal (n = 2), hepatocellular (n = 2), prostate (n = 1), sarcoma (n = 1), small cell lung (n = 1), thymus (n = 1), and uterine carcino-sarcoma (n = 1). Response to therapy was measured using the PET radiotracer 3'-deoxy-3'-18F-fluorothymidine (18F-FLT). As a surrogate of cellular proliferation, 18F-FLT is emerging as a promising candidate for chemotherapy response assessment as demonstrated in patients with lymphoma, breast cancer, and glioma (23–25). Patients were injected intravenously with approximately 240 MBq (6.5 mCi) of 18F-FLT and underwent whole-body PET/CT scans at baseline (pretreatment), during treatment, and after treatment using a Discovery LS PET/CT scanner (GE Healthcare). 18F-FLT was synthesized following the method described by Martin et al., with slight modifications (26). PET/CT imaging began 47 ± 4 min after injection and extended inferiorly from the base of the skull to the distal femora. Acquisition mode was 2-dimensional, and acquisition time was 10 min per bed position to minimize image noise. PET images were reconstructed on a 128 × 128 grid over a 50-cm field of view using the ordered-subset expectation maximization algorithm with 2 iterations, 24 subsets, 5-mm gaussian loop (interiteration) filter, 3-mm gaussian postprocessing filter, and CT attenuation correction. On average, patient weight changed over a 50-cm field of view using the ordered-subset expectation maximization algorithm. 

Quantification of Tumor PET Response

PET activity concentrations (MBq/mL) were converted to SUVs by normalizing by the decay-corrected injected activity per patient mass. 18F-FLT-avid lesions (~2/patient) were segmented on PET images by an experienced nuclear medicine physician. Lesion boundaries were delineated on transverse images where uptake level was visually elevated above background. These segmentations were used to generate a 3-dimensional volume of interest (VOI) for each lesion. The location and number of lesions were as follows: lung, 11; mediastinum, 5; liver, 3; abdomen, 3; adrenal, 1; gastrointestinal, 2; pelvis, 1; gluteus, 1; uterus, 1; and arm, 1. Tumor volumes ranged from 1 to 530 mL, with an average volume of 66 mL.

For an individual lesion (n), SUVtotal, SUVmean, and SUVtotal were defined as follows:

\[
SUVT_{total} = \frac{\sum_{k=1}^{K} SUV_k}{K} \quad \text{Eq. 1}
\]

\[
SUVT_{mean} = \frac{\sum_{k=1}^{K} SUV_k}{K} \quad \text{Eq. 2}
\]

\[
SUVT_{max} = \max\{SUVT_1, SUVT_2, ...SUVT_n\} \quad \text{Eq. 3}
\]

Here, n is an individual tumor, SUVt is the SUV of an individual voxel (k) within the physician-delineated tumor VOI, and K is the total number of voxels in the VOI. SUVT was defined as the average SUV within a 1 cm^3 sphere centered in the highest uptake region of the tumor (11). With 18F-FLT PET, SUVT represents total lesion proliferation, which is similar to total lesion glycolysis with 18F-FDG PET.

\[
(R(t))^{18F}-FLT = \frac{SUVT_{max}(t) - SUVT_{max}(baseline)}{SUVT_{max}(baseline)} \times 100\% \quad \text{Eq. 4}
\]

\[
(R(t))^{18F}-FLT_{peak} = \frac{SUVT_{peak}(t) - SUVT_{peak}(baseline)}{SUVT_{peak}(baseline)} \times 100\% \quad \text{Eq. 5}
\]

\[
(R(t))^{18F}-FLT_{mean} = \frac{SUVT_{mean}(t) - SUVT_{mean}(baseline)}{SUVT_{mean}(baseline)} \times 100\% \quad \text{Eq. 6}
\]

\[
(R(t))^{18F}-FLT_{total} = \frac{SUVT_{total}(t) - SUVT_{total}(baseline)}{SUVT_{total}(baseline)} \times 100\%. \quad \text{Eq. 7}
\]

Here, n is an individual tumor, SUVmean (baseline) is the baseline value of an SUV measure, SUVmean (t) is the value of the SUV measure at time point t (mid- or posttreatment), and (R(t))^{18F}-FLT is the 18F-FLT PET-based tumor proliferative response associated with an SUV measure at time point t. Unless otherwise noted, PET response is used subsequently to refer to 18F-FLT PET-based proliferative response in this study.

The 4 different SUV measures (SUVmax, SUVpeak, SUVmean, and SUVtotal) yielded 4 different PET responses for each tumor at mid-treatment and at posttreatment. At each time point, a mean PET response for each tumor was determined (mean intratumor PET response, Eq. 8), and the variation of the 4 PET responses about the mean PET response was quantified using SD and range.

\[
(R(t))^{18F}-FLT = \frac{(R(t))^{18F}-FLT_{max} + (R(t))^{18F}-FLT_{peak} + (R(t))^{18F}-FLT_{mean} + (R(t))^{18F}-FLT_{total}}{4} \quad \text{Eq. 8}
\]

Here, n is an individual tumor and (R(t))^{18F}-FLT is the mean intratumor PET response at time point t (mid- or posttreatment).
In addition, a population average PET response (Eqs. 9–12) was determined for each SUV measure by averaging the PET responses of all tumors at midtreatment or at posttreatment.

$$\langle R(t) \rangle_{SUV_{\text{mean}}} = \frac{1}{N} \sum_{n=1}^{N} (R(t))_{SUV_{\text{mean}}}^n$$  \hspace{1cm} \text{Eq. 9}

$$\langle R(t) \rangle_{SUV_{\text{peak}}} = \frac{1}{N} \sum_{n=1}^{N} (R(t))_{SUV_{\text{peak}}}^n$$  \hspace{1cm} \text{Eq. 10}

$$\langle R(t) \rangle_{SUV_{\text{mean}}} = \frac{1}{N} \sum_{n=1}^{N} (R(t))_{SUV_{\text{mean}}}^n$$  \hspace{1cm} \text{Eq. 11}

$$\langle R(t) \rangle_{SUV_{\text{total}}} = \frac{1}{N} \sum_{n=1}^{N} (R(t))_{SUV_{\text{total}}}^n$$  \hspace{1cm} \text{Eq. 12}

Here, \( n \) is an individual tumor, \( N \) is the total number of tumors, and \( \langle R(t) \rangle_{SUV_{\text{measure}}} \) is the population average PET response associated with an SUV measure at time point \( t \). Variation of the population average PET responses was measured using SD and range.

One-way ANOVA was used to test whether the changes in the different SUV measures resulted in statistically significant differences in tumor PET responses. Means were compared using Tukey honestly significant difference test. Differences were considered statistically significant at an \( \alpha \)-level less than 0.05, after adjustment for multiple comparisons.

Correlations between the variation of tumor PET response and other tumor characteristics were tested using the Pearson correlation coefficient \( (r) \) and considered statistically significant at an \( \alpha \)-level less than 0.05.

**Association of PET Response with Clinical Endpoint**

A Cox proportional hazards survival regression was used to associate the change in each SUV measure at each imaging time point (mid- and posttreatment) with the clinical endpoint, PFS. PFS was defined as the time to disease progression, either radiographic progression on CT or clinical progression of symptoms related to disease. Time to disease progression ranged from 2 to 22 mo, with a mean of 7.3 mo. Hazard ratio, covariate coefficient, and survivor function along with \( \chi^2 \) statistic and \( P \) value were determined for each SUV measure at each imaging time point. Statistical significance was achieved at an \( \alpha \)-level less than 0.05.

**RESULTS**

**Individual Tumors**

PET responses of individual tumors were sensitive to the SUV measure used to quantify the response. On average, different SUV measures resulted in substantial variation of individual tumor PET response (average SD, 20%; average range, 40%; Figs. 1–3). In individual tumors, the most extreme variation between SUV response measures was 90% (largest SD, 44%). On average, differences in tumor PET response were greatest between SUV$_{\text{mean}}$ and SUV$_{\text{total}}$ (average difference, 28%) and smallest between SUV$_{\text{max}}$ and SUV$_{\text{peak}}$ (average difference, 13%). Results at midtreatment and at posttreatment were similar (Figs. 2 and 3).

Variation of individual tumor PET response is highlighted for a uterine tumor in Figure 1. Pre- to midtreatment, all 4 SUV measures decreased by different amounts (\( R_{\text{SUV}_{\text{max}}} = -48\% \); \( R_{\text{SUV}_{\text{peak}}} = -35\% \); \( R_{\text{SUV}_{\text{mean}}} = -55\% \); and \( R_{\text{SUV}_{\text{total}}} = -77\% \)) yet all SUV response measures indicated a PET PR (using PERCIST thresholds for different PET response categories). Pre- to posttreatment, there was wide variation associated with the changes of the different SUV measures. SUV$_{\text{max}}$ and SUV$_{\text{peak}}$ increased (\( R_{\text{SUV}_{\text{max}}} = +42\% \); \( R_{\text{SUV}_{\text{peak}}} = +53\% \)) whereas SUV$_{\text{mean}}$ and SUV$_{\text{total}}$ decreased (\( R_{\text{SUV}_{\text{mean}}} = -34\% \); \( R_{\text{SUV}_{\text{total}}} = -22\% \)), resulting in multiple PET response classifications of this uterine tumor. SUV$_{\text{max}}$ and SUV$_{\text{peak}}$ indicated PET PD, SUV$_{\text{total}}$ indicated PET stable disease, and SUV$_{\text{mean}}$ indicated PET PR. Similar ambiguous categorizations of tumor PET response arose in more than 80% of tumors assessed in this study (Fig. 2).

There was no significant correlation between tumor size and the variation of individual tumor PET response (Fig. 2, tumors ordered by size). Furthermore, there was no significant correlation between the degree of PET response (i.e., PET PD, PET stable disease, or PET PR) and the variation of individual tumor PET response.

For each SUV measure at each response time point, individual tumor PET responses were tested for strength of association with the clinical endpoint of PFS (Table 1). PET response determined posttreatment using SUV$_{\text{total}}$ (\( R_{\text{posttreatment}}(R_{\text{SUV}_{\text{total}}}) \)) was significantly associated with PFS (\( P = 0.046 \)). In addition, PET response determined posttreatment using SUV$_{\text{peak}}$ (\( R_{\text{posttreatment}}(R_{\text{SUV}_{\text{peak}}}) \)) was marginally associated with PFS (\( P = 0.071 \)). However, all other PET response measures failed to achieve statistically significant association with PFS. In general, posttreatment PET response was more strongly associated with PFS than midtreatment PET response.

**FIGURE 1.** Variation of individual tumor PET response using different SUV measures. (A) $^{18}$F-FLT PET/CT images of uterine tumor (white outline), pretreatment (top), and posttreatment (bottom). After treatment, SUV$_{\text{max}}$ (arrows) increased by 40% whereas SUV$_{\text{mean}}$ decreased by 35%. (B) SUV measures (normalized to baseline) changed throughout therapy. At midtreatment, decreases of SUV measures varied but all measures indicated partial PET response (below –30%, green line). Posttreatment, wide variation of changes of SUV measures resulted in multiple, ambiguous PET response classifications, including PET PR, PET stable disease, and PET PD (above +30%, red line). mid-tx = midtreatment; post-tx = posttreatment; pre-tx = pretreatment; wk = week.

**TABLE 1.** Association of PET Response with Clinical Endpoint

<table>
<thead>
<tr>
<th>SUV Measure</th>
<th>Association with PFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUV$_{\text{max}}$</td>
<td>Strong</td>
</tr>
<tr>
<td>SUV$_{\text{peak}}$</td>
<td>Strong</td>
</tr>
<tr>
<td>SUV$_{\text{mean}}$</td>
<td>Marginal</td>
</tr>
<tr>
<td>SUV$_{\text{total}}$</td>
<td>None</td>
</tr>
</tbody>
</table>

**REFERENCES**


Population Average

The use of different SUV measures resulted in small variation of the population average PET response (average SD, 7%; average range, 16%; Figs. 3 and 4). There was slightly greater variation of the population average PET response at midtreatment (SD, 8%; range, 20%) than at posttreatment (SD, 6%; range, 12%). Differences in population average PET responses were greatest between SUVmean and SUVtotal (average difference, 16%) and smallest between SUVmax and SUVpeak (average difference, 2%). Differences between the populations of PET response associated with each SUV measure were not statistically significant.

The minimal variation of the population average PET response is shown in Figures 3 and 4. Pre- to midtreatment, all 4 SUV measures decreased by similar amounts ($R_{SUV_{max}} = -28\%$; $R_{SUV_{peak}} = -27\%$; $R_{SUV_{mean}} = -16\%$; and $R_{SUV_{total}} = -36\%$), and accordingly almost all measures classified the midtreatment population average PET response as PET stable disease (SUVtotal response fell slightly below the PET PR/PET stable disease cutoff). Pre- to posttreatment, changes of the different SUV measures varied even less ($R_{SUV_{max}} = 3\%$; $R_{SUV_{peak}} = 0\%$; $R_{SUV_{mean}} = 2\%$; and $R_{SUV_{total}} = -10\%$), and consequently all SUV response measures indicated PET stable disease.

**DISCUSSION**

The SUV measure used to determine treatment response had a dramatic effect on the quantification of PET response. On average, different SUV measures caused a 20% variation of individual tumor PET response, and this variation ranged as high as 90%. Large variation can lead to different categorizations of PET response using established response criteria where fixed thresholds separate different PET response categories (e.g., EORTC response criteria (6) or PERCIST (11)). One such case is illustrated in Figure 1 where the posttreatment PET response (week 3) was classified either as PET PD, PET stable disease, or PET PR, depending on the SUV measure used to quantify the response. Such ambiguous PET response categorizations arose in more than 80% of the tumor PET responses assessed in this study (Fig. 2). These ambiguities remained using either the EORTC or PERCIST thresholds (which are slightly different) that

**TABLE 1**

Association of SUV Response Measures with PFS

<table>
<thead>
<tr>
<th>SUV response measure</th>
<th>Assessment time point</th>
<th>Hazard ratio</th>
<th>Covariate coefficient*</th>
<th>SE</th>
<th>$\chi^2$ statistic</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$SUV_{max}</td>
<td>Midtx</td>
<td>0.990</td>
<td>-0.010</td>
<td>0.007</td>
<td>2.3</td>
<td>0.132</td>
</tr>
<tr>
<td>$\Delta$SUV_{peak}</td>
<td>Midtx</td>
<td>0.991</td>
<td>-0.009</td>
<td>0.007</td>
<td>1.6</td>
<td>0.205</td>
</tr>
<tr>
<td>$\Delta$SUV_{mean}</td>
<td>Midtx</td>
<td>0.994</td>
<td>-0.006</td>
<td>0.009</td>
<td>0.4</td>
<td>0.551</td>
</tr>
<tr>
<td>$\Delta$SUV_{total}</td>
<td>Midtx</td>
<td>0.998</td>
<td>-0.002</td>
<td>0.004</td>
<td>0.1</td>
<td>0.705</td>
</tr>
<tr>
<td>$\Delta$SUV_{max}</td>
<td>Posttx</td>
<td>1.011</td>
<td>0.011</td>
<td>0.007</td>
<td>2.5</td>
<td>0.105</td>
</tr>
<tr>
<td>$\Delta$SUV_{peak}</td>
<td>Posttx</td>
<td>1.011</td>
<td>0.011</td>
<td>0.006</td>
<td>3.3</td>
<td>0.073</td>
</tr>
<tr>
<td>$\Delta$SUV_{mean}</td>
<td>Posttx</td>
<td>1.014</td>
<td>0.014</td>
<td>0.009</td>
<td>2.1</td>
<td>0.145</td>
</tr>
<tr>
<td>$\Delta$SUV_{total}</td>
<td>Posttx</td>
<td>1.010</td>
<td>0.010</td>
<td>0.006</td>
<td>4.0</td>
<td>0.048</td>
</tr>
</tbody>
</table>

*Covariates are SUV response measures.

$\Delta$ = change; Midtx = midtreatment; Posttx = posttreatment.
The considerable variation of quantification of PET response of individual tumors using different SUV measures underscores the pressing need for systematic selection of those measures that are most effective for assessment of treatment response. Ideally, these measures should be predictive of clinical outcome and robust to imaging uncertainties. As an example, SUV response measures in this study were correlated with a clinical endpoint using a Cox proportional hazards model. Despite small patient numbers, the posttreatment change in SUV_{total} was identified as significantly associated with PFS. Larger clinical trials are necessary to establish the superiority of specific PET measures (SUV or non-SUV) for quantification of response to therapy. These trials should determine and compare the correlation of different PET response measures with clinical outcome. Combinations of PET response measures could also be explored to ascertain whether they offer improved predictive power over individual measures. Further-
more, these trials should investigate the sensitivities of these measures to a variety of imaging factors including image noise, scan acquisition and image reconstruction parameters, partial-volume effects, tumor motion, and others. Ultimately, the most predictive and robust PET measures (or combination of measures) should be selected for quantification of treatment response.

Unlike individual tumors, the population average PET response was relatively insensitive to the SUV measure used to quantify the response. On average, different SUV measures caused only a 7% variation in the population average PET response. This is consistent with the findings of Krak et al. and Yap et al. who demonstrated a high correlation of PET-based treatment responses using different SUV measures averaged over many tumors and patients (8,16). Because of an averaging effect, this variation is expected to be further reduced as more tumors are included in the population average. The minimal variation resulted in almost all SUV measures categorizing the population average PET response as PET stable disease at mid- and post-treatment (Fig. 4). This robustness of the population average highlights the strength of PET imaging for quantification of the average response to therapy. Using large numbers of patients, the population average PET response could be applied to establish clinically validated thresholds for more accurate response classification.

18F-FLT, rather than 18F-FDG, was selected as a radiotracer in this study because of the antiproliferative nature of the molecule-targeted therapy. Furthermore, 18F-FLT may be more effective for PET-based assessment of treatment response than 18F-FDG (32–34). Imaging of tumors using both 18F-FLT and 18F-FDG has revealed somewhat higher SUV and broader SUV range with 18F-FLT (32) than with 18F-FDG (23,35,36). Thus, compared with 18F-FLT, 18F-FDG is expected to result in similar if not greater variation of tumor PET response using different SUV measures.

PERCIST thresholds were applied to the 18F-FLT PET imaging response data in this study. However, PERCIST and EORTC response criteria are both based on 18F-FDG PET imaging studies. The PET response thresholds (percentage change in SUV measure) are slightly more stringent (larger) for PERCIST than for EORTC to better account for the uncertainties and variability associated with PET imaging (11,37). Minimally, PET response thresholds (e.g., percentage ± 30% in PERCIST) must be greater than these uncertainties for PET response data to achieve a meaningful level of significance. These uncertainties plague PET imaging regardless of the specific radiotracer being imaged (37). Consequently, in this study, the PERCIST thresholds were applied to the 18F-FLT PET imaging response data to account for the associated uncertainties and variability. Furthermore, uncertainties are likely to be similar for 18F-FDG PET and 18F-FLT PET because the 18F radionuclide is common to both radiotracers. In addition, PET response thresholds of ± 30% (as in PERCIST) are supported by a variety of other 18F-FLT PET-based response assessment studies (25,38,39). Ultimately, future and more refined PET response criteria may depend on the specific response metric, disease, radiotracer, imaging time point, and other relevant factors.

In this study, all SUV measures were determined using body weight (SUVBW) and not lean body mass (SUVLBM, recommended by PERCIST). However, on average, patient weight changed only 1.5% among the 3 PET scans, which would result in approximately 0.6% difference between PET response determined using SUVBW and SUVLBM. Consequently, in this study, approximately the same variation of tumor PET response is expected using either SUVBW or SUVLBM.

**CONCLUSION**

PET-based quantification of treatment response was affected substantially by the SUV measure used to assess response. Different SUV measures resulted in a 20% variation of individual tumor PET response, and this variation ranged as high as 90%. Consequently, classification of individual tumor PET response strongly depended on the SUV measure, because different SUV measures resulted in different categorizations of response in more than 80% of tumors. This substantial uncertainty in individual patient PET response was present despite the concurrent robustness of the population average PET response. Given these uncertainties, PET-based quantification of treatment response should be optimized for accurate response assessment in individual patients. Clinical trials are necessary to select the most predictive, robust SUV measures (or combinations of measures) and associated response thresholds that should be used for assessment of treatment response.

**DISCLOSURE**

The costs of publication of this article were defrayed in part by the payment of page charges. Therefore, and solely to indicate this fact, this article is hereby marked “advertisement” in accordance with 18 USC section 1734. This work was financially supported by NIH grant R01 CA136927. No other potential conflict of interest relevant to this article was reported.

**ACKNOWLEDGMENTS**

We acknowledge the PET technologists Chris Jaskowiak and Mark McNall for scanning patients after hours as well as the UW Cyclotron Research Group for producing the 18F-FLT used in the study.
REFERENCES


