Complexity of Response Evaluation During Primary Systemic Therapy of Breast Cancer: Scoring Systems and Beyond–Preliminary Results

TÍMEA TŐKÉS1, GYÖNGYVÉR SZENTMÁRTONI1, LÁSZLÓ TORGYIK1, KRISZTIÁN SOMLAI1, JANINA KULKA2, ZSOLT LENGYEL3, TAMÁS GYÖRKE4,5 and MAGDOLNA DANK1

1Oncology Division, 1st Department of Internal Medicine, 2nd Department of Pathology, and 4Department of Nuclear Medicine, Semmelweis University, Budapest, Hungary; 3Pozitron PET/CT Center, Budapest, Hungary; 4Department of Nuclear Medicine, Semmelweis University, Budapest, Hungary; 5Scanomed Ltd., Budapest, Hungary

Abstract. Background: Precise and standardized response evaluation enables clinicians to tailor primary systemic therapy (PST). Patients and Methods: Breast cancer patients underwent 18F-fluorodeoxyglucose positron emission tomography and computerized tomography (FDG-PET/CT) before and after PST. Response was assessed by maximal Standardized Uptake Value (SUVmax); morphological changes and Ki-67 labeling index (LI). In parallel response assessment was performed by European Organization for Research and Treatment of Cancer (EORTC); PET Response Criteria in Solid Tumors (PERCIST); World Health Organization (WHO); Response Evaluation Criteria in Solid Tumors (RECIST); Chevallier and Sataloff classifications, and by a novel Ki-67 score. Accuracy of different scoring systems was evaluated. Results: In the 42 enrolled patients, SUVmax, size, and Ki-67 LI decreased significantly on PST. Significant differences between patients with versus those without pathological complete response were observed for pre-treatment Ki-67 LI and SUVmax and for post-treatment Ki-67 LI, SUVmax and size. Change in Ki-67 LI was the best predictor of pathological complete response. Correlation patterns of the directly measured metabolic, morphological, and proliferation responses differed from those determined by scoring methods. Conclusion: During PST, FDG-PET/CT enables for robust assessment of treatment efficacy, but more reliable scoring systems are still needed for more precise response evaluation.

Despite the widely accepted and successful screening methods for malignant breast disease, a significant proportion of patients are still diagnosed with locally advanced breast cancer (1, 2) for which the initial treatment should be preoperative (neoadjuvant) primary systemic therapy (PST). The main goal of PST is to downstage tumors to make patients eligible for surgery (ideally breast-conserving surgery) and, in addition, PST is an in vivo test of therapeutic effect (3-8). A precise and standardized evaluation of tumor response to PST could help clinicians to promptly identify patients who do not respond to the applied treatment, thus allowing for modification of their regimen at the earliest possible opportunity (3, 9-11).

The first standardized response evaluation score was the World Health Organization (WHO) system based on comparative measurement of tumor size (12, 13). The Response Evaluation Criteria in Solid Tumors (RECIST) Group refined this system to take into account the benefit of improved imaging techniques (14). The most recent version of RECIST (version 1.1) was published in 2009 (15) and became the basis of response evaluation not only in clinical trial settings but also in daily clinical practice.

During this period a new imaging modality came into daily use in oncology: 18F-fluorodeoxyglucose positron-emission tomography (FDG-PET) and computerized tomography (CT) imaging. FDG-PET/CT has proven accuracy in the staging of solid tumors, including breast cancer, and has also been successful in the measurement of the therapeutic response (16-18). The European Organization for Research and Treatment of Cancer (EORTC) score, published in 2000, was the first PET scoring system for solid tumors (14), and was followed in 2009 by the PET Response Criteria in Solid Tumors (PERCIST; version 1.0) for hybrid PET/CT imaging (19).

In the current study, we hypothesized that consensus is required not only on the histological definition of tumor remission (16), but more broadly on all clinically-used...
indicators of tumor response. Therefore, we investigated correlations between tumor markers (tumor morphology, metabolism, proliferation and pathological response) using conventional imaging modalities, FDG-PET/CT, and pathological methods. We also tried to clarify the role of current scoring systems in measuring these response parameters. Our aim was to investigate if a scoring system has superiority in response measurement. We also tested a new scoring system for tumor proliferation defined by the Ki-67 labeling index (LI), and our first experiences with this score are discussed.

Patients and Methods

Patients. We retrospectively identified patients diagnosed with primary breast cancer and treated at the Oncological Division of the Semmelweis University between 2008 and 2012. For all patients, breast cancer was confirmed with core-biopsy, which allowed the initial evaluation of the biological behavior of the tumors. Ethical approval for the study was given by the Semmelweis University Institutional Review Board (SE TUKEB No.:120/2013, date: 12 June 2013). Written informed consent was not mandatory and was waived for this retrospective study.

PET/CT examinations. The staging FDG-PET/CTs were performed before and after the PST (just before surgery), as required by Hungarian guidelines (20). Informed consent for the PET/CT scans was routinely obtained from all patients before the examinations. The PET/CT scans were performed with dedicated whole-body PET/CT scanners (Siemens Biograph™ TruePoint™ HD, Siemens Healthcare, Siemens Medical Solutions, Malvern, PA, USA; GE Discovery™ ST 8 GE Healthcare, GE Medical Systems, Waukesha, WI, USA) following the standard protocols and guidelines (19, 21-23). Patients were exposed to the routine amount of FDG (5-10 mCi/185-370 MBq), which was dependent on body weight. Whole-body scans from skull base to mid-thigh were performed approximately 60 min after the tracer injection (19). For each patient, the pre- and post-therapeutic measurements were carried out with the same scanner.

Imaging assessment. Visual assessment of images was performed with the same software and workstation, staging and restaging examinations were evaluated together retrospectively. Focally increased FDG uptake was considered positive if it was higher than the background activity, which was defined by measurement of blood-pool activity (19). The regions of interest were located manually over the primary tumor and any axillary lymph node metastases (19), and the maximal Standardized Uptake Value (SUVmax) was measured. If the tumor showed complete morphological or metabolic remission on the restaging scans then the region of interest was located in the original position of the tumor (which was defined earlier on the staging examination). The absolute and percentage changes in SUVmax were calculated; PERCIST and EORTC scoring was carried out (24, 25).

Tumor size measurements were made according to the WHO and the RECIST (version 1.1) on the CT scans performed before the PET examinations (15). If the measurement was inaccurate on the native scans due to dense glandular breast tissue, the rater used the FDG-avid regions on the PET scans to demarcate the malignant lesion from the surrounding breast tissue. If necessary, in these cases the investigator could be unblinded to the results of the conventional imaging modalities, especially the results of the breast ultrasonography.

Clinical TNM stage was retrospectively determined by two oncologists according to the American Joint Committee on Cancer classification (seventh edition) (1) using results from the staging physical examination, X-ray mammography and breast ultrasonography, complemented by the results of the PET/CT examinations.

Histopathological analysis and pathological response evaluation. Histological analyses of the tumor tissue were performed routinely on the core-biopsy specimens (from before PST), as well as on the surgical samples taken after completion of PST, and these were evaluated retrospectively. Detailed histological characterization of tumors was performed on the core biopsy samples (histological type, nuclear grade, tubule formation, mitotic index, inflammatory cell infiltrate, presence or absence of in situ carcinoma component and lymphovascular invasion). In surgical samples, if residual tumor was present, the detailed histological characterization was repeated and in addition tumor size and nodal stage were also assessed.

Immunohistochemistry (IHC) was performed on paraffin-embedded tissue samples to evaluate hormone receptor (HR) (estrogen or progesterone) and human epidermal growth factor receptor 2 (HER2) expressions, as well as the Ki-67 labeling index (LI) and p53 protein status. HR positivity was confirmed if the Alford score was ≥3 with IHC; HER2 overexpression was defined as IHC 3+. For IHC 2+ samples, fluorescence in situ hybridization was performed to confirm gene amplification. HER2 1+ or 0 tumors were considered to be HER2 negative. Using these parameters, the biological subtype of the tumors was defined according to the recommendations of the 13th St. Gallen International Breast Cancer Conference (26).

Pathological response was evaluated using the Chevallier (27) and Sataloff (28) classifications, which were determined at our pathological facility for every patient. Of note, ductal carcinoma in situ (DCIS) was considered as pathological complete response (pCR) by definition (29). To evaluate the ‘proliferation-based response rate’ of the tumors, we created a new scoring system on the basis of the changes in the Ki-67 LI to facilitate the comparison of data and also to avoid bias in this marker (see below). Four groups were defined, cut-off values were based on the recommendations of St. Gallen International Breast Cancer Conferences (26, 30): Score 1: Ki-67 LI of 0% or Ki-67 nuclear staining is ≤5% of the tumor cells in the surgical specimens; Score 2: >30% decrease in the Ki-67 LI, and post-therapeutic Ki-67 LI is >5%, but ≤20%; Score 3: ≤30% decrease in the Ki-67 LI, or post-therapeutic Ki-67 LI is >20% (high proliferation category) or the Ki-67 LI is stable or increased <5%; Score 4: the post-therapeutic Ki-67 LI increased compared to the LI in the core-biopsy specimens (and the increase is higher than 5%).

Statistical analysis. Data are expressed as the mean ± standard deviation (SD) or median (interquartile range). Paired t- or Wilcoxon sign-rank tests (depending on the results of the normality tests) were used to compare absolute values of the response evaluation markers before and after PST. We also divided our patient population according to whether or not they achieved pCR. Comparisons between these response groups were made using t-tests or Mann–Whitney tests (depending on the results of the
normality tests). Receiver operating characteristic (ROC) analysis was also performed to identify the marker that best described pCR.

Correlations were tested with Spearman rank correlation analysis. All tests were two-sided and \( p \)-values of less than 0.05 were considered significant. Microsoft Excel 2010 (Microsoft Corp., Redmond, Washington, USA), SigmaPlot 11.0 (Systat Software Inc., San Jose, California, USA), and Statistica 64 11 (StatSoft Inc., Tulsa, Oklahoma, USA) software were used for the data collection and processing.

**Results**

**Patients’ characteristics.** In total, 449 patients were reviewed and 42 were enrolled in our study (41 female, one male). The mean ± SD age was 48.3 ± 10.77 years. PST typically comprised of taxane-based regimens (\( n = 39 \)), most frequently in combination with anthracycline therapy (\( n = 20 \)), in 3-week schedules for six cycles (Table I). After completion of PST, all patients gave their consent for surgery: 26 patients (61.9%) underwent mastectomy and 16 (38.1%) had breast-conserving surgery (sectorectomy or quadrantectomy), with axillary block dissection (73.8%) or sentinel lymph node biopsy (26.2%). All primary tumors were metabolically active, but in four patients, for the morphological measurements, FDG-avidity was used to demarcate the tumor (see above). Axillary lymph nodes were clinically positive in 28 patients, morphologically measurable (according to RECIST 1.1) in 17 patients, and considered metabolically active in 24 patients.

**Pre-treatment tumor characteristics.** There was significant correlation between the core biopsy Ki-67 LI and the initial SUVmax in primary tumors (\( r = 0.31; p = 0.04 \)); pre-treatment SUVmax was even more highly correlated with tumor size (\( r = 0.498; p = 0.0008 \)). The tumor size and core-biopsy Ki-67 LI did not correlate significantly (\( p = 0.531 \)).

**Response evaluation: absolute values.** In the primary tumor, SUVmax and tumor size decreased significantly following PST (Table II). In the axillary lymph node region, we
detected 24 FDG-avid lesions; the measured SUVmax was reduced significantly after PST. The size of 17 measurable lymph nodes also decreased significantly. The reduction of the Ki-67 LI in the primary tumors proved to be significant. By the final histological examination, 18 patients (42.9%) had achieved pCR, including three with DCIS described in the final histological report. In the remaining 24 patients, residual tumor tissue was found after PST. The overall pathological response rate (pCR and patients with partial tumor remission) was 83.3%; seven patients did not show any pathological response to the applied treatment (e.g. Chevallier IV/Sataloff T-D stage, Table III).

We investigated the prognostic value for pCR of the initial Ki-67 LI, SUVmax and tumor size (Figure 1A). We found significant differences between patients who achieved pCR and those who did not in core-biopsy Ki-67 LI and in the initial SUVmax value, but not in the initial size of the tumors (Table IV).

We also analyzed the changes in the markers of tumor response in the surgical samples and post-treatment imaging results (Figure 1B). Patients achieving pCR had significantly lower Ki-67 expressions after the PST, with lower post-
treatment SUVmax and smaller tumor size than did patients who did not achieve pCR (Table IV).

**Correlation patterns: Response by absolute values.** There was a significant correlation between percentage changes of the Ki-67 LI (ΔKi-67) and SUVmax (ΔSUVmax) (r=0.51; p=0.0007). Interestingly, changes in tumor size (Δsize) also showed quite good correlation with the ΔSUVmax (r=0.452; p=0.00282), and the ΔKi-67 (r=0.49; p=0.00088).

Based on these results, ROC analysis was performed to find the marker that described primary tumor pCR most accurately. The ROC curves of ΔKi-67, ΔSUVmax and Δsize for patients who achieved pCR are presented in Figure 2. The areas under the ROC curve (AUCs) were similar for these changes, however, the AUC of ΔKi-67 LI was the largest (AUC=0.84; good accuracy), followed by ΔSUVmax (AUC=0.82; good accuracy) and Δsize (AUC=0.74; fair accuracy).

**Response evaluation and correlation patterns: scoring systems.** Thereafter, we grouped patients according to the commonly used response evaluation scoring systems (WHO, RECIST, EORTC, PERCIST, Chevallier, Sataloff) and our own Ki-67 LI score (Table III). Pathological remission, as defined by Chevallier and Sataloff response evaluation scores, showed the best correlation with the RECIST and WHO scores. Significant correlations were found between the Ki-67 LI score and the PERCIST and EORTC scores; however, interestingly, our novel Ki-67 LI scoring system correlated slightly better with the RECIST and WHO scores. The Ki-67 LI score did not correlate well with pathological remission, particularly as assessed with the Sataloff-score (detailed correlation pattern: Table V).

**Discussion**

The relationship between the FDG uptake and the clinicopathological features of breast cancer, especially the two potentially applicable and widely investigated biological markers, tumor morphology and proliferation, were reviewed in our study to elucidate their correlation matrices. It has been reported that FDG uptake correlates best with tumor proliferation, namely with the Ki-67 LI (31, 32), a routinely used predictive and prognostic marker in breast cancer pathology (33). Less-differentiated tumors, with higher Ki-67 LI, had significantly higher FDG uptake than tumors with lower levels of tumor proliferation (34-38). Higher initial Ki-67 LIs were also associated with better response to cytotoxic treatment and relatively high pCR rates (39-42). In our current study, analysis of tumor properties

---

**Table III. Response according to the different scoring systems.**

<table>
<thead>
<tr>
<th>Scoring system</th>
<th>Response, n (%)</th>
<th>Overall response rate, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CR</td>
<td>PR</td>
</tr>
<tr>
<td>WHO</td>
<td>18 (42.9)</td>
<td>8</td>
</tr>
<tr>
<td>RECIST</td>
<td>20 (47.6)</td>
<td>11</td>
</tr>
<tr>
<td>EORTC</td>
<td>28 (66.7)</td>
<td>11</td>
</tr>
<tr>
<td>PERCIST</td>
<td>28 (66.7)</td>
<td>10</td>
</tr>
<tr>
<td>Ki-67</td>
<td>29 (69.0)</td>
<td>2</td>
</tr>
<tr>
<td>Chevallier</td>
<td></td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>Sataloff</td>
<td>T-A</td>
<td>T-B</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>5</td>
</tr>
</tbody>
</table>


**Table IV. Differences in pre- and post-treatment Ki-67 labeling index (LI), maximal Standardized Uptake Value (SUVmax), and tumor size between patients who achieved pathological complete response (pCR) versus those who did not (Non-pCR).**

<table>
<thead>
<tr>
<th>Marker</th>
<th>Non-pCR</th>
<th>pCR</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ki-67 LI (%)</td>
<td>Mean±SD</td>
<td>63.3±24.3</td>
<td>65.6±27.6</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>33 (18-60)</td>
<td>70 (40-90)</td>
<td></td>
</tr>
<tr>
<td>SUVmax</td>
<td>Mean±SD</td>
<td>10.8±5.8</td>
<td>15.8±10.8</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>11.2 (5.7-14.6)</td>
<td>14.5 (12.0-16.3)</td>
<td></td>
</tr>
<tr>
<td>Size (mm)</td>
<td>Mean±SD</td>
<td>30.2±13.6</td>
<td>34.9±16.3</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>25.0 (21.5-36.5)</td>
<td>30 (25-44)</td>
<td></td>
</tr>
</tbody>
</table>

| Post-treatment | | | |
| Ki-67 LI (%) | Mean±SD | 16.7±22.1 | 8.9±24.7 | <0.001 |
| Median (IQR) | 5 (1-28) | 0 (0-0) |
| SUVmax | Mean±SD | 3.2±3.0 | 1.4±0.6 | 0.004 |
| Median (IQR) | 1.8 (1.5-3.8) | 1.4 (1-1.6) |
| Size (mm) | Mean±SD | 15.2±15.0 | 7.6±13.0 | 0.045 |
| Median (IQR) | 14.5 (3.5-22.0) | 0 (0-20) |

IQR: Interquartile range, SD: standard deviation.
revealed a significant correlation between core-biopsy KI-67 LI and pre-treatment, initial SUVmax.

According to the results of Buck et al. (32) and Groheux et al. (37) there is no significant correlation between FDG uptake and tumor size. In contrast, Kumar et al. (43), Gil-Rendo et al. (35) and Song et al. (44) found significant correlation between tumor size and FDG uptake (smaller tumors show smaller FDG avidity). In our study, we also found a significant positive correlation between initial tumor size and FDG uptake.

In our study, nearly half (47.6%) of the patients were under 50 years old and more than 60% of the patients were pre-menopausal at the time of diagnosis. Most malignant breast lesions were high grade (31% grade 2; 69% grade 3) invasive ductal carcinomas; two breast carcinomas were classified as inflammatory breast cancer. The lesions were mostly (85.7%) of T-stage 2 or higher and two-thirds of the patients had axillary metastases. Ki-67 LI was high in 88.1% of the patients. Almost one-third of patients had triple-negative breast cancer and nearly one-quarter had luminal B1 subtype. Therefore, our patients were representative of the average target group for PST. The pCR rate was 42.9% and the overall pathological response rate was 83.3% in our patient group, which compares favorably with reports in the literature (pCR rate: 4-27%, and pathological response rate: 70%) (5, 45-48). The importance of pCR has been highlighted by several studies showing that patients achieving pCR after PST have significantly longer disease-free and overall survival than those not achieving pCR (45, 47, 49, 50).

The patients who achieved pCR in our study had significantly lower post-treatment KI-67 LI and SUVmax, and had smaller tumor size than those without a pCR. Earlier published studies have already confirmed the importance of decreased proliferation activity of tumors (i.e. Ki-67 LI) after PST (51-55) and also underlined the sensitivity of the post-therapeutic SUVmax in the prediction of pCR and the detection of response to PST (17, 56-60). In further analyses of this patient group, we also observed significant differences between patients according to initial FDG uptake (61, 62) and tumor proliferation (53, 54). However, initial tumor size was not predictive for pCR.

ROC analysis revealed that changes in the Ki-67 LI and FDG uptake were significantly associated with pCR rate, with better accuracy than that provided by morphological changes. The relatively small sample size in our study did not allow definition of exact cut-off values for these markers with satisfactory statistical power. However, Neubauer et al. found a similar relationship between reduction of Ki-67 LI and pCR attainment (63), and García García-Esquinas et al. found significant differences in the changes in SUVmax between patients with versus those without pCR (64), which reinforce the relevance of our results. According to Dose-Schwarz et al., all imaging modalities have distinct limitations in the assessment of residual tumor tissue when compared with histopathology (65). Controversy remains as to whether the clinical complete response (cCR) rate should be defined by morphology or metabolism. In our patient group, cCR rate ranged from 47.6% (according to morphological assessment) to 66.7% (metabolism-defined complete remission), whereas the actual pCR rate was 42.9%. Our results compare favorably with those from the meta-analysis of Mauri et al., which reported cCR rates of 7-65% after PST (48).

When examining the scoring systems, we hypothesized that the same good correlation would be found between the scoring system ranks as between the absolute values describing tumor proliferation and metabolism. Our results suggest that the pathological remission scores correlated best with RECIST and WHO scores (morphological remission). We also found a good correlation between our novel Ki-67 LI score and the PERCIST remission scores; however, our proliferation score showed a slightly better correlation with the morphological response scores. The differences between the response evaluation by the score systems and by the direct measurements could result from the nature and biological diversity of the used response markers. These discrepancies underlined the bias in tumor response evaluation by the absolute values of response markers (absolute measurements may be affected by sampling and
testing methodologies, image resolution, and observer subjectivity) which errors might become balanced with the deliberate application of the score systems.

Despite its strong predictive and prognostic value (66), Ki-67 LI has limitations. The known intra-tumoral heterogeneity of the marker (67) necessitates a correctly implemented core-biopsy sampling method. Moreover, the intra- and interobserver variability, especially for grade 2 carcinomas, and the differences in the IHC results caused by the different antibodies applied limit the utility of absolute Ki-67 LI values (41, 68). These considerations led us to develop a new Ki-67 LI scoring system to overcome these pitfalls. The preliminary results with our new Ki-67 LI scoring system are promising, but the system requires evaluation in a larger patient population to allow an accurate cut-off value to be determined to ensure the most sensitive and specific scoring of the response.

Regarding morphological and metabolic response, PET/CT imaging is currently the best option for tumor staging and response evaluation in breast cancer. Our findings underline the fact that glycolytic metabolism as an imaging biomarker can be indicative of tissue response after treatment, and, more importantly, might allow for prompt identification of patients with non-responsive disease (69). The biggest rival to PET/CT in this role is contrast-enhanced magnetic resonance imaging (MRI), which has high specificity and positive predictive value (18, 70). Several studies suggested that MRI detects pCR satisfactorily after PST, based on morphological remission but is secondary compared to evaluation with PET/CT imaging (65, 71).

The main limitation of our study is the relatively low number of patients included. However, in our daily clinical practice, patients eligible for PST are selected by a multi-disciplinary tumor board (which comprises of clinical oncologists, pathologists, diagnostic imaging specialists, surgeons and radiation therapists), therefore allowing us to work with a highly representative and homogeneous group of patients treated in the neoadjuvant setting.

**Conclusion**

To the best of our knowledge, our study is the first to report a comparative investigation of the relationship between the FDG uptake described by response evaluation scores and the other response criteria which are available for breast cancer. Our data suggest that response evaluation based on functional imaging techniques are generally consistent with cellular responses within the tumor after PST when assessed using direct measurement of these markers. Metabolic remission correlated well with pathological response and tumor proliferation, and to a greater extent than it did with morphological remission. However, on the basis of the scoring systems, we found that the morphological response evaluating systems were more accurate than metabolism-based scores. Further studies are required to unequivocally elucidate these differences. Nevertheless, scoring is important in balancing the bias associated with evaluations based on absolute parameters of tumor response markers. Absolute measurements may be affected by sampling and testing methodologies, image resolution, and observer subjectivity. Moreover, our results confirm the need for a complex response evaluation scoring system that encompasses the different aspects of tumor response, and thereby accuracy can be significantly improved.

**Acknowledgements**

The whole research group expresses its gratitude for editorial support from Caroline Landon and Jennifer Kelly (Medi-Kelsey Limited). This support was funded by the ‘Spectrum Radiológáival az Egészségért’ Foundation. The study was presented in part at the American Society of Clinical Oncology (ASCO) Annual Meeting 2014 (abstract No. e12010).

**Conflicts of Interest**

The Author(s) declare that they have no competing interests with regard to this study.

**References**


Received January 14, 2015
Revised May 23, 2015
Accepted May 26, 2015