Measuring tumor response and shape change on CT: esophageal cancer as a paradigm

L. H. Schwartz1, J. A. C. Colville1, M. S. Ginsberg1, L. Wang1, M. Mazumdar2, J. Kalaigian1, H. Hricak1, D. Ilson3 & G. K. Schwartz3*

Department of 1Radiology, 2Biostatistics and Epidemiology and 3Medicine, Memorial Sloan-Kettering Cancer Center, New York, USA

Received 23 December 2005; revised 17 February 2006; accepted 20 February 2006

Background: Accurate response assessment is essential for evaluating new cancer treatments. We evaluated the impact of Response Evaluation Criteria in Solid Tumors (RECIST), World Health Organization (WHO) criteria and tumor shape on response assessment in patients with metastatic esophageal cancer.

Patients and methods: In 19 patients with metastatic esophageal cancer in a phase II trial of bryostatin-1 and paclitaxel, response was retrospectively assessed for 89 lesions with RECIST and WHO criteria on baseline and serial follow-up CT scans. The eccentricity factor (EF) was introduced for measuring the degree to which tumor shape diverges from a perfect sphere [EF = \( \sqrt{1-(LPD/MD)^2} \), where LPD is the largest perpendicular diameter and MD is the maximal diameter].

Results: The disagreement rate in best overall response categorization between RECIST (unidimensional) and WHO (bidimensional) criteria was 26.3%. Change in eccentricity was significantly greater (\( P < 0.01 \)) for patients with disagreement (mean 0.31, range 0–0.91). When the short axis was used for unidimensional lymph node measurement, disagreement between WHO and RECIST lessened.

Conclusions: Response assessment by WHO and RECIST differs substantially. Greater change in eccentricity is associated with greater discordance between WHO and RECIST. The discordance between WHO and RECIST may impact on how effective a therapy is judged to be.

Key words: esophageal cancer, RECIST, response criteria, World Health Organization

Introduction

Computed tomography (CT) is increasingly used to evaluate and quantify response to new chemotherapy regimens. The criteria for assessing therapy response with CT are still evolving. The 1999 guidelines of the Response Evaluation Criteria in Solid Tumors (RECIST) Group recommend measuring only the maximal dimension of each tumor deposit (i.e. unidimensional measurement) and using the sum of these measurements to evaluate response assessment. In contrast, the method previously proposed by the World Health Organization (WHO) calls for measurement of the maximal tumor dimension and greatest perpendicular dimension (i.e. bidimensional measurement); these measurements are then multiplied and the resulting cross products for all tumor deposits are summed to assess response. Since the introduction of the RECIST guidelines, investigators have critically assessed the concordance between WHO and RECIST in various cohorts of patients and specific tumors [1–3].

A key component of evaluating cancer treatments in clinical trials is reporting the response rate and, therefore, accurate measurement of response is essential. Small differences in the response rate can affect the outcome of phase I and II clinical trials. We were particularly interested in determining whether the use of the RECIST criteria (required by the NCI) rather than the WHO criteria for response assessment would affect our decision to consider a promising novel therapy (bryostatin-1) in a phase II study sufficiently active in combination therapy to justify the continuation of its clinical development. We chose to study 19 patients with metastatic esophageal cancer enrolled in a phase II trial of bryostatin-1 and paclitaxel. In this phase II trial, determination of response rate is critically important for deciding if further drug development is needed. We elected to focus our analysis on the phase II clinical trial of weekly paclitaxel and bryostatin-1 in the treatment of patients with metastatic esophageal cancer. Bryostatin-1 is an inhibitor of protein kinase C and also inhibits the cyclin dependent kinases by inducing p21 protein expression [4]. Previously, we reported that bryostatin-1 enhanced the effect of paclitaxel in a mouse mammary carcinoma model [5]. In addition, in our phase I study of paclitaxel and bryostatin-1, we observed promising clinical activity in patients with esophageal cancer who were treated weekly with this combination therapy [6].

In the phase II clinical trial of weekly paclitaxel and bryostatin-1 in esophageal cancer, we employed a two-stage
biostatistical design created by Simon [7]. This study design was based on the response rate of 17% detected in a multicenter phase II trial of weekly, single-agent paclitaxel in patients with esophageal cancer, reported by Kelsen et al. [8]. However, the single-agent study by Kelsen et al. used WHO criteria rather than RECIST criteria to determine response. The two-stage design for our phase II study of weekly bryostatin-1 and paclitaxel required seeing four or more responses using RECIST in the first stage of the study (19 patients). If we observed less than this, the combination was to be declared ‘non-promising’ for further study.

The aim of the present study was to retrospectively evaluate the best overall response using both the WHO and RECIST criteria and explain differences in the response rates generated with the two sets of criteria. This is the first study, to our knowledge, that specifically compared these criteria in metastatic esophageal cancer.

**patients and methods**

**patient population**

Nineteen patients with esophageal carcinoma were treated in a clinical trial with bryostatin-1 and paclitaxel. The institutional review board approved the study and all participating patients gave informed consent. Inclusion criteria for the study included histologic confirmation of esophageal cancer, stage IV disease and unidimensionally measurable metastatic lesions. All patients in the study had assessment of disease by CT scans, which were reviewed on the Picture Archiving and Communication System (PACS). During the chemotherapy protocol, CT scans were obtained every 8 weeks to assess tumor response. RECIST response criteria were used prospectively during this trial.

**tumor measurement protocol**

All tumor sites in each patient were measured at baseline and on all follow-up CT scans. Intravenous and oral contrast were both administered in all but one case (due to a previous allergic reaction, one patient was not able to have intravenous contrast). The tumor measurements were obtained retrospectively using a PACS workstation (GE Healthcare, Chicago, IL). An electronic caliper tool allowed the user to draw a thin electronic line on the computer monitor to measure the maximal tumor diameter and the largest perpendicular diameter. The maximal tumor diameter is defined as the longest diameter of the tumor that can be drawn through a lesion in its entirety. Only tumor tissue should be measured, no intervening normal tissue should be included in the calculation of the maximal tumor diameter. The largest perpendicular diameter is the maximal line drawn perpendicular to the maximal tumor diameter. The largest perpendicular need not go through the midline of the maximal tumor diameter. Images could be magnified and window/level settings adjusted at the radiologist’s discretion to best display each tumor deposit; the settings were recorded by one radiologist (J.A.C.C.). The tumor site, the maximal diameter and the largest perpendicular diameter of each tumor deposit were recorded.

**response categories**

Unidimensional (maximal dimension) per cent changes were categorized according to the following RECIST classification criteria: complete response (CR), the disappearance of all disease; partial response (PR), a decrease ≥30% in the sum of the maximal tumor dimensions of all measurable disease; stable disease (SD), a decrease <30% and an increase ≤20% in the sum of the maximal tumor dimensions; and progressive disease (PD), an increase >20% in the sum of the maximal tumor dimensions.

Bidimensional (cross-product) per cent changes were categorized according to the following WHO classification criteria: CR, the disappearance of all disease; PR, a decrease ≥50% in the sum of the cross products; SD, a decrease <50% and an increase ≤25% in the sum of the cross products; and PD, an increase >25% in the sum of the cross products or new metastatic lesions.

**assessment of outcome end point**

For each patient, the response was calculated retrospectively for each follow-up scan using the standard response categories and the rules of categorization for both the conventional bidimensional (WHO) and unidimensional (RECIST) criteria. The ‘best overall response’ out of all the follow-up scans was recorded. Possible overall responses, from ‘best’ to ‘worst’, were CR, PR, SD and PD. Overall rates of disagreement between RECIST and WHO criteria and the disagreement rate for each response category were calculated.

**eccentricity**

The ‘Eccentricity Factor’ (EF) was devised as a measure of the ‘roundedness’ of a tumor and to account for change in tumor shape. The EF mathematically calculates the overall change in tumor shape based on the following equation: EF = \(1 - \frac{PD^2}{MD^2}\) where PD is the greatest perpendicular diameter and MD is the maximal diameter. For example, a lesion measuring 2 x 2 cm would have an EF of 0 whereas a lesion measuring 2 x 1 cm would have an EF of 0.86. The EF values must fall between zero and one; the higher the eccentricity value, the flatter the shape of the tumor (Figure 1).

The eccentricity was calculated for every lesion at baseline and follow-up scans, as was the change in eccentricity for each lesion. The patient population was divided into two groups: the first group consisted of 14 patients who demonstrated concordance between the bidimensional and the unidimensional measurement response assessments, and the second group consisted of five patients who showed discordance.

\[
\text{EF} = \sqrt{1 - \left(\frac{LPD}{MD}\right)^2}
\]

\[\text{LPD}, \text{largest perpendicular diameter; MD, maximal diameter.}\]

**Figure 1.** Mathematical formula for the eccentricity factor and diagram illustrating two different eccentricity values. LPD, largest perpendicular diameter; MD, maximal diameter.
short-axis lymph node response assessment
Lymph node metastases are an important prognostic factor in esophageal cancer, as a large proportion of the metastases found in esophageal cancer patients will be in lymph nodes. These lymph nodes are generally not spherical and it has been shown that the short-axis diameter is prognostically a better parameter for tumor involvement [9–14]. Therefore we chose to perform a novel secondary analysis for lymphadenopathy using a unidimensional measurement of the largest perpendicular diameter (the short-axis measurement for lymph nodes) and substituting this measurement for the maximal diameter in the RECIST classification system when measuring lymph nodes. For all patients, the best overall response determined with this revised unidimensional classification system was then compared to the best overall response determined with the conventional bidimensional (WHO) classification system.

results

patients and lesions
We evaluated 16 male patients and three female patients with a mean age of 60 years (range 30–79 years). Evaluations were performed after treatment with bryostatin-1 and paclitaxel in 19 patients. Sixty-six scans were analyzed. All 19 patients had a baseline scan and at least one follow-up scan, which were all retrospectively reviewed; on average, 3.1 follow-up scans were performed on each patient (range 2–7). Scans were obtained at the same time intervals in all patients. There was an average of 3.4 tumor deposits (lesions) at baseline (range 2–7 deposits per patient). Sites of metastases included lymph node (n = 52), liver (n = 30), adrenals (n = 3), lung (n = 3) and peritoneum (n = 1). A total of 434 tumor measurements were performed.

unidimensional analysis versus bidimensional analysis
Best overall response was determined for the 19 patients both unidimensionally and bidimensionally (Table 1). The best overall response classification differed in five of the 19 patients; three (15.8%) were categorized as SD with unidimensional criteria and PD with bidimensional criteria, and two were SD with unidimensional and PR with bidimensional criteria. Four of the discrepancies occurred when comparing the baseline scan with the first follow-up scan; the remaining difference occurred when comparing the baseline with the second follow-up scan (Figure 2). One patient demonstrated discrepancies at two time points (at both the first and second follow-up scans).

eccentricity
There was no significant difference in mean eccentricity between the concordant group (0.68, range 0–0.95) and the discordant group (0.75, range 0–0.93). However, the difference in the change in eccentricity was statistically significant (P < 0.01), with a greater change in the discordant group (mean 0.31, range 0–0.91) than in the concordant group (mean 0.16, range 0–0.82).

To understand better the reason for the discordance between the unidimensional and the bidimensional classifications, all the lesions in the discordant group were further evaluated.

<table>
<thead>
<tr>
<th>Bidimensional (WHO)</th>
<th>Unidimensional (RECIST)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR</td>
<td>CR</td>
<td>Total</td>
</tr>
<tr>
<td>PR</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>SD</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>PD</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>14</td>
</tr>
</tbody>
</table>

Grey boxes indicate discrepancies between RECIST and WHO. CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

Figure 2. Contrast-enhanced CT scans of the upper abdomen demonstrating a superior celiac lymph node at baseline (A) and at follow-up (B). At baseline the node measures 2.0 × 2.0 (EF = 0) and at follow-up it measures 1.9 × 1.4 (EF = 0.68).

Specifically, the changes in the eccentricity of lymph nodes were compared to the changes in the eccentricity of non-lymph node lesions. In the discordant group, 18 of the 24 lesions were in lymph nodes. The mean change in the eccentricity of these lymph nodes was 0.31 (range 0–0.93), while the change in the eccentricity of the non-lymph node lesions was 0.20 (range 0–0.66); the difference was statistically significant (P < 0.01).
short-axis lymph node response assessment

In four of the five (80%) patients with discordance between bidimensional (WHO) and unidimensional (RECIST) response assessments, discordance was no longer present when the maximal diameter was replaced with the short axis (the largest perpendicular diameter) for the unidimensional measurement of lymph nodes (Table 2). This finding proved consistent with the findings in the 14 patients of the concordant group, as no new discrepancies were found when the maximal diameter was replaced with the short-axis measurement. Overall concordance with use of the short-axis unidimensional measurement in lymph nodes was 94.7%, with agreement in 18 of the 19 patients. No patients were prospectively recruited into the study based upon RECIST measurements whose tumors did not meet WHO measurement criteria.

Discussion

It is critical to have a standardized method of assessing a patient’s response to cancer therapy, not only for the patient, but also for evaluating the efficacy of drugs in clinical trials. In drug development, the outcomes of clinical trials frequently need to be compared to ascertain the impact of potential new treatments. In fact, the FDA has an expedited approval process for selected anticancer drugs with documentation of tumor shrinkage. It is therefore imperative to assess tumor response accurately. The only way to accurately compare patient response between trials is to ensure that the response assessment criteria have been applied in the same way in the studies being compared. In the late 1970s the World Health Organization devised a bidimensional method of assessing tumor response [15]. It was used routinely until the introduction of the RECIST criteria, which have since become popularized in clinical trials. The RECIST criteria were formulated in an attempt to unify the response assessment criteria, to define which lesions should be evaluated and to allow new imaging modalities to be employed as the use of plain film declined [16].

Both WHO and RECIST criteria use the same four response categories, namely: complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD). They vary in the percentage of change required for each classification: WHO criteria require a decrease ≥50% in the sum of the cross products for PR; a decrease <50% and an increase ≤25% in the sum of the cross products for SD; and an increase >25% in the sum of the cross products for PD. In contrast, RECIST criteria require a decrease ≥30% in the sum of the maximal dimensions of all measurable disease for PR; a decrease <30% and an increase ≤20% in the sum of the maximal tumor dimensions for SD; and an increase >20% in the sum of the maximal tumor dimensions for PD.

Since the introduction of the RECIST criteria, which were based upon historical data sets criteria, many researchers have assessed their validity. The original paper presenting the RECIST criteria, published by James et al. [17], compared the bidimensional and unidimensional measurements in 569 patients with eight types of malignancy. Results found a concordance of 96% ± 3%, demonstrating that the maximal diameter of each tumor and the sum of the maximal diameters are all that are required to assess tumor response. Hilsenbeck et al. [18, 19], however, pointed out that there were many flaws to the original paper by James et al. [17], the most important one being the statement ‘proportional change in diameter performs well in estimating similar proportional changes in log cell kill’, which is in fact incorrect. A recent retrospective paper comparing WHO and RECIST criteria in 79 patients [20] found a 92% concordance between the two sets of criteria. Overall, nine patients were reclassified using the RECIST criteria; six of these patients were reclassified as SD, having been classified as PD by WHO criteria. This type of variation in classification has implications for patient management in clinical trials. More recently, other publications have found that in sarcoma and lung cancer there is a high degree of concordance between RECIST and WHO [1, 2]. Others have found that in lung cancer, for instance, greater changes are seen when measuring tumors volumetrically compared to unidimensionally or bidimensionally [3].

In the original RECIST publications, it does not appear that any trials included patients with esophageal cancer, which may limit the applicability of RECIST criteria to this disease. In our retrospective study, we demonstrated only a 73.7% concordance between WHO and RECIST. Five of the 19 patients were reclassified when RECIST was used (26.3%). Out of a total of five patients called PR by WHO, two were classified as PR and three were classified as SD under RECIST. Also, out of a total of five patients classified as PD by WHO, two were classified as SD under RECIST.

To try to explain the variations in response category, we introduced a new parameter, the ‘eccentricity factor’ (EF). The eccentricity factor was used to calculate the change in tumor shape, taking into account that tumors are not routinely spherical (only 4.4% of the lesions measured were spherical), and thus requiring both the maximal diameter and the largest perpendicular diameter to be used. It was noted that a total of four of the five patients who had discordance between WHO and RECIST had predominately lymph node disease. In the patients with discordance, the mean change in eccentricity of the lymph nodes was significantly greater than the mean change in the non-lymph node lesions. To evaluate the

Table 2. Best overall response assessment comparing unidimensional and bidimensional criteria, using the largest perpendicular diameter for unidimensional lymph node assessment

<table>
<thead>
<tr>
<th>Bidimensional (WHO)</th>
<th>Unidimensional (largest perpendicular diameter)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>PR</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>SD</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>PD</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>

Grey boxes indicate previous areas of discrepancy between RECIST and WHO.

CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

RECIST response rate for both RECIST and WHO was 26.3%.
importance of the largest perpendicular diameter (i.e. the greatest short axis of lymph nodes), we substituted it for the maximal diameter in the unidimensional measurement of lymph nodes. Our results were almost exactly the same as if we had used the bidimensional measurement. These results suggest the importance of the largest perpendicular diameter when calculating tumor response, especially in lymph nodes.

In esophageal cancer and other diseases, the short axis of lymph nodes has been proven to be a better predictor of malignancy than the long axis [9,11,12,14,21]. It may also be the case that the short axis is a better predictor of response or progression. Studies have looked at sensitivity and specificity of lymph node size in determining presence or absence of tumor involvement. In many primary tumors, it is the short axis or a ratio of long to short axis that is a better predictor of malignancy. The recommendation of the use of short axis is based upon radiologic surgical correlatives demonstrating higher sensitivities and specificities using the short-axis measurement and a predefined cut-off value for metastatic lymphadenopathy [22,23]. For many other solid tumors that have a large proportion of lymph node metastases, response assessment may be greatly affected by measurement of lymph nodes. This issue is also of particular importance in patients with lymphoma, where response is based on a decrease in enlarged lymph node or confluent lymph node masses. Specific response criteria for non-Hodgkin’s lymphoma have been published [24,25]; however, the measurement of these lesions with the new criteria and its implications for response rates and approval of novel drugs is still being evaluated.

There are potentially other tumor types and sites of metastases that may be problematic for the RECIST criteria. For instance, in mesothelioma, it is difficult to reproducibly measure the long axis of a tumor and assess change. Modifications to the RECIST criteria for mesothelioma have been proposed and are currently used [26]. In some tumor types, other biomarkers may be superior or offer different information for assessing response such as in ovarian cancer and prostate cancer [27]. Combined use of tumor measurements using criteria such as RECIST and other biomarkers will play an important role in a more accurate and meaningful assessment of response.

Our findings support the proposition that the results obtained from unidimensional tumor assessment are in fact not equivalent to the results obtained from bidimensional tumor assessment. When using the RECIST criteria, we observed only two responses in the first stage of this clinical trial. However, with the WHO criteria, we observed five responses in the same patient population. Based on the two-stage design of this phase II trial, in which we were required to observe four or more responses in the first stage, only the WHO criteria would have given a response rate great enough to consider the tested combination of therapy promising enough to proceed to the second stage of the trial. Thus, the selection of RECIST criteria for determining response rate would have a significant impact on determining the effectiveness of this combination therapy for future clinical use. We suspect that this conclusion is not specific for the combination of weekly paclitaxel and bryostatin-1, but will apply to many clinical trials underway today in advanced solid tumors with other new and exciting combination therapies. This issue appears particularly critical when assessing response of solid tumors for diseases in which local or distant lymph nodes are a dominant site of metastasis.

Acknowledgements

Supported in part by RO1-001826 from the National Cancer Institute (to GKS) and in part by FDA-RO1-001826 from the Food and Drug Administration Orphan Drug Program (to GKS). Also supported by Mr William H. Goodwin and Mrs Alice Goodwin and the Commonwealth Cancer Foundation for Research, The Experimental Therapeutics Center of Memorial Sloan-Kettering Cancer Center.


References