Developing a Common Language for Tumor Response to Immunotherapy: Immune-Related Response Criteria Using Unidimensional Measurements

Mizuki Nishino, Anita Giobbie-Hurder, Maria Gargano, Margaret Suda, Nikhil H. Ramaiya, and F. Stephen Hodi

Abstract

Purpose: Immune-related response criteria (irRC) was developed to adequately assess tumor response to immunotherapy. The irRC are based on bidimensional measurements, as opposed to unidimensional measurements defined by Response Evaluation Criteria in Solid Tumors, which has been widely used in solid tumors. We aimed to compare response assessment by bidimensional versus unidimensional irRC in patients with advanced melanoma treated with ipilimumab.

Experimental Design: Fifty-seven patients with advanced melanoma treated with ipilimumab in a phase II, expanded access trial were studied. Bidimensional tumor measurement records prospectively conducted during the trial were reviewed to generate a second set of measurements using unidimensional, longest diameter measurements. The percent changes of measurements at follow-up, best overall response, and time-to-progression (TTP) were compared between bidimensional and unidimensional irRC. Interobserver variability for bidimensional and unidimensional measurements was assessed in 25 randomly selected patients.

Results: The percent changes at follow-up scans were highly concordant between the 2 criteria (Spearman r: 0.953–0.965, first to fourth follow-up). The best immune-related response was highly concordant between the 2 criteria (κw = 0.881). TTP was similar between the bidimensional and unidimensional assessments (progression-free at 6 months: 70% vs. 81%, respectively). The unidimensional measurements were more reproducible than bidimensional measurements, with the 95% limits of agreement of (−16.1%, 5.8%) versus (−31.3%, 19.7%), respectively.

Conclusion: irRC using the unidimensional measurements provided highly concordant response assessment compared with the bidimensional irRC, with less measurement variability. The use of unidimensional irRC is proposed to assess response to immunotherapy in solid tumors, given its simplicity, higher reproducibility, and high concordance with the bidimensional irRC. Clin Cancer Res; 19(14); 3936–43. ©2013 AACR.
**Translational Relevance**

Given the increasing evidence of the benefits of immunotherapeutic agents in patients with melanoma and other solid malignancies, unifying the strategy to assess response to immunotherapy is essential to provide a "common language" to describe treatment results and provide basis for further advances in cancer immunotherapy. By systematically investigating the tumor measurements record during a prospective phase II trial of ipilimumab in patients with advanced melanoma, the present study showed that immune-related response criteria (irRC) using unidimensional, longest diameter measurements provide highly concordant response assessment with better reproducibility compared with the irRC using bidimensional measurements as originally proposed. The study provides a basis for the direction toward unidimensional irRC, which is simple and practical, and provides response assessment that can be directly compared with the results from other trials based on unidimensional Response Evaluation Criteria in Solid Tumors–assessment in the past decade.

"immune-related response criteria (irRC)" in 2009, based on the discussion by 200 oncologists, immunotherapists, and regulatory experts (7). The irRC were evaluated in large, multinational studies, involving 487 patients with advanced melanoma who received ipilimumab (7). Recent phase II trial of ipilimumab in non–small cell lung cancer (NSCLC) used irRC to assess response and define endpoints (6).

The irRC published in 2009 was based on the modified WHO criteria and use bidimensional tumor measurements of target lesions, which is obtained by multiplying the longest diameter and the longest perpendicular diameter of each lesion (7). However, most trials of solid tumors in the past decade have used RECIST guidelines, which uses unidimensional, longest diameter measurements (9–11). To directly compare the efficacy and effectiveness of immunotherapeutic agents, which are simpler and more reproducible, and provide response assessment that can be directly compared with the results from trials in the past decade.

**Patients and Methods**

**Patients**

The study population included 57 patients (36 men and 21 women; mean age, 64 years; range, 39–87 years), with advanced melanoma treated with ipilimumab at Dana-Farber Cancer Institute (Boston, MA) in a phase II, multicenter treatment protocol for expanded access of ipilimumab monotherapy in subjects with histologically confirmed unresectable stage III or IV melanoma, whose prospective tumor measurement tables at baseline and at least one follow-up computed tomographic (CT) scan were available for review. In this expanded access program, the dose of ipilimumab was 10 mg/kg initially and then changed to 3mg/kg. The protocol was approved by the Dana-Farber/ Harvard Cancer Center Institutional Review Board, and all patients provided written informed consent.

**Tumor response assessment**

Tumor measurements were conducted prospectively during the trial by staff radiologists at Dana-Farber Cancer Institute at the baseline and at every follow-up CT. Follow-up scans were conducted at every 12 weeks in principle, whereas shorter interval follow-up (i.e., 4 weeks) were conducted if necessary for the purposes such as confirmation of response or progression. Tumor measurement records included the number of the treatment cycle, the date of assessment, the method of imaging, the target lesion description and bidimensional measurements, the sum of the target lesion measurements (and new lesions if any), descriptions of non-target lesions, and the presence or absence of new lesions with their bidimensional measurements if present. These records were retrospectively reviewed by a board-certified radiologist (M. Nishino) with 8 years of experience in oncologic imaging, to generate a second set of tumor measurements using the unidimensional, longest diameter measurements (7, 16).

As emphasized in the publication of WHO criteria by Miller and colleagues in 1981 in Cancer, tumor response criteria were developed due to the necessity of a "common language" to describe the results of cancer treatment and provide basis for advances in cancer therapy (15). Given the promising efficacy of newer immunotherapeutic agents, such as anti-PD-1 antibody in melanoma as well as in other solid tumors including NSCLCs and renal cell carcinoma (RCC), it is necessary to develop a "common language" for immune-related tumor response assessment to further move the field forward.

In the present study, we hypothesized that the irRC using unidimensional measurements can provide response assessment concordant with the original irRC with bidimensional measurements. We also hypothesized that the unidimensional measurements has less measurement variability than the bidimensional measurements. If these hypotheses are proven, we propose to use unidimensional, longest diameter measurements in irRC to assess efficacy and effectiveness of immunotherapeutic agents, which are
Table 1. Summary of measurement and response assessment approaches for bidimensional and unidimensional assessment based on irRC

<table>
<thead>
<tr>
<th>Bidimensional assessment (the original irRC [7])</th>
<th>Unidimensional assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measurable lesions</strong></td>
<td></td>
</tr>
<tr>
<td>≥5 × 5 mm² by bidimensional measurements</td>
<td>≥10 mm in the longest diameter</td>
</tr>
<tr>
<td><strong>Measurement of each lesion</strong></td>
<td></td>
</tr>
<tr>
<td>The longest diameter × the longest perpendicular</td>
<td></td>
</tr>
<tr>
<td>diameter (cm²)</td>
<td></td>
</tr>
<tr>
<td><strong>The sum of the measurements</strong></td>
<td></td>
</tr>
<tr>
<td>The sum of the bidimensional measurements of all</td>
<td>The sum of the longest diameters of all target lesions and new lesions if any</td>
</tr>
<tr>
<td>target lesions and new lesions if any</td>
<td></td>
</tr>
<tr>
<td><strong>Response assessment</strong></td>
<td></td>
</tr>
<tr>
<td>PD: &gt;25% increase from the nadir</td>
<td>PD: &gt;20% increase from the nadir</td>
</tr>
<tr>
<td>PR: &gt;50% decrease from baseline</td>
<td>PR: &gt;30% decrease from baseline</td>
</tr>
<tr>
<td>CR: Disappearance of all lesions</td>
<td>CR: Disappearance of all lesions</td>
</tr>
<tr>
<td><strong>New lesions</strong></td>
<td></td>
</tr>
<tr>
<td>The presence of new lesion(s) does not define progression. The measurements of the new lesion(s) are included in the sum of the measurements.</td>
<td></td>
</tr>
</tbody>
</table>
Interobserver variability was assessed using concordance correlation coefficients (CCC), mean relative difference (%), and 95% limits of agreement (%) for the unidimensional, longest diameter (cm) and the bidimensional measurements. CCC was used to assess reproducibility of 2 measurements, as described previously (13–14). Assuming 2 measurements have mean $\bar{u}_1$ and $\bar{u}_2$, with variance $\sigma_1^2$ and $\sigma_2^2$ and covariance $\sigma_{12}$ and $CCC = \frac{\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\bar{u}_1 - \bar{u}_2)^2}$. CCCs were composed of a measure of precision (how far each pair of measurements deviates from the best-fit line through the data) and a measure of accuracy (the distance between the best-fit line and the 45 line through the origin). A value of 1 indicates perfect agreement and −1 indicates perfect reversed agreement (18). Agreement in the 2 measurements was shown visually using Bland–Altman plots with 95% limits of agreement and the average relative difference, computing the mean relative difference (%) between the 2 measurements ($100 \times \frac{M_2 - M_1}{M_1}$) = measurements during trial, $M_2$ = measurements by the radiologist in this study; ref. 14). All $P$-values are based on a 2-sided hypothesis. $P < 0.05$ was considered to be significant.

**Results**

**Bidimensional versus unidimensional tumor response assessment**

Figure 1 shows the percent changes according to bidimensional and unidimensional measurements at each follow-up scan, including the 1st to 17th follow-up (f/u) scans. The percent changes by 2 measurements were highly concordant, with Spearman correlation coefficient of 0.959 (95%CI, 0.93–0.98) for the 1st f/u ($n = 57$); 0.963 (0.92–0.98) for the 2nd f/u ($n = 33$); 0.953 (0.88–0.98) for the 3rd f/u ($n = 21$); and 0.965 (0.87–0.99) for the 4th f/u ($n = 12$). The number of patients were too small (≤5) after the 4th follow-up to obtain a reliable estimate. Response assessment results by 2 measurements on the first 3 follow-up scans had almost perfect agreement, with $\kappa_w$ values of 0.844 for the 1st ($n = 57$), 0.830 for the 2nd ($n = 33$), and 0.861 ($n = 21$) for the 3rd follow-up (Figs. 1 and 2).

The best immune-related response according to two measurements showed almost perfect agreement between the 2 criteria ($\kappa_w = 0.881$, Table 2). Best response assessments by 2 criteria were identical in 53 of 57 patients (93%). The remaining 4 patients (7.0%) had discordant results, including 3 with irPD by bidimensional measurements and irSD by unidimensional measurements and one with irSD by bidimensional measurements and irPD by unidimensional measurements. Forty-one patients (72%) had irSD as the best immune-related response according to both measurements.

Kaplan–Meier estimates of TTP are shown in Fig. 3. At 6 months, 70% of patients were found to be free of progression using the bidimensional assessment, compared with 81% using the unidimensional assessment. Estimates of the 25th percentile (time point at which 75% are free of progression) were 5.3 months (95% CI, 3.5–6.9) by

---

**Figure 1.** The percent changes according to bidimensional and unidimensional measurements at each follow-up scan from the 1st to 17th follow-up scans. The orange dashed lines represent the cut-off values for response and progression (−50% and +25% for bidimensional measurements, −30% and +20% for unidimensional measurements). The observations within the top left, middle center, and top right boxes have concordant assessment between two measurements, whereas observations in other boxes have discordant assessment. The purple dashed line represents a 44% change for bidimensional measurements, which corresponds to a 20% change for unidimensional measurements, which was given to visually demonstrate that more observations are concordant if this cut-off value is used. The percent changes presented in the figure are in comparison with baseline measurements when tumors are decreasing to assess response and in comparison with the nadir (the smallest measurement since baseline) when tumors are increasing to assess progression. These values are displayed as they are used to define response/progression in patients at the time of response assessment.
bidimensional assessment versus 9.1 months (95% CI, 3.7–∞) by unidimensional assessment. On the basis of the almost identical confidence intervals for the 25 percentile, there is no evidence of a difference in TTP between the 2 methods of assessment.

Reproducibility of bidimensional versus unidimensional measurements

In 25 randomly selected patients, the CCCs between the measurements conducted during the trial and the measurements by the radiologist conducted in this study were 0.986 (95% CI, 0.972–0.993) for bidimensional measurements and 0.995 (95% CI, 0.989–0.998) for unidimensional measurements (Table 3).

Bland–Altman plots with 95% limits of agreement and the average relative difference are shown in Fig. 4. The 95% limits of agreement of bidimensional measurements were (−31.3%, 19.7%), that were twice wider compared with (−16.1%, 5.8%) for unidimensional measurements.

Discussion

The present study showed that the immune-related response assessment using unidimensional, longest diameter measurements was highly concordant with the assessment based on bidimensional measurements in patients with advanced melanoma treated in a clinical trial of ipilimumab. The unidimensional measurements had less measurement variability than bidimensional measurements. The results of the study provide a basis for using unidimensional measurements in immune-related tumor response assessment. The study also serves as an initial step to further optimize response assessment in patients treated with immunotherapeutic agents, toward developing a “common language” for immune-related response.

Highly concordant response assessment at each follow-up between bidimensional and unidimensional measurements was noted, with almost perfect agreement between response assessment categories by 2 assessments at the first 3 follow-up scans, which was consistent with our initial expectation. Of note, the high concordance was showed despite of the difference of the cutoff value scales for progression according to bidimensional and unidimensional assessment. Twenty per cent increase in unidimensional measurements corresponds to 44% increase in bidimensional measurements, according to the mathematical conversion provided by RECIST (9). As shown in Fig. 1, the use of the scaled value of 44% for progression by bidimensional measurements would have resulted in even higher agreement between the 2 assessments. On the other hand, 25% increase by bidimensional measurements corresponds to approximately 12%
increase by unidimensional measurements. We did not apply this scaled value due to the concern that 12% unidimensional increase is within the measurement variability and therefore can be attributed to measurement error rather than true increase of tumor, which was supported by the reproducibility results of the present study.

Best immune-related response had almost perfect agreement by weighted kappa analysis, which was consistent with our hypothesis. Most patients (41 of 57, 72%) in the study had the best response of irSD by both assessments, because of the requirement of confirmation for irCR, irPR, and irPD. All 4 patients with discordant best immune-related response were in irPD versus irSD categories, with 3 patients having irPD by bidimensional assessment, whereas these had irSD by unidimensional assessment. Among these 3 patients, one patient was alive after 36.4 months since the initiation of therapy, which was 3 times longer than the median OS of 10.1 months (95% CI, 8.0–13.8) in a phase III trial of ipilimumab in patients with melanoma (1). Other 2 patients died after 13.3 months and after 8.4 months, which were within the 95% CIs of the reported median OS (1). One patient with irSD by bidimensional assessment and irPD by unidimensional assessment died after 22.5 months since the initiation of therapy. The data from the small cohort evaluated by this retrospective study are limited to address the important question of association between survival and response assessment. The question needs to be addressed in a larger prospective cohort. The discordance could also be related to the difference in cutoff values, as bidimensional 25% increase may require smaller increase than unidimensional 20% increase. Requiring smaller increase for progression is subject to higher rate of misclassification due to measurement variability, especially when the cutoff values are within the range of measurement errors (12).

There was no evidence of a difference in TTP by 2 criteria; however, the majority of patients did not progress during the study and therefore censored by both assessments. This is partly due to the requirement of confirmation for all categories except for irSD, which is one of the unique features of irRC. Because of the same reason, median TTP could not be obtained, which is one of the limitations of the present study. We followed this requirement as it was implemented to capture additional response pattern specific to immunotherapy, that is, decrease of tumor burden after initial progression.

Unidimensional measurements were more reproducible than bidimensional measurements, which was concordant with our initial hypothesis as well as previous reports (12–14). The 95% limits of agreement for bidimensional measurements were twice larger than those for unidimensional measurements. It should also be noted that 25% change for bidimensional measurements are within the measurement error and therefore cannot be reliably used to define progression. On the other hand, the cutoff values for the percent change applied for the unidimensional measurements (−30% for PR and +20% for PD) were beyond the range of measurement variability and therefore can be considered to reflect true change of tumor burden, rather than measurement error (12–14). The cutoff values used for unidimensional measurements in the present study were based on RECIST guidelines (−30% for PR and +20% for PD; refs. 9–10). We chose these cutoff values because (i) these values are widely accepted in response assessment using unidimensional measurements and (ii) the results obtained using these values can be directly compared with the results of prior trials and studies based on RECIST (10). The capability of directly comparing the trial results in patients with other solid tumors with other systemic anti-cancer agents is becoming increasingly important as newer immunotherapeutic agents are tested and approved for a variety of solid tumors (19, 20).

The current study assessed the measurement variability of 25 randomly selected patients. We based this approach on

<table>
<thead>
<tr>
<th>Number</th>
<th>Events</th>
<th>% progression-free at 6 months (S.E.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bidimensional</td>
<td>57</td>
<td>10</td>
</tr>
<tr>
<td>Unidimensional</td>
<td>57</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 3. Interobserver measurement variability

<table>
<thead>
<tr>
<th></th>
<th>CCC (95% CI)</th>
<th>Mean relative difference, %</th>
<th>95% limits of agreement, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bidimensional</td>
<td>0.986 (0.972–0.993)</td>
<td>–5.8</td>
<td>–31.3, 19.7</td>
</tr>
<tr>
<td>Unidimensional</td>
<td>0.995 (0.989–0.998)</td>
<td>–5.1</td>
<td>–16.1, 5.8</td>
</tr>
</tbody>
</table>

Figure 3. TTP according to bidimensional versus unidimensional assessment.
past investigations showing that unidimensional measurements were more reproducible than bidimensional measurements. Measurement variability is an important issue in the context of defining the adequate cutoff value for response and progression and remains to be systematically investigated in a larger population of patients during immunotherapy.

Limitations for this analysis include the retrospective design for the unidimensional response assessment. However, the tumor measurement records used in the study were prospectively acquired during the trial. The number of patients included in the analysis was relatively small and was from a single institution. The association between clinical outcome and response assessment results needs to be investigated, which constitutes an important next step to establish an appropriate surrogate marker in cancer immunotherapy.

In conclusion, the irRC using unidimensional tumor measurements provided highly concordant response assessment and had less measurement variability compared with the irRC with bidimensional measurements. Additional investigation is warranted to in a larger cohort with correlations with clinical outcomes and assessments by multiple radiologists for reproducibility to propose the longest axis measurements for tumor response assessment during immunotherapy. It is also necessary to test our observations in patients with other solid tumors treated with other immunotherapeutic agents to evaluate the broader applicability of the results. We are currently planning to validate the observation in a larger cohort and to systematically investigate the measurement variability to determine adequate cutoff values for response and progression to accurately characterize immune-related response and progression during immunotherapy.

Disclosure of Potential Conflicts of Interest
F.S. Hodi has served as a non-paid consultant to Bristol-Myers Squibb and has received clinical trial support from Bristol-Myers Squibb. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions
Conception and design: M. Nishino, N.H. Ramaiya, F.S. Hodi
Development of methodology: M. Nishino, N.H. Ramaiya, F.S. Hodi
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M. Nishino, M. Gargano, M. Suda, F.S. Hodi
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M. Nishino, A. Giobbie-Hurder, F.S. Hodi
Writing, review, and/or revision of the manuscript: M. Nishino, A. Giobbie-Hurder, M. Gargano, M. Suda, N.H. Ramaiya, F.S. Hodi
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M. Gargano, M. Suda, F.S. Hodi
Study supervision: N.H. Ramaiya

Grant Support
The investigator M. Nishino was supported by 1K23CA157631 (NCI) and Dana-Farber Cancer Institute Fellowship for the Eleanor and Miles Shore 50th Anniversary Fellowship Program.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 29, 2013; revised May 16, 2013; accepted May 23, 2013; published OnlineFirst June 6, 2013.

References
2. Weber J, Thompson JA, Hamid O, Minor D, Amin A, Ron I, et al. A randomized, double-blind, placebo controlled, phase II study comparing the tolerability and efficacy of ipilimumab administered with or 3. A randomized, double-blind, placebo controlled, phase II study comparing the tolerability and efficacy of ipilimumab administered with or


Developing a Common Language for Tumor Response to Immunotherapy: Immune-Related Response Criteria Using Unidimensional Measurements

Mizuki Nishino, Anita Giobbie-Hurder, Maria Gargano, et al.


Updated version Access the most recent version of this article at: doi:10.1158/1078-0432.CCR-13-0895

Cited articles This article cites 20 articles, 7 of which you can access for free at: http://clincancerres.aacrjournals.org/content/19/14/3936.full.html#ref-list-1

Citing articles This article has been cited by 15 HighWire-hosted articles. Access the articles at: /content/19/14/3936.full.html#related-urls

E-mail alerts Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.