Volume-based Response Evaluation with Consensual Lesion Selection: A Pilot Study by Using Cloud Solutions and Comparison to RECIST 1.1

Estanislao Oubel, PhD, Eric Bonnard, MD, Naoko Sueoka-Aragane, MD, PhD, Naomi Kobayashi, MD, Colette Charbonnier, PhD, Junta Yamamichi, MSc, MPH, Hideaki Mizobe, MSc, Shinya Kimura, MD, PhD

Rationale and Objectives: Lesion volume is considered as a promising alternative to Response Evaluation Criteria in Solid Tumors (RECIST) to make tumor measurements more accurate and consistent, which would enable an earlier detection of temporal changes. In this article, we report the results of a pilot study aiming at evaluating the effects of a consensual lesion selection on volume-based response (VBR) assessments.

Materials and Methods: Eleven patients with lung computed tomography scans acquired at three time points were selected from Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) and proprietary databases. Images were analyzed according to RECIST 1.1 and VBR criteria by three readers working in different geographic locations. Cloud solutions were used to connect readers and carry out a consensus process on the selection of lesions used for computing response. Because there are not currently accepted thresholds for computing VBR, we have applied a set of thresholds based on measurement variability (−35% and +55%). The benefit of this consensus was measured in terms of multiobserver agreement by using Fleiss kappa ($k_{fleiss}$) and corresponding standard errors (SE).

Results: VBR after consensual selection of target lesions allowed to obtain $k_{fleiss} = 0.85$ (SE = 0.091), which increases up to 0.95 (SE = 0.092), if an extra consensus on new lesions is added. As a reference, the agreement when applying RECIST without consensus was $k_{fleiss} = 0.72$ (SE = 0.088). These differences were found to be statistically significant according to a z-test.

Conclusions: An agreement on the selection of lesions allows reducing the inter-reader variability when computing VBR. Cloud solutions showed to be an interesting and feasible strategy for standardizing response evaluations, reducing variability, and increasing consistency of results in multicenter clinical trials.

Key Words: Clinical trials; RECIST; cloud computing; consensus; lesion volume; biomarkers; volume thresholds.

©AUR, 2015

According to recent statistics of the World Health Organization, cancer is the leading cause of death worldwide accounting for 8.2 million deaths in 2012 (1). Lung cancer accounted for 1.6 million deaths (19.4%), which makes it the first cause of cancer death even ahead of liver (9.1%) and stomach (8.8%) cancers. Computed tomography (CT) is currently the standard imaging modality for assessing the response to treatment in patients with solid tumors. In general, the quantification of the response is performed by using Response Evaluation Criteria in Solid Tumors (RECIST) (2,3). In summary, this standard establishes the way of measuring lesions and provides a set of thresholds to classify the response into partial response, stable disease, and progressive disease. In the context of this article, the term RECIST refers to its revised version (1.1) (3).

Even when RECIST criteria have been broadly adopted in clinical trials, they present some drawbacks that are a source of measurement variability. This is quite inconvenient because the consistency in the production of trial results is necessary for comparison purposes (4). For example, the lesion size is measured as its longest axial diameter, which is not a robust measure in case of complex lesions and creates problems of accuracy and precision (5). To cope with this drawback, the use of volume is currently being considered as a promising direction to make tumor measurements more accurate and consistent, which would enable an earlier detection of temporal changes (6–9). The benefit of using volume as biomarker has already been reported in the literature (5,10); however, the use of volume also presents some
drawbacks like the long segmentation time in case of big lesions (in particular for thin-slice acquisitions) and the lack of accepted thresholds for computing response. Regarding this last point, it is worth mentioning the intensive efforts performed by the Quantitative Imaging Biomarkers Alliance (11) to better understand volumetric biomarkers and their sources of variability. In the context of this article, the term volume-based response (VBR) refers to the response estimated exactly as established by RECIST, except for the use of volume of lesions and volume-specific thresholds instead of diameters. The use of VBR was preferred to names like 3D-RECIST to avoid confusion with the application of RECIST with 3D-extended thresholds and also because currently there is no formal definition of the volumetric version of RECIST.

Another reported source of variability of RECIST is the difference in the target lesions (TLs) selected for computing response (12,13). TLs are selected on the basis of their size and their suitability for reproducible measurements (6). Even when it is relatively easy to measure the lesion size required for its inclusion, it is more difficult to assess measurement repeatability by visual inspection only. For example, the edges of irregular or infiltrating lesions are often difficult to define and, in some cases, they are even impossible to measure (14). Another commonly found obstacle is the presence of peritumoral fibrosis, which is difficult to distinguish from tumor spread and adds further uncertainty to measurements. A very interesting discussion about limitations of RECIST can be found in (15).

To cope with the problems of variability mentioned before, regulatory authorities have been recommending the use of an Independent Central Review (ICR). The original purpose of ICR as recommended by the Food and Drug Administration (FDA) was to eliminate the bias associated with local evaluation (LE). This is thought to be relevant for studies in which the investigator knows the patients enrolled in different trial arms. Another advantage of ICR is the possibility of standardizing evaluations because all patients are evaluated by a restricted number of reviewers (typically one reviewer and one adjudicator) using the same software solutions. However, the advantages of ICR over LE have been questioned in the literature (16,17,18), and some alternatives to ICR like the use of audit tools are currently being considered (19,20). The main criticism is that ICR does not remove bias completely, and in some cases, it may introduce bias by itself (eg, by informative censoring). Finally, the implementation and management of the ICR process is costly and burdensome for sites and sponsors.

Several analyses, notably the one by the Pharmaceutical Research and Manufacturers Association Working Group (21), show that no systematic bias is introduced by LE. The same conclusions have been presented recently by the FDA, who proposes the use of ICR only as an audit tool to detect evaluation bias in LE assessments (19). However, in an LE approach, there is an intrinsic variability among sites because evaluation protocols cannot be guaranteed to be the same or equivalent. Therefore, a standardization of the evaluation could make results from different sites more consistent.

Cloud computing offers an opportunity for reducing the variability in the application of RECIST in a LE approach because it can facilitate the interaction between different stakeholders and provide common tools for image analysis (which reduces the variability coming from the use of different measurement systems). The application of cloud computing in the context of clinical trials has become possible, thanks to the availability of high-capacity networks, low-cost computers, and storage services. The huge amount of data generated by medical imaging acquisition systems makes cloud computing an interesting alternative for processing, storing, and sharing images. Besides, cloud free customers (eg, hospitals) from the responsibility of installing and maintaining hardware and basic computational services because these tasks are performed by cloud providers.

In this article, we have evaluated a cloud-based system to perform a consensus on the selection of lesions used for computing VBR. As suggested by previous publications (12,13), this is expected to reduce the inter-reader variability in LE approaches. Even when in this article we focus on the specific problem of lesion selection, the proposed workflow can be extended to control other sources of variability. The final objective is to make the application of RECIST more consistent among sites. Several contributions are provided in this article. First, we show the feasibility of applying cloud computing in the context of multicenter clinical trials. To do this, we propose a novel cloud-based workflow enabling a more consistent application of RECIST among participant sites. Second, we applied this workflow to investigate the potential benefits of a lesion consensus on the response. This is only one specific application among others that illustrates how the system could be used, and the advantages provided. Third, we compare the proposed approach with respect to the current standard and LE. Finally, we analyze the impact on the RECIST compliance, which is expected to be higher after a consensus process.

MATERIALS AND METHODS

Data Set

Eleven patients (aged 62 years in average; six men and 5 women) presenting solid lesions were retrospectively selected from RIDER (22) (6 patients) and Median Technologies’ proprietary databases (5 patients). The selection criteria were the type of lesion (solid), location (lung), number of acquired time points (at least at baseline, 3, and 6 months), and homogeneity of image acquisition parameters (eg, filter type). Chest CT scans with contrast were performed by using General Electric LightSpeed and Siemens Sensation scanners. Image acquisition was performed at 120kVp and (354 ± 103 mAs). The slice thickness was ≤2.5 mm (1.93 ± 0.63 mm), and the in-plane resolution was ≤0.78 mm (0.71 ± 0.06 mm). A filtered back projection reconstruction method was used in all cases with B60f (Siemens) and Lung (GE) convolution kernels.

Lesions were selected and measured by two oncologists with 27 and 12 years of experience (the on-site readers) and
a radiologist with 5 years of experience (the independent reviewer [IR]). The oncologists are specialized in clinical trials and perform RECIST evaluations in the clinical routine. Figures 1–4 show the distribution of the selected lesions according to different characteristics.

Cloud-based Setup System

Figure 5 shows the set up cloud solution and the different stakeholders involved in the reading process, which can be explained in a general manner as follows:

1) The Data Managers (DMs) perform a quality control of images and, if the quality control test is successful, they are transferred to the data center where the processing required before analysis is performed (image identification, image reconstruction, lung mask extraction, and so forth). Once the processing is completed, the DMs make the images available to the IR and readers for analysis.

2) The reader recovers the images of a particular patient from the data center and performs a lesion segmentation and response assessment according to RECIST and VBR. The performed analysis is saved into the data center for it to be reviewed by the IR.

3) The reader and IR start exchanging opinions about the selected TLs, and the reader is allowed to modify his selection if the feedback provided is considered to be relevant. This process named “consensus” is coordinated by the DMs, who are in charge of keeping records of changes in evaluations.

4) Once an agreement or maximum number of iterations (set to four in this study) is reached, the final response is calculated. For clarity and readability purposes, sometimes it is necessary to emphasize under which conditions (with or without consensus) the response was calculated. Therefore, we use a superscript (asterisk) to distinguish between responses computed with and without consensus. In this way, RECIST* and VBR* refer to responses calculated...
The geographic location of the different stakeholders was different, which allowed evaluating the solution for a potential use in multicountry clinical trials. DMs were based in Sophia Antipolis (France), the readers in Saga (Japan) and Glasgow (United Kingdom), the IR in Nice (France), and the data center (Canon IT Solutions Inc.) was installed in Tokyo (Japan). A more detailed workflow of the reading process is shown in Figure 6.

Volume Measurement

The lesion volume was computed from segmentations performed by using Lesion Management Solutions (Median Technologies, Sophia Antipolis, France). If necessary, the initial results provided by the semiautomatic segmentation were corrected by using interactive segmentation tools implementing shape-based interpolation methods (23). To remove normal lung tissue from the lesion, a final threshold step is automatically applied after each correction performed by the user.

Response Agreement

Target Lesions. We have evaluated the impact of the consensus on the inter-reader response agreement by comparing TL-based responses before and after consensus. As explained before, the response without consensus is always assessed before IR’s feedback, and it is potentially modified after this. Therefore, both responses are available, and a comparison of agreements can be carried out. We have quantified the inter-reader agreement by using the Fleiss Kappa ($k_{fleiss}$) statistic (24). The Fleiss Kappa allows computing the agreement between multiple observers when assigning categorical ratings to a number of statistical units.

The inter-reader agreement was also computed for the tumor burden (sum of lesion volumes) and its change over time. The objective of this analysis was to evaluate the impact of the consensus on the measure itself, independently of thresholds. This is important because the use of thresholds could hide differences in the continuous values, which is an interesting result even if no changes are observed in the categorical data. This agreement was quantified by using reproducibility coefficient values (25), which can be computed for multiple observers. A Shapiro-Wilk test of normality (26) was performed, and a logarithmic transformation was applied to data in case of null hypothesis rejection for normalization purposes.

New Lesions. The consensus process was performed on TLs only. However, the impact of new lesions (NLs) on the response is strong because the appearance of at least one NL changes automatically the response status to progress disease according to RECIST, independently of changes in TLs. Therefore, the effect of a potential consensus on NL on the response deserves special consideration. To analyze this, we simulated a consensus on NL by assigning to all readers a NL identified by at least one reader. For example, suppose that the reader #1 identifies an NL at a given time point of a given patient, but this lesion was missed by both reader #2 and IR, then this NL is also taken into account for computing the responses of reader #2 and IR. The Fleiss Kappa was also used to measure the inter-reader agreement.
**Volume Thresholds.** Currently, there are no accepted thresholds for assessing response based on volume changes. One could simply apply RECIST thresholds after extrapolation to three dimensions, but, according to Mozley et al. (10), this is known to work well only in some situations where the tumor morphology is simple, but it cannot be applied in a general manner. Therefore, we preferred to apply thresholds empirically estimated from measurements of interobserver variability and established to +55% for progression and −35% for response (27). The rationale behind the use of this set of thresholds is that the actual differences in lesion size can be hidden by the variability of the measure.

Then, for a difference in size to be measured significantly, it must be necessarily larger than this variability. This way of estimating thresholds is similar to the approach followed by Mozley et al. (10), the main difference is the application of the Geary–Hinkley transformation to model response evaluation.

**Data Analysis.** All statistical analyses were performed with the R software (28). The Fleiss kappa was computed by using the irr package (29), and the Shapiro-Wilk test of normality was computed by using the stats package. A z-test on original and bootstrapped data was applied to assess the statistical significance of differences in $\kappa_{\text{Fleiss}}$.

---

**Figure 6.** Unified Model Language diagram of the reading process. The sequence part inside the loop corresponds to the consensus process aiming to apply RECIST uniformly. IR, independent reviewer.
TABLE 1. Fleiss Kappa and Standard Errors (Between Parentheses) According to Different Types of Consensus and Criteria

<table>
<thead>
<tr>
<th>Consensus</th>
<th>TL Only Response</th>
<th>RECIST</th>
<th>VBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL</td>
<td>—</td>
<td>X</td>
<td>0.72 (0.808)/SuA</td>
</tr>
<tr>
<td>NL</td>
<td>—</td>
<td>—</td>
<td>0.72 (0.090)/SuA</td>
</tr>
<tr>
<td>X</td>
<td>—</td>
<td>X</td>
<td>0.72 (0.088)/SuA</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0.72 (0.088)/SuA</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>—</td>
<td>0.81 (0.089)/APA</td>
</tr>
</tbody>
</table>

APA, almost perfect; NL, new lesion criteria; RECIST, Response Evaluation Criteria in Solid Tumors; SuA, substantial; TL, target lesion; VBR, volume-based response.

Consensus on TL, NL, and TL only response has been marked with “X” when they are applied and “—” when they are not. The interpretation of values according to Landis and Koch (30) has been added as follows: slight/fair/moderate/substantial/almost perfect agreement. The highest value is shown in bold.

RESULTS

Response Agreement

Table 1 shows the inter-reader agreement measured by Fleiss kappa agreements (24) for different types of consensus and response criteria. The table shows that the use of volume provides higher levels of agreement than RECIST. The table also shows that, differently from one-dimensional measurements, the VBR does benefit from the consensus on TLs (RECIST and RECIST* provide similar results). A paired sample z-test for mean of paired differences between VBR* and RECIST rejected the null hypothesis in favor of the alternative hypothesis of higher agreement for VBR* (z-value = 1.77; P value = .038). This means that the difference found between both approaches is statistically significant. A z-test was also applied to compare VBR* and VBR (ie, to measure the effects of consensus only). In this case, the agreement with consensus was significantly higher than without consensus at a 10% level (z-value = 1.46; P value = .072). Finally, when comparing RECIST versus VBR, differences were not statistically significant in favor of volume (z-value = 0.31; P value = .378).

We have also performed z-tests on 1000 bootstrapped combinations of the original ratings. In this case, the z-test rejected the null-hypotheses with P value <.01 for all three comparisons performed before.

Table 2 shows reproducibility index values (25) for the sum of lesion volumes (SLV), its change with respect to baseline (ΔSLVBL), and its change with respect to nadir (ΔSLVNADIR). A Shapiro–Wilk test of normality showed that the distributions of these parameters were not normal. To reduce these deviations from normality, the values were transformed by using the logarithmic function. The success of this correction was verified by a new application of the normality test which showed that the null hypothesis could not be rejected. Table 2 shows that the reproducibility index values decrease after consensus (ie, the inter-reader agreement increases), which is consistent with the results presented in Table 1. These results provide further evidence about the reduction of variability when there is an agreement on the lesions to be used for response assessment.

Comparison Between VBR and RECIST

Sometimes, the response agreement may be meaningless. For example, if thresholds are too high, only one type of response and 100% of agreement are obtained; however, this result is meaningless because the set of thresholds is not discriminative. As RECIST is an already validated and broadly accepted criteria by the scientific community, a new response criteria is expected to provide similar results for a large population (but not exactly the same for the new criteria to be interesting). The kappa value between RECIST and VBR* on TLs and NLS was equal to 0.53 (SE = 0.152), which corresponds to a moderate agreement according to the classification by Landis and Koch (30).

Besides the kappa value itself, it is interesting to analyze the contingency table used for its calculation. Table 3 shows that the matrix of contingency presents some asymmetries deserving further analysis. Even when the lack of ground truth precludes an analysis of accuracy, a comparison between methods can be performed. When rows and columns of the contingency matrix are ordered by decreasing order of response, this matrix becomes close to upper triangular,
which means that VBR* presents a positive bias with respect to RECIST. In other words, for each category of response provided by VBR*, the response provided by RECIST is equal or inferior. For example, when a partial response is provided by VBR*, RECIST provides either partial response or stable disease; when VBR* is stable disease, RECIST says stable disease or progressive disease. Of course, this is not valid for the last row because this matrix is not strictly upper triangular.

**Number of Lesions**

Some articles in the literature (31,32) suggest that the number of lesions is important when computing response. More specifically, the use of a large number of lesions has been shown to reduce the interobserver variability. We have analyzed this aspect for our data by using volume measurements before consensus to assess if this is actually a source of variability. This is interesting because the number of selected lesions is a parameter of lesion selection that could be potentially controlled during the consensus process, which would be an interesting extension of the workflow presented in this article. For example, if there is a difference in the number of selected lesions between readers and if an influence on the variability of such differences is proven, the IR could ask one of the readers to select additional lesions to match the number of lesions selected by the second reader.

We have compared the responses between readers by taking two readers at the time. The results presented in Table 4 confirm that the number of lesions used for computing response is important for reducing variability. In cases where four lesions were chosen, 100% of response agreement was achieved. This supports the claims by Moskowitz et al. (31) who propose a minimum number of five lesions for avoiding response misclassification. These results are also in full agreement with Darkeh et al. (32), who discourage the use of less than four lesions because inter-rater discrepancies may be introduced.

**RECIST Compliance**

Before consensus, we have found 12% of response evaluations noncompliant with RECIST, whereas after consensus all evaluations were RECIST-compliant. The following RECIST nonconformities have been found as follows:

- Number of lesions per organ higher than two. For example, one reader selected three pulmonary TLs, but the maximum number of lesions per organ established by RECIST is two.
- Lesion location mistakes. For example, pulmonary lesions labeled as mediastinal.
- Measurability problems. For example, one of the readers chose lesions with ill-defined contours next to the pulmonary artery, which could not be measured accurately.

Noncompliant RECIST evaluations (and sources of deviation) have been already reported in the literature (33), and therefore, these results confirm and are consistent with previously published results.

**DISCUSSION**

Table 1 suggests that the RECIST-based response agreement is independent of the consensus on TLs. Indeed, the choice of the same TLs does not guarantee a lower variability. This situation may seem to be paradoxical, but the following example provides an explanation to this observation. Let us consider, for example, the lesion shown in Figure 7. This lesion is quite irregular and the measured diameters are very different, which leads to an also high response variability even if all readers choose only this lesion as target. On the other hand, the same Table 1 shows that a consensus on NLs improves the agreement, which owns probably to the strong effect of this type of lesions on the response.

Table 1 shows that the consensus on TLs has an impact on the VBR agreement; however, no differences were observed for RECIST. These differences may be explained by a higher reproducibility of volume measurements with respect to diameter (5,6,10). The inter-reader agreement for VBR* was significantly higher than RECIST and VBR. This is in agreement with results published recently by Zhao et al. (13) and Khul et al. (12).

We have found a moderate agreement between VBR* and RECIST. However, this agreement is not a metric of performance because RECIST is not a gold standard for response. In fact, if a new response index provides the same results as RECIST ($\kappa = 1$), it would be completely useless. On the other hand, $\kappa = -1$ is not wished either because this would mean that the new index is saying the opposite than RECIST. This would mean in turn that the proposed response index presents major drawbacks, which would preclude its use in clinical trials. Therefore, a $\kappa = 0.53$ seems to be a good compromise because it implies neither perfect agreement nor completely opposed results and completely random differences.

Without consensus, the selection of lesions was noncompliant with RECIST requirements in 12% of cases. The existence of non-compliances has already been reported by Skougaard et al. (33) who found 46% of inter-reader discrepancies associated with non-conformities in the application of RECIST. In this study, 100% of evaluations were RECIST- compliant.
compliant after consensus. This is an important aspect of the workflow explained in the section on cloud-based setup system for quality control purposes.

In this article, we have applied cloud-based solutions for achieving a consensus on the selection of lesions. However, the proposed framework can be applied for the standardization of clinical trials in a broader sense. For example, the lesion segmentation could be performed only on specific images of the whole set of images corresponding to a given time point. Another example is the use of similar visualization settings (zoom, window level, and so forth) for all readers. The number of lesions used for computing response could also be controlled because we have observed in the section on number of lesions that this is a source of variability. These standardizations could potentially reduce the variability of measurements and can be easily included as a part of the proposed workflow.

One limitation of this study is the low number of patients of the data set. This is the result of applying multiple image selection criteria when creating the data set as described in the section on data set. The idea was to remove the maximum number of sources of variability related to image characteristics to focus only on the effects of lesion selection. It is important to take into account that the statistical analyses were performed on 66 responses (three time points per patient evaluated by three readers), which is an acceptable number of statistical units to perform such analyses. For the purposes of a pilot study like this, the number of patients included was quite convenient because it avoided the complexity associated with the management of large data sets and to test the workflow promptly.

The consensus process increases necessarily the reviewing time. However, in some clinical centers, this additional time is spent anyway in discussions between radiologists and oncologists to agree on lesion selection. Another aspect to take into account is that, in general, physicians follow different and more time-consuming protocols when evaluating patients in the context of a clinical trial (use of specific software, different measurements, and filling of case report forms); therefore, the time invested in performing a consensus seems to be compliant with clinical trial protocols. Finally, this additional time may reduce the intersite variability, which in turn may eventually improve the quality of the generated statistical data.

In a real clinical context, several sites must be managed. The use of cloud solutions reduces the installation complexity because, differently from a LE approach, it is not necessary to perform specific installations at each site. However, to manage the consensus, additional DMs and IRs might be required. Clouds also allow simplifying the workflow and the communication between stakeholders. For example, reports can be signed electronically by the adjudicator and transmitted immediately to the sponsor, needless to exchange reports in paper format. Another example is the update of the sponsor's database: once the radiologist finishes an evaluation, the site can connect directly to the sponsor's database to update it with the results for the corresponding patient. In this way, data are immediately available for analysis. The same idea is applicable to images: once a scan is available for a specific patient, it can be imported into the system through a computer terminal and then transmitted to the sponsor to be included in a central database. This solves the problem of sending images by using support materials like digital versatile disks (DVDs) or hard drives.

One of the main concerns with cloud solutions is confidentiality. To keep data confidential, the cloud infrastructure and data were centrally hosted in a data center managed by a vendor guaranteeing security and confidentiality. The data center provides early warning systems to identify potential attacks. These types of infrastructures are generally more secure than hospital infrastructures. Besides, our cloud service requires strong authentication to identify users and provides an audit trail to record the actions performed by authorized users. Our system is compliant with regulations such as CFR21 Part 11 (FDA—Electronic records, electronic signature) and International Committee for Harmonization—Good Clinical Practices to guarantee the application of security, confidentiality, and safety best practices. Finally, the communication between the client and the server hosted by the data center is encrypted over the Secure Sockets Layer/Transport Layer Security protocol. We are aware that cloud security is a major concern because data are transferred across the network, and therefore, we use state-of-art technology to make data transfer as safe as possible. We think that, with the use of such technology, data are probably as well protected in a cloud infrastructure as in a local infrastructure.

Figure 7. Differences in assessment of longest axial diameter between readers. The figure shows the slice containing the LAD (blue line) and the short axis diameter (red line). (a) Reader #1, (b) Reader #2, and (c) independent reviewer. (Color version of figure is available online.)
CONCLUSIONS
In this article, we have set up a cloud-based system aiming at standardizing response evaluations in multicenter clinical trials. The use of cloud solutions showed to be an interesting and feasible strategy for standardizing response evaluations and reducing variability. The use of VBR increased the inter-reader agreement and the RECIST compliance, and therefore, it seems to be a promising alternative for evaluating the response to a treatment. For VBR, the consensus improved the inter-reader response agreement and the RECIST compliance, and therefore, it is an interesting phase to be considered in a clinical trial protocol. Finally, the consensus on NLs showed to play an important role because of their strong impact on the global response, and therefore, it deserves special attention as a topic of future research.

REFERENCES