How Does Target Lesion Selection Affect RECIST?

A Computer Simulation Study

Tareco Bucho, Teresa M; Tissier, Renaud L M; Groot Lipman, Kevin B W; Bodalal, Zuhir; Delli Pizzi, Andrea; Nguyen-Kim, Thi Dan Linh; Beets-Tan, Regina G H; Trebeschi, Stefano

Published in:
Investigative Radiology

DOI:
10.1097/RLI.0000000000001045

Publication date:
2024

Document version:
Final published version

OPEN

How Does Target Lesion Selection Affect RECIST? A Computer Simulation Study

Teresa M. Tareco Bucho, MSc, Renaud L.M. Tissier, PhD, Kevin B.W. Groot Lipman, MSc, Zuhir Bodalal, MD, MSc, Andrea Delli Pizzi, MD, PhD, Thi Dan Linh Nguyen-Kim, MD, Regina G.H. Beets-Tan, MD, PhD, and Stefano Trebeschi, PhD

Objectives: Response Evaluation Criteria in Solid Tumors (RECIST) is grounded on the assumption that target lesion selection is objective and representative of the change in total tumor burden (TTB) during therapy. A computer simulation model was designed to challenge this assumption, focusing on a particular aspect of subjectivity: target lesion selection.

Materials and Methods: Disagreement among readers and the disagreement between individual reader measurements and TTB were analyzed as a function of the total number of lesions, affected organs, and lesion growth.

Results: Disagreement rises when the number of lesions increases, when lesions are concentrated on a few organs, and when lesion growth borders the thresholds of progressive disease and partial response. There is an intrinsic methodological error in the estimation of TTB via RECIST 1.1, which depends on the number of lesions and their distributions. For example, for a fixed number of lesions at 5 and 15, distributed over a maximum of 4 organs, the error rates are observed to be 7.8% and 17.3%, respectively.

Conclusions: Our results demonstrate that RECIST can deliver an accurate estimate of TTB in localized disease, but fails in cases of distal metastases and multiple organ involvement. This is worsened by the "selection of the largest lesions," which introduces a bias that makes it hardly possible to perform an accurate estimate of the TTB. Including more (if not all) lesions in the quantitative analysis of tumor burden is desirable.

Key Words: RECIST, reader variability, simulation, cancer imaging

(*Invest Radiol* 2024;59: 465–471)

The Response Evaluation Criteria in Solid Tumors (RECIST) consists of a standardized methodology used in early and phase II clinical trials to evaluate tumors' response to therapy,[1,2] by defining end points surrogate for overall survival, namely, progression-free survival and overall response rate.[3,4] Buyse et al[5] have undertaken extensive research focusing on statistical frameworks for the validation of surrogate end points, encompassing their definition, validation, and the substantiating evidence on the applicability of progression-free survival and overall response rate as surrogate end points for overall survival.[5–7]

The most recent version of the criteria, RECIST 1.1,[8] states that a maximum of 5 measurable lesions, and no more than 2 per organ, should be selected as target lesions and measured at baseline. On follow-up, the same target lesions should be identified and remeasured. Response to therapy is then classified into 4 categories, primarily based on the percent change of the sum of the sizes of target lesions between baseline (or at nadir) and follow-ups. Nontarget lesions are also considered in case of their unambiguous progression.

Target lesion selection assumes that target lesions can be objectively and reproducibly identified and measured.[3] However, intervariability and intravariability when measuring tumor size and identification of new lesions are known factors contributing to inconsistency in the application of RECIST.[9–12]

Particularly, disagreement in the selection of target lesions is 1 of the leading causes of variability in RECIST.[13–16] According to the guidelines, the largest lesions in diameter, or the ones that "lend themselves to reproducible repeated measurements" should be chosen.[8] However, the interpretation of these guidelines is reader-dependent,[14] and different readers might end up selecting different target lesions, even if correctly applying RECIST criteria. In other words, even when excluding the factor of diameter-measuring error (eg, through medical segmentation software), RECIST still shows intrinsic variability if a limited number of target lesions has to be selected.

RECIST strives to make the comparison of clinical trial outcomes possible and reproducible.[3] However, this is grounded on the assumption that the target lesions are objectively identified and that this subset of the total tumor burden (TTB) of a patient is sufficient to adequately represent response to therapy.[14,15,17] Nevertheless, these assumptions may not hold in practice.

In this study, we aim to show that target lesion selection introduces large inconsistency in RECIST assessments and that its role as a surrogate of TTB is not sustained. Motivated by a previous study by Moskowitz et al,[18] where the effect on response assessment of the number of lesions measured was studied, we created a computer simulation model to investigate target lesion selection variability in RECIST across different patient characteristics, and its validity as a proxy for TTB. Compared with observational studies, where we are limited to the characteristics of the retrospective cohort (eg, number of patients, lesions per patient, average tumor growth, unavailability of TTB measurements), a simulation model yields advantages. It allows us to run a controlled experiment, in which we can generate any virtual cohort of patients with precise characteristics, and with which we can study the influence of each of them on the outcome: the RECIST assessment.

METHODS

Simulation Model

We created a computer simulation model, represented schematically in Figure 1. With this model, we generate cohorts of patients

Received for publication August 29, 2023; and accepted for publication, after revision, September 26, 2023.

From the Radiology Department (T.T.B., K.G.L., Z.B., T.D.L.N.-K., R.B.-T., S.T.), Biostatistics Unit (R.T.), and Thoracic Oncology (K.G.L.), Netherlands Cancer Institute, Amsterdam, the Netherlands; GROW School for Oncology and Reproduction, Maastricht University, Maastricht, the Netherlands (T.T.B., K.G.L., Z.B., R.B.-T., S.T.); Institute for Advanced Biomedical Technologies, Gabriele d'Annunzio University of Chieti-Pescara, Italy (A.D.P.); Department of Innovative Technologies in Medicine and Dentistry, Gabriele d'Annunzio University of Chieti-Pescara, Italy (A.D.P.); Institute of Diagnostic and Interventional Radiology, University Hospital of Zurich, Zurich, Switzerland (T.D.L.N.-K.); Institute of Radiology and Nuclear Medicine, Stadtspital Zürich, Zurich, Switzerland (T.D.L.N.-K.); and Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark (R.B.-T.).

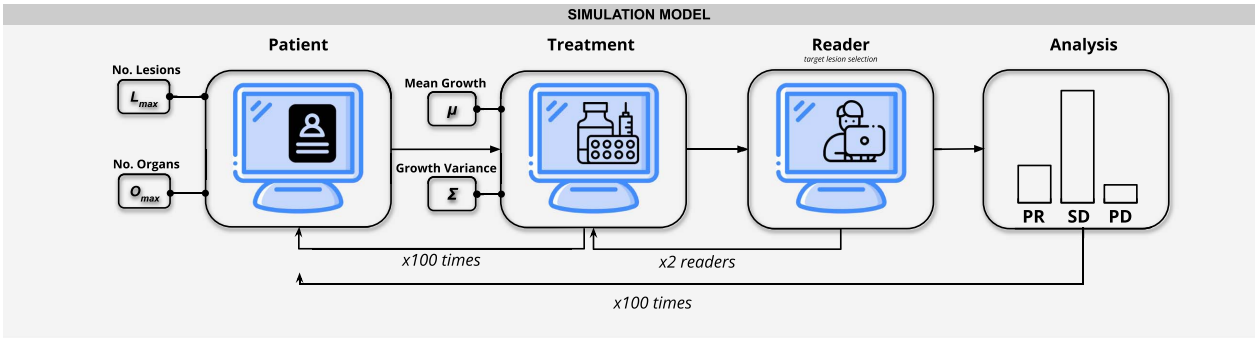
Conflicts of interest and sources of funding: none declared.

Posted history: An earlier version of this manuscript, with preliminary results, was previously posted to bioRxiv (https://www.biorxiv.org/content/10.1101/2022.04.14.488203v2).

Correspondence to: Stefano Trebeschi, PhD, GROW School for Oncology and Reproduction, Maastricht University, Universiteitssingel 40, Maastricht, Limburg, 6229 ER, the Netherlands. E-mail: stefano.trebeschi@maastrichtuniversity.nl.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).

ISSN: 0020-9996/24/5906–0465
DOI: 10.1097/RLI.0000000000001045



EXAMPLE OF A SIMULATED PATIENT AND READOUT

Input	
L_{max}	15
O_{max}	8
μ	-10
ε_O	11.6
ε_R	36.6

Patient				
Lesion	Organ	d_{BL}(mm)	Δ (%)	d_{FU}(mm)
1*	1	19	3.6	19.7
2	1	78	-6.4	73.0
3	1	52	1.3	52.7
4	2	80	-60.1	31.9
5	2	75	-11.3	66.5
6	2	83	-44.2	46.3

Reader 1 Target Lesions			
Lesion	Organ	d_{BL}(mm)	d_{FU}(mm)
2	1	78	73.0
3	1	52	52.7
4	2	80	31.9
6	2	83	46.3
Target Lesions' Growth		-30.4%	
RECIST Category		PR	

Reader 2 Target Lesions			
Lesion	Organ	d_{BL}(mm)	d_{FU}(mm)
2	1	78	73.0
3	1	52	52.7
5	2	75	66.5
6	2	83	46.3
Target Lesions' Growth		-17.2%	
RECIST Category		SD	

FIGURE 1. Top, schematic representation of the simulation model. Number of lesions (L) and number of organs (O) are, for each patient, a random number between [1, L_{max}] and [1, O_{max}], respectively. Each lesion's percent growth (Δ) is sampled from a truncated multivariate normal distribution with mean μ and covariance matrix Σ (a combination of ε_P, ε_O, and ε_R). Bottom, example of a simulated patient with L = 6 lesions distributed among O = 2 organs, and 2 virtual readers. Only large lesions are eligible for selection as target lesions, so lesion 1 (*) is excluded. The 2 readers disagree on the choice of target lesions, resulting in different RECIST categories. CR is unlikely to be attributed because lesion growth is sampled from a truncated multivariate normal distribution with nonsmall variance.

undergoing a virtual clinical trial. A cohort of patients is characterized by 4 parameters: maximum number of lesions (L_{max}), maximum number of organs involved (O_{max}), mean tumor percent growth (μ), and growth variance (Σ). To test the influence of each of the parameters on the RECIST assessment, we fix, in succession, all but 1 of the 4 input parameters to a default value, whereas the remaining variable changes within a certain range: L_{max} between 1 and 20, O_{max} between 1 and 10, μ between −100% and 200%, and μ between 10% and 100%. Σ is composed of 3 different variances, of which 2 change (ε_O and ε_R, see next section), with values ranging between 10^2 and 100^2 ‰, and one is kept at its default value (ε_P, see *Virtual Patient* and Supplement 1, http://links.lww.com/RLI/A860). We set a large range of possible values, which could include rare situations, such as only 1 candidate lesion per patient or, controversially, 20 lesions or 10 organs, to intentionally consider a wide variety of possible scenarios where RECIST could fail.

Virtual Patient

Each simulated patient can be described in 4 steps. First, the number of organs (O) and lesions (L) of each patient is drawn randomly between [1 to O_{max}] and [1 to L_{max}], respectively. The lesions are distributed among the organs with a multinomial distribution, where the probability of a lesion getting assigned in each organ is the same (if O > L, then some organs get assigned no lesions). Second, for each lesion, a random baseline tumor size (d_{BL}) between 10 and 100 mm is sampled from a random normal distribution. Third, the growth or shrinkage of each lesion (Δ) is simulated via a truncated multivariate normal distribution with mean μ and variance-covariance matrix Σ. Σ is composed of 3 different variance parameters to account for the growth correlation between lesions within patients (ε_P), within organs (ε_O), and the residual variance (ε_R). Fourth, based on the sampled percent growth, Δ, and the baseline diameter, d_{BL}, of each lesion, we calculate the follow-up diameter (d_{FU}). We only simulate these 2 time points of the virtual clinical trial: this will suffice to explore the effects of RECIST variability on the estimation of the TTB without overcomplicating the simulation unnecessarily. An example of a simulated patient can be seen in Figure 1.

The default values of μ, ε_P, ε_O, and ε_R were estimated from 3 real datasets: n = 61 patients with melanoma treated with immunotherapy, n = 44 urothelial cancer patients treated with immunotherapy, and n = 37 patients with non–small cell lung cancer treated with chemotherapy, already reported in previous work.[19,20] ε_P has no impact on target lesion selection, and thus, its impact on variability was not analyzed. For all datasets, the diameters of all lesions present at both baseline and follow-up were available. From that, the respective growth percent was estimated, on which linear mixed-effects model was fit. L_{max} and O_{max} were set to 3 different pairs of values (5, 2), (10, 4), and (15, 8), aiming to portray different stages of disease—low, medium, and high, respectively.

Virtual Reader

Each patient was evaluated by 2 independent computer-generated virtual readers. These virtual readers emulate the selection process of radiologists that select up to 5 target lesions, no more than 2 per organ, according to the RECIST criteria. To account for the RECIST guidelines suggesting that the largest lesions should be selected as target lesions, we restricted the choice of target lesions to a pool of lesions with size within 20% of the size of the largest lesion in each organ. When no lesions were under this condition, we added the second largest lesion to the pool, as well as all the lesions within 20% of its size (thus allowing at least 2 lesions to be chosen per organ). For each patient, we calculate the overall percent growth (or shrinkage) based on the sum of diameters of target lesions selected by each reader and assign the respective RECIST category. A percent growth above 20% (and minimum absolute increase of 5 mm) corresponds to progressive disease (PD), a percent decrease below 30% to partial response (PR), and otherwise, stable disease (SD).[8]

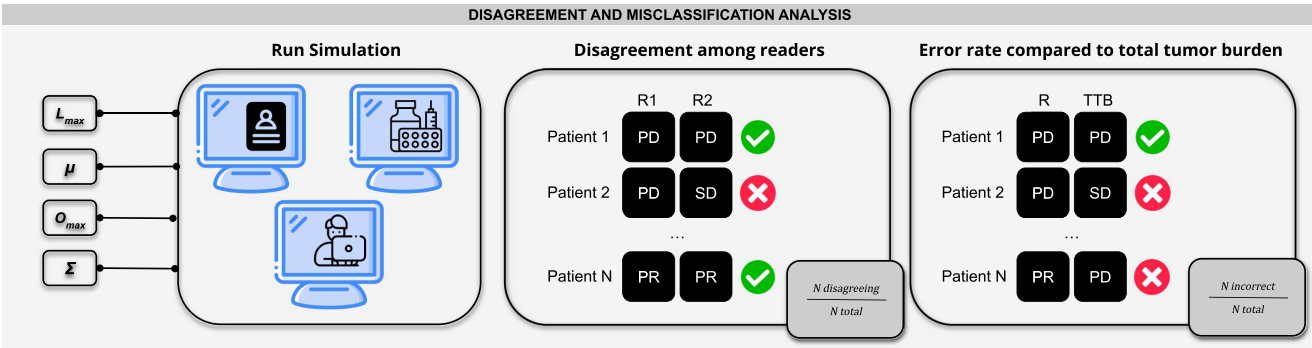


FIGURE 2. Metrics for analyzing RECIST variability in the simulated cohorts of patients. Disagreement corresponds to the percentage of patients in a cohort with inconsistent RECIST readings (ie, different response categories attributed by different readers). Disagreement is then analyzed as a function of the input parameters (L_{max}, O_{max}, μ). The error rate, computed per reader, corresponds to the percentage of incorrectly attributed response categories, taking TTB as reference. The error rate is analyzed as a function of L_{max}, and for different variations of RECIST, see *Results* section. R, Reader; TTB, total tumor burden.

Data Analysis

To study the relationship between readers' disagreement and patients' characteristics, we simply record the percentage of patients with inconsistent RECIST readings across the readers for a specific set of cohort characteristics (ie, model parameters). To minimize the effect of random noise on the analysis, each experiment is repeated and averaged over 100 runs. The disagreement levels for different combinations of patient cohort characteristics and treatment effects are recorded and analyzed.

To study the accuracy of RECIST as an estimator of the changes in the TTB, we turn the analysis into a classification problem. Given that, for each simulated patient, we know the true change in TTB (and therefore the true response class of PD, SD, or PR), we simply count the number of RECIST readings that resulted in an erroneous prediction of TTB-derived class of response. As in the analysis above, we repeat and average over 100 runs to minimize the effect of random noise, and study the trend of RECIST misclassification of TTB as a function of the (fixed) number of lesions (L_{max} = L). Figure 2 shows a schematic representation of the analysis. The code of the simulation model is publicly available on our GitHub repository.*

RESULTS

Examining Disagreement Between Readers

First, the disagreement between readers was investigated as a function of patient characteristics. Figure 3 depicts RECIST variability across different cohort characteristics. A linear relation between the maximum number of lesions per patient (L_{max}) and disagreement is evident, which behaves independently from the extent of the disease spread (Fig. 3). Even when the maximum number of lesions is less than 5, some disagreement is still evident, since only a maximum of 2 lesions per organ can be chosen.

There is a nonlinear relation between disagreement levels and the number of affected organs (O_{max}). When O_{max} varies from 1 to 10, there is high disagreement initially, because many lesions are concentrated on very few organs, and due to the imposition of a maximum of 2 lesions per organ, readers are forced to make different choices of target lesions. Disagreement then decreases because lesions spread out enough over multiple organs, limiting the number of choices, but increases again due to the limit of a maximum of 2 target lesions per organ, forcing a choice of target lesions and, implicitly, target organs. When L_{max} = 5 (green line in the middle plot of Figure 3), disagreement approaches zero when O_{max} increases, because all lesions can be selected if no organ contains more than 2 lesions (Fig. 3). When O_{max} increases beyond the number of lesions, some organs do not get assigned any lesions. Disagreement continues to decrease merely due to the chances of the 5 lesions being located in different organs increases (indicated by the dashed line). This scenario, where $O_{max} > L_{max}$, can also occur in other curves but to a negligible degree.

Disagreement among readers is higher when the average tumor growth (μ) borders the thresholds of PD and PR (+20% and −30%, respectively). When distancing from the thresholds, disagreement decreases because it becomes more likely for either PD or PR to be attributed. A nonlinear relation exists between average tumor growth and disagreement, which, looking across the different groups of disease stages, seems to be amplified by the stage or spread of the disease. The disagreement levels as a function of the input variances can be found in Supplement 2, http://links.lww.com/RLI/A860.

Evaluating RECIST in Estimating Alterations in Total Tumor Burden

The performance of RECIST to predict the true class of response, as defined by the TTB, as a function of the number of lesions, was then investigated. In Figure 4A, it can be observed that the error rate plateaus above 15% for all patients with 10+ lesions and reaches levels approximately 20% for patients with 20 lesions, for all levels of O_{max}. For comparison, the experiments were rerun with different variations of RECIST: RECIST 1.0, which allows the selection of up to 10 target lesions; and RECIST-random, where the readers are free to choose any target lesions, and are not limited by the largest ones (Fig. 4B). RECIST 1.0 reaches the lowest error rate, staying approximately 10%, even for patients with 20 lesions. Both RECIST-random 1.1 and RECIST-random 1.0 perform only slightly worse than the original counterparts, with RECIST-random 1.0 achieving a better estimation of the true response class than standard RECIST 1.1.

Finally, an additional setting, RECIST 1.1 with an adjudicator, was added: 2 readers independently perform the assessment, and if divergent, a third reader, that is, the adjudicator, is tasked to independently perform the conclusive assessment. As observed in Figure 4B, the error rate with an adjudicator is overall lower than that of RECIST 1.1, yet higher than RECIST 1.0.

Assessing the Impact of Nontarget Lesion Progression

Thus far, RECIST has been evaluated overlooking nontarget lesion behavior. As per guidelines, unequivocal nontarget lesion progression should prompt a PD classification. These guidelines, however, do not specify a minimum increase or require explicit radiological measurements. To address this in our model, it was assumed that a minimal growth, detectable by observation alone, exists. Different thresholds of "minimal identifiable growths" of 5%, 10%, 20%, and 50% were tested. Growth equal to or beyond these thresholds in any nontarget lesion will trigger the PD classification. The experiments from *Examining Disagreement Between Readers* and *Evaluating RECIST in Estimating*

*https://github.com/nki-radiology/recist-variability

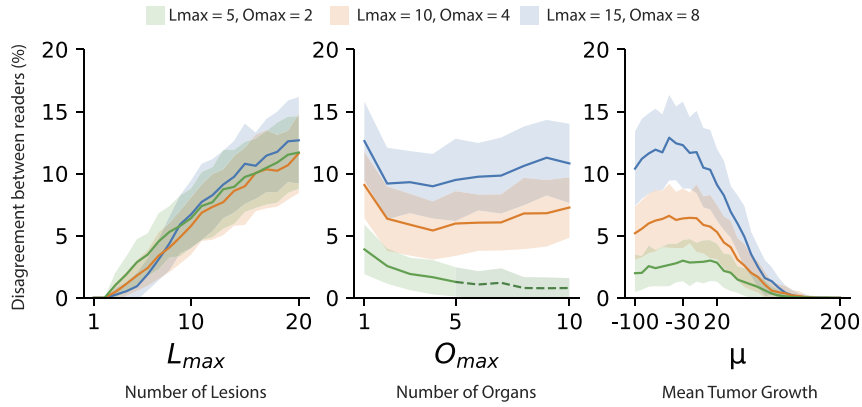


FIGURE 3. Mean disagreement levels as a function of the number of lesions, number of affected organs, mean lesion growth, with respective confidence intervals (±standard deviation). The different curves represent different combinations of the default values of L_{max} and O_{max}, respectively: green (5, 2); orange (10, 4); and blue (15, 8), representing different stages of disease. In plots A and B, only the default values of O_{max} and L_{max}, respectively, change. Dashed line: when $O_{max} > L_{max}$, disagreement decreases as the likelihood of the lesions becoming more spread out increases.

Alterations in Total Tumor Burden were repeated in this fashion, keeping the middle group ($O_{max} = 4$, $L_{max} = 10$) as default reference.

Figures 5A–C illustrates the disagreement as a function of the number of lesions, organ involvement, and tumor growth rate. In terms of tumor extent (ie, number of lesions and number of organs), the introduction of nontarget lesions PD decreases the level of disagreement between readers, compared with the case where the nontarget lesions are completely ignored. The opposite trend is observed with respect to tumor growth, where it is observed that the introduction of nontarget lesions PD increases the disagreement between readers that tumors are responding to therapy (growth $\mu < 0$). No trend is observed among the different thresholds with respect to different tumor extents, but it is evident with respect to tumor growth, where lower thresholds of identifiable growth result in higher disagreement between readers.

Figure 5D illustrates the error rate when estimating the TTB response class, as a function of the maximum number of lesions per patient. It is observed that the identification of progression within nontarget lesions leads to a rise in the cumulative error rate. The increase in error rate correlates proportionally with the reader's sensitivity to detecting alterations in lesion size, namely, the higher the sensitivity of the reader in detecting small changes, the higher the error rate. Only when the reader's sensitivity in detecting the progression of nontarget lesions diminishes to the point where the reader is able to only identify increases of at least 50% is a minor reduction in the error rate observed. This is in comparison to the scenario where nontarget lesions are entirely disregarded or overlooked.

DISCUSSION

This study aimed to investigate target lesion selection as a cause of variability in RECIST, in a controlled experiment, by means of simulated models. Previous studies have identified and studied target lesion selection as a source of variability in response classification in real cohorts of patients.[13–15] The advantage of a controlled experiment over an observational study is that it allows us to set the effects of the treatment and the characteristics of the patient cohort, and analyze the particular conditions under which RECIST is inconsistent, despite being correctly used. This helps us understand where the focus of future developments in tumor size–based response assessment should be—in polishing the already existing RECIST guidelines or in searching for better alternatives. Overall, our findings suggest a complex function linking patient characteristics with RECIST variability. In other words, although it is possible to explain the behavior of RECIST variability in relation to different characteristics of the patient cohorts, it seems impossible to draw a general rule that defines the expected level or behavior of RECIST variability in a simple fashion.

We observed that disagreement between readers increases linearly (and nearly independently from the number of organs) when the total number of lesions per patient increases, because the chances of different readers selecting the exact same target lesions decreases. RECIST recommends choosing target lesions based on size or on how reproducible their measurement is, which should help dilute differences between readers and lead to reduced variability. Selecting the largest lesions in diameter implicitly requires that lesions are sufficiently distinct in size such

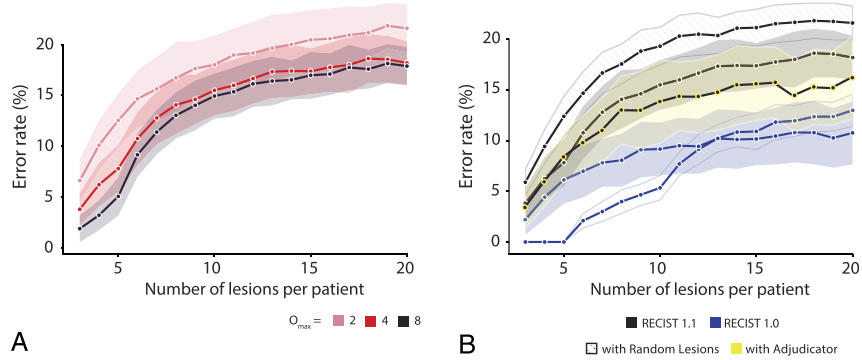


FIGURE 4. A, Error rate in classifying response with RECIST compared with the true tumor response, as defined by TTB, for different values of O_{max}. B, Error rate for $O_{max} = 4$, comparing variations of RECIST: 1.1, 1.0, with random lesions selected as target, and with adjudicator reader used when 2 readers disagree on the same case.

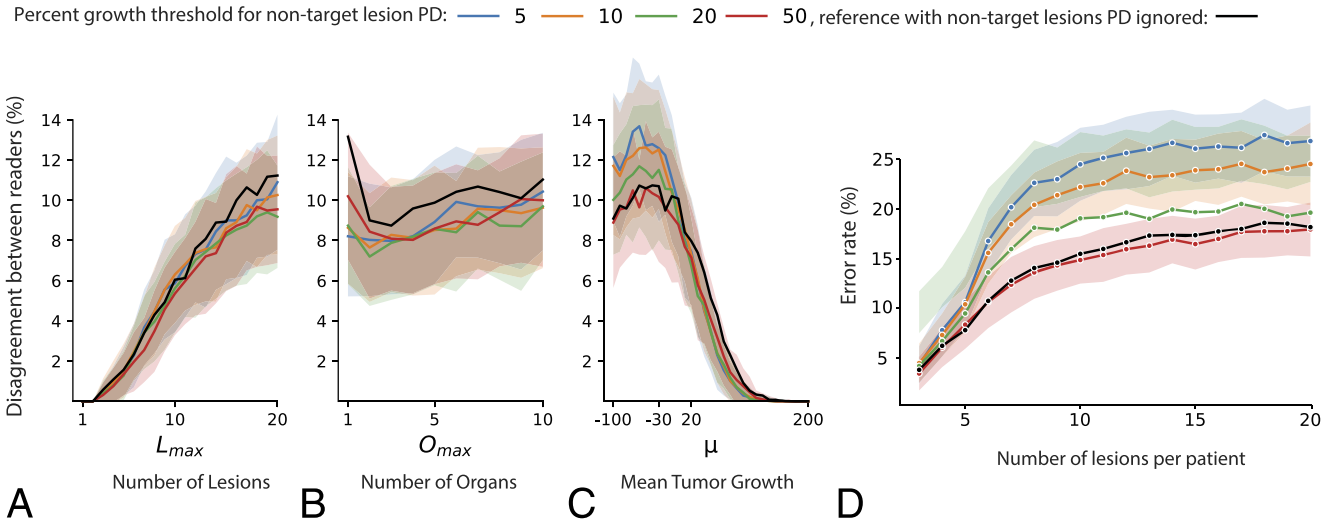


FIGURE 5. A–C, Mean disagreement levels as a function of the number of lesions, number of affected organs, mean lesion growth, with respective confidence intervals (±standard deviation). The different curves represent different thresholds for nontarget progression (5%, 10%, 20%, 50%) and no threshold for reference; default values of L_{max} and O_{max} are 10 and 4, respectively. D, Error rate for $O_{max} = 4$, comparing the different nontarget progression thresholds.

that the ones with the largest diameters are easily and uniquely identified, or that actual measurements of all the lesions are available such that the largest can be objectively chosen.[18] Even when same-lesion actual measurements are carried out, there can be critical interobserver and intraobserver measurement discrepancies.[9,13,21,22] Therefore, if the selection of the largest lesions has to be performed by visual inspection, it is expected that disagreements will aggravate further,[18] especially if the lesions have complex shapes. In our simulation, we allowed the virtual readers to select lesions from a pool of "large lesions," which we set to be all those within 20% of the size of the largest lesion, to account for the possibility of readers diverging in their discernment of what the largest lesions should be. Selecting the most reproducible lesions is also heavily subject to personal judgment. Readers need to decide if a lesion is "reproducible enough" or even of malignant nature[16,17] and if it should be chosen in favor of a larger lesion. Furthermore, selecting the largest or the lesions with the most well-defined boundaries for measurement may imply that the lesions selected are the ones responding to treatment in a similar way,[23] while largely overlooking the (nontarget) lesions that could be as or more determinant for treatment response assessment.[24] Despite the criteria guiding the choice of target lesions, this is one of the main contributors to RECIST's inconsistency.[13–15] As noted by Kuhl,[23] although additional recommendations would indeed help reproducibility, these might also just be masking the fundamental problems of RECIST and constraining readers from selecting the same lesions, even if these are not representative of the true tumor burden.

The substantial disagreement in response assessment by different readers suggests that 5 lesions alone are not sufficient to assure an objective response assessment. This is confirmed in our experiments by the relatively high error rate of RECIST in predicting the true response class (based on the TTB), in comparison to the substantially lower error rate reported for RECIST 1.0 (which allows double the number of target lesions in the analysis) and the relatively small difference in performance with an alternative, hypothetical RECIST (which allows the selection of random target lesions, instead of the largest ones). Seeing that the selection of random lesions (instead of the largest ones) did not degrade the performance as much as decreasing the number of target lesions (from 10 to 5) suggests that the inclusion of the largest lesions had a small effect. We found that incorporating an adjudicator led to a decrease in the error rate when estimating the TTB. The adjudicator is introduced specifically only for patients for which the initial readers disagree, that is, for which there are already relatively high chances of disagreement. By including the adjudicator, we identify a response class that garnered agreement from at least 2 readers. The reduction in error rate suggests that this response class is closer to the actual true response class, based on TTB, which highlights the importance of having an adjudicator as suggested by RECIST guidelines. It is important to add, however, that the adjudication still did not provide the same accuracy as in selecting more target lesions, observed in comparison to RECIST 1.0.

Including more, if not all, lesions in the quantitative analysis would be the best way to reduce variability and help capture the true TTB, likely leading to a more representative evaluation of response to therapy,[14,23] and possibly fewer patients required in clinical trials to prove the efficacy of the treatment. Because percent change is computed by taking into account the sum of the sizes of all target lesions, more importance is given to changes in larger lesions. This explains why we observe an increase in error when selecting random lesions compared with only large ones, which is aggravated if the number of target lesions is more limited.

In the simulation study of Moskowitz et al,[18] the impact of the number of target lesions measured (10, 5, 3, 2, or 1) on response assessment was investigated. The authors agreed that measuring a smaller number of lesions led to a larger percentage of misclassified patients. Nevertheless, it was concluded that measuring 5 lesions was the best compromise between a good enough proxy for TTB and the labor-intensive task of assessing many lesions, because assessing 10 lesions did not provide added benefit. In the study of Schwartz et al,[17] the authors defend that the number of lesions to be measured should be established based on the specific context of the study and intended comparisons with other studies.

In some therapies, clinical outcomes are associated with the site of metastasis.[25–27] As patients usually have more than 1 site of measurable disease, spreading the choice of lesions across the body (by imposing a maximum of 2 target lesions per organ) allows for the selection of a complete set of lesions with varying degrees of response to therapy. Nevertheless, as we observed, when many lesions are concentrated in a single or few organs, this imposition creates high disagreement between readers. Including all measurable lesions together with a subanalysis of the tumor burden per organ could be warranted.

Disagreement between readers varied nonlinearly with the average tumor growth but increased linearly with disease stage. It was aggravated when the mean growth of single lesions was centered around the small percentages of growth or shrinkage. The further away mean

lesion growth of the set of target lesions is from the cutoffs for PD (20%) and PR (−30%), the clearer the attribution of the response category is. However, when mean lesion growth is closer to these cutoffs, classifying response is more troublesome and completely dependent on the selected lesions. The coarse compartmentalization of response into 4 categories has been the target of critique. It has been proposed that a system that describes lesion growth as a continuous variable[18,23,28] would be more appropriate for assessing response to therapy.[29,30] Furthermore, although percent change allows us to compare patients with different levels of baseline tumor burden easily, one could also question the appropriateness of this metric. For example, when assessing the significance of a doubling in size of a lesion, it becomes essential to consider whether such a change holds equal clinical relevance for lesions with vastly different baseline sizes. A lesion with a very small diameter at baseline experiencing a doubling in size may not carry the same clinical implications as a larger lesion exhibiting a similar change. Consequently, the influence of time as a crucial factor cannot be overlooked. Estimating the rate of growth, therefore the factor of time between measurements, and incorporating it into the analysis become imperative to obtain a comprehensive and accurate understanding of the response assessment.

When the threshold for nontarget progression was set at a smaller value, the error rate in estimating the TTB increased. This results in a paradox wherein the identical lesion, exhibiting the same percent growth, would have different implications depending on its target/nontarget classification. If classified among the target lesions, it would have to be considered in conjunction with the remaining target lesions. However, if it was categorized as a nontarget lesion, it would independently constitute progression. In general, we can conclude that a minor progression in nontarget lesions should not automatically classify a patient as having PD, which aligns with the RECIST guidelines. However, the absence of a more quantitative assessment of nontarget disease, combined with possible differing opinions among readers on what constitutes unequivocal progression, can still lead to variations in the application of RECIST.

Our results must be interpreted in the light of the default values chosen for the input parameters, and the overall behavior of the simulation curves should be the focus point. We estimated the default values of mean lesion growth and variance from 3 real datasets of patients, as a way of approximating 2 possible scenarios of growth patterns between baseline and first follow-up. For example, if patient, organ, and residual default variances were very small, we would expect to see 2 disagreement peaks centered around the exact cutoff percentages for PD and PR. Because of the nature of the simulation, where lesion growth is sampled from a truncated multivariate normal distribution with nonsmall variance, we do not observe CR, and we still observe some disagreement when $\mu = -100\%$. Furthermore, in the case of malignant lymph nodes, it would be necessary to introduce additional complexity to the simulation model by specifying the exact organs where lesions are situated. This arises from the criterion that classifies lymph node lesions as nonmalignant if they measure less than 10 mm on their short axis. The current model does not explicitly specify the organs, operating on the premise that a complete response is observed when all lesions are no longer present, which does not hold true for lymph nodes. Regardless of whether lymph nodes are involved, classifying a patient with CR does not rely on the selection of target lesions.

In our simulation model, we did not explicitly take into account selection variability arising from the identification of what should be considered "reproducible" lesions. The readers were simulated such that they had to choose a maximum of 5 and no more than 2 lesions per organ, and only from the largest pool of lesions. However, there was no other explicit rule to force the readers to spread their choice of lesions as much as possible through the organs. Reader experience and possible appearance of new lesions were also not considered. This model also does not evaluate measurement variability, as the diameters of the lesions generated are assumed to be "accurately measured."

Volumetric measures have been described as a better discriminator of tumor size changes than RECIST's linear measurements,[31] related to the capacity of tumor volumetry better describing the size of irregular lesions[32,33] and to the reduced interobserver variability.[34–36] Although manual tumor segmentation is very labor-demanding, advances in artificial intelligence for automatic volumetric tumor segmentation of whole-body scans[37–39] might soon be integrated in clinical practice. Such a development could change the landscape of response assessment[2] by replacing target lesion selection with TTB estimation and rethinking the coarse threshold-based compartmentalization of response. Furthermore, it would also become feasible to analyze individual lesions, addressing the potential variability in how different lesions of the same patient respond to treatment.[40]

In conclusion, response assessment according to RECIST 1.1 experiences large variability due to the selection of target lesions, especially in a metastatic setting. Even if readers were to agree on the same set of target lesions, it cannot be guaranteed that these are representative of TTB and response to treatment. Recognizing that measuring more, if not all, measurable lesions in clinical practice is unrealistic, future efforts should focus on incorporating whole-body automatic segmentation models to achieve a more objective and accurate response assessment.

REFERENCES

1. Seymour L, Ivy SP, Sargent D, et al. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. *Clin Cancer Res*. 2010;16:1764–1769.
2. Litière S, Bogaerts J. Imaging endpoints for clinical trial use: a RECIST perspective. *J Immunother Cancer*. 2022;10:e005092.
3. Villaruz LC, Socinski MA. The clinical viewpoint: definitions, limitations of RECIST, practical considerations of measurement. *Clin Cancer Res*. 2013;19:2629–2636.
4. Mushti SL, Mulkey F, Sridhara R. Evaluation of overall response rate and progression-free survival as potential surrogate endpoints for overall survival in immunotherapy trials. *Clin Cancer Res*. 2018;24:2268–2275.
5. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*. 1998;54:1014–1029.
6. Michiels S, Saad ED, Buyse M. Progression-free survival as a surrogate for overall survival in clinical trials of targeted therapy in advanced solid tumors. *Drugs*. 2017;77:713–719.
7. Buyse M, Burzykowski T, Saad ED. The search for surrogate endpoints for immunotherapy trials. *Ann Transl Med*. 2018;6:231–231.
8. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45:228–247.
9. Skougaard K, McCullagh MJD, Nielsen D, et al. Observer variability in a phase II trial—assessing consistency in RECIST application. *Acta Oncol*. 2012;51:774–780.
10. Abramson RG, McGhee CR, Lakomkin N, et al. Pitfalls in RECIST data extraction for clinical trials: beyond the basics. *Acad Radiol*. 2015;22:779–786.
11. Sridhara R, Mandrekar SJ, Dodd LE. Missing data and measurement variability in assessing progression-free survival endpoint in randomized clinical trials. *Clin Cancer Res*. 2013;19:2613–2620.
12. Beaumont H, Evans TL, Klifa C, et al. Discrepancies of assessments in a RECIST 1.1 phase II clinical trial—association between adjudication rate and variability in images and tumors selection. *Cancer Imaging*. 2018;18:50.
13. Tovoli F, Renzulli M, Negrini G, et al. Inter-operator variability and source of errors in tumour response assessment for hepatocellular carcinoma treated with sorafenib. *Eur Radiol*. 2018;28:3611–3620.
14. Keil S, Barabasch A, Dirrichs T, et al. Target lesion selection: an important factor causing variability of response classification in the response evaluation criteria for solid tumors 1.1. *Invest Radiol*. 2014;49:509–517.
15. Kuhl CK, Alparslan Y, Schmoee J, et al. Validity of RECIST version 1.1 for response assessment in metastatic cancer: a prospective, multireader study. *Radiology*. 2019;290:349–356.
16. Iannessi A, Beaumont H, Liu Y, et al. RECIST 1.1 and lesion selection: how to deal with ambiguity at baseline? *Insights Imaging*. 2021;12:36.
17. Schwartz LH, Mazumdar M, Brown W, et al. Variability in response assessment in solid tumors: effect of number of lesions chosen for measurement. *Clin Cancer Res*. 2003;9:4318–4323.
18. Moskowitz CS, Jia X, Schwartz LH, et al. A simulation study to evaluate the impact of the number of lesions measured on response assessment. *Eur J Cancer*. 2009;45:300–310.

19. Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol*. 2019;30:998–1004.
20. Trebeschi S, Bodalal Z, van Dijk N, et al. Development of a prognostic AI-monitor for metastatic urothelial cancer patients receiving immunotherapy. *Front Oncol*. 2021;11:637804.
21. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976;38:388–394.
22. Yoon SH, Kim KW, Goo JM, et al. Observer variability in RECIST-based tumour burden measurements: a meta-analysis. *Eur J Cancer*. 2016;53:5–15.
23. Kuhl CK. RECIST needs revision: a wake-up call for radiologists. *Radiology*. 2019;292:110–111.
24. Coy HJ, Douek ML, Ruchalski K, et al. Components of radiologic progressive disease defined by RECIST 1.1 in patients with metastatic clear cell renal cell carcinoma. *Radiology*. 2019;292:103–109.
25. Bianchi JJ, Zhao X, Mays JC, et al. Not all cancers are created equal: tissue specificity in cancer genes and pathways. *Curr Opin Cell Biol*. 2020;63:135–143.
26. Lu LC, Hsu C, Shao YY, et al. Differential organ-specific tumor response to immune checkpoint inhibitors in hepatocellular carcinoma. *Liver Cancer*. 2019;8:480–490.
27. Schmid S, Diem S, Li Q, et al. Organ-specific response to nivolumab in patients with non–small cell lung cancer (NSCLC). *Cancer Immunol Immunother*. 2018;67:1825–1832.
28. Mercier F, Consalvo N, Frey N, et al. From waterfall plots to spaghetti plots in early oncology clinical development. *Pharm Stat*. 2019;18:526–532.
29. Jain RK, Lee JJ, Ng C, et al. Change in tumor size by RECIST correlates linearly with overall survival in phase I oncology studies. *J Clin Oncol*. 2012;30:2684–2690.
30. Wang M, Chen C, Jemielita T, et al. Are tumor size changes predictive of survival for checkpoint blockade based immunotherapy in metastatic melanoma? *J Immunother Cancer*. 2019;7:39.
31. Hayes SA, Pietanza MC, O'Driscoll D, et al. Comparison of CT volumetric measurement with RECIST response in patients with lung cancer. *Eur J Radiol*. 2016;85:524–533.
32. Fenerty KE, Folio LR, Patronas NJ, et al. Predicting clinical outcomes in chordoma patients receiving immunotherapy: a comparison between volumetric segmentation and RECIST. *BMC Cancer*. 2016;16:672.
33. Oubel E, Bonnard E, Sueoka-Aragane N, et al. Volume-based response evaluation with consensual lesion selection. *Acad Radiol*. 2015;22:217–225.
34. Rothe JH, Grieser C, Lehmkuhl L, et al. Size determination and response assessment of liver metastases with computed tomography—comparison of RECIST and volumetric algorithms. *Eur J Radiol*. 2013;82:1839–1839.
35. Wulff AM, Fabel M, Freitag-Wolf S, et al. Volumetric response classification in metastatic solid tumors on MSCT: initial results in a whole-body setting. *Eur J Radiol*. 2013;82:e567–e573.
36. Zimmermann M, Kuhl C, Engelke H, et al. Volumetric measurements of target lesions: does it improve inter-reader variability for oncological response assessment according to RECIST 1.1 guidelines compared to standard unidimensional measurements? *Pol J Radiol*. 2021;86:e594–e600.
37. Jemaa S, Fredrickson J, Carano RAD, et al. Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks. *J Digit Imaging*. 2020;33:888–894. doi:10.1007/s10278-020-00341-1.
38. Tang Y, Cai J, Yan K, et al. Weakly-supervised universal lesion segmentation with regional level set loss. *arXiv:210501218 [cs, eess]*. Available at: http://arxiv.org/abs/2105.01218. Published online May 3, 2021. Accessed May 6, 2021
39. He J, Zhang Y, Chung M, et al. Whole-body tumor segmentation from PET/CT images using a two-stage cascaded neural network with camouflaged object detection mechanisms. *Med Phys*. 2023;mp.16438. doi:10.1002/mp.16438.
40. Kumar R, Qi T, Cao Y, et al. Incorporating lesion-to-lesion heterogeneity into early oncology decision making. *Front Immunol*. 2023;14:1173546.