How does target lesion selection affect RECIST? A computer simulation study

Teresa T. Bucho[1,2], Renaud Tissier[3], Zuhir Bodalal[1,2], Thi Dan Linh Nguyen-Kim[4], Regina Beets-Tan[1,2,5], Stefano Trebeschi[1]

[1] Radiology Department and [3] Biostatistics Unit, Netherlands Cancer Institute, Amsterdam, The Netherlands; [2] GROW School for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands; [4] Institute of Diagnostic and Interventional Radiology, University Hospital of Zurich, Zurich, Switzerland; [5] Faculty of Health Sciences, University of Southern Denmark, Denmark

Abstract

RECIST is grounded on the assumption that target lesion selection is objective and representative of the total tumor burden's response to therapy. A computer simulation model was designed to study target lesion selection's impact on response assessment. Readers' disagreement as a function of the total number of lesions, affected organs, and lesion growth was analyzed. Disagreement aggravates when the number of lesions increases, when lesions are concentrated on few organs, and when lesion growth borders the thresholds of progressive disease and partial response. In a metastatic setting, RECIST displays a non-linear, unpredictable behavior, not reliable for response assessment due to target lesion selection variability. Including more (if not all) lesions in the quantitative analysis of tumor burden is desirable.

Keywords: RECIST; Reader Variability; Simulation; Cancer Imaging

1 INTRODUCTION

The Response Evaluation Criteria in Solid Tumors (RECIST) consists of a standardized methodology used in clinical trials to evaluate tumors' response to therapy (Seymour et al. 2010), by defining endpoints surrogate for overall survival, namely progression-free survival and overall response rate (Villaruz and Socinski 2013; Mushti, Mulkey, and Sridhara 2018). The most recent version of the criteria, RECIST 1.1 (Eisenhauer et al. 2009), states that a maximum of five measurable lesions, and no more than two per organ, should be selected as target lesions and measured at baseline. On follow-up, the same target lesions should be identified and re-measured. Response to therapy is then classified into four categories, primarily based on the percent change of the sum of the sizes of target lesions between baseline (or at nadir) and follow-ups.

RECIST strives to make the comparison of clinical trials outcomes possible and reproducible (Villaruz and Socinski 2013). However, it is grounded on the assumptions that the target lesions can be objectively and reproducibly identified and measured (Villaruz and Socinski 2013), and that these lesions alone are sufficient to adequately represent the total tumor burden of a patient (Schwartz et al. 2003). However, these assumptions may not hold in practice. Inter- and intra-variability when measuring tumor size, and identification of new lesions are among the factors contributing to inconsistency in the application of RECIST (Skougaard et al. 2012; Abramson et al. 2015; Sridhara, Mandrekar, and Dodd 2013; Beaumont et al. 2018).

Particularly, disagreement in the selection of target lesions is one of the leading causes of

variability in RECIST (Tovoli et al. 2018; Keil et al. 2014; Kuhl et al. 2019; Iannessi et al. 2021). According to the guidelines, the largest lesions in diameter, or the ones that "lend themselves to reproducible repeated measurements" should be chosen (Eisenhauer et al. 2009). However, the interpretation of these guidelines is reader-dependent (Keil et al. 2014), 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 (e.g. through medical segmentation software), RECIST still shows intrinsic variability if a limited number of target lesions has to be selected.

In this study, we create a computer simulation model to investigate target lesion selection variability in RECIST across, and in relation to, different patient characteristics, and its validity as a surrogate for total tumor burden (TTB). Compared to an observational study, where we are limited by the characteristics of the retrospective cohort (e.g. number of patients, lesions per patient, average tumor growth) a simulation model allows to run a controlled experiment, in which we can generate any virtual cohort of patients, which precisely displays a set of characteristics, and which we can use to study the interaction and influence of each of these characteristics on the outcome: the RECIST assessment in our case.

2 METHODS

2.1 Simulation Model

We created a computer simulation model, represented schematically in Figure 1. With this model, we generate cohorts of patients undergoing virtual clinical trials. A cohort of patients is characterized by the six variables with which the model is parametrized: maximum number of lesions (L_{max}), maximum number of organs involved (O_{max}), mean tumor percent growth (μ), and growth variances ($\mathcal{E}_P$, $\mathcal{E}_O$ and $\mathcal{E}_R$). To test the influence of each of the parameters of the simulation model on the RECIST assessment, we fix, in succession, all but one of the six input parameters to a default value, while the remaining variable is made to change within a certain range: L_{max} between 1 and 20, O_{max} between 1 and 10, μ between −100% and 300%; and μ, $\mathcal{E}_P$, $\mathcal{E}_O$ and $\mathcal{E}_R$ each between 10^2 and 100^2 ‰. The default values of the μ, $\mathcal{E}_P$, $\mathcal{E}_O$ and $\mathcal{E}_R$, (12.8%, 44.6^2 , 11.6^2 , 36.6^2, respectively) were estimated from a real dataset of patients with melanoma and urothelial cancer (Trebeschi et al. 2019; Trebeschi et al. 2021), treated with immunotherapy at the NKI-AVL hospital, by fitting a linear mixed-effects model, whereas L_{max} and O_{max} were set to 3 different pairs of values (5, 2), (10, 4) and (15, 8), respectively — low, medium and high disease stage, respectively.

2.2 Virtual Patient

Each simulated patient can be described in 4 steps: 1) for each patient, the number of organs (O) and lesions (L) is drawn randomly from a vectors of possible values, i.e. from 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 each organ is the same (p=1/O; if O > L then some organs get assigned no lesions); 2) for each lesion, a random baseline tumor size (d_{BL}) between 10 and 100 mm is sampled from a random normal distribution; 3) 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 the 3 different input variance parameters ($\mathcal{E}_P$, $\mathcal{E}_O$ and $\mathcal{E}_R$) in order to take into account the correlation existing between lesions within patients ($\mathcal{E}_P$) and within organs ($\mathcal{E}_O$), and a residual

variance ($\mathcal{E}_R$); 4) 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 two time points of the virtual clinical trial. The variability of RECIST should be evident when considering only these two time points, so the model complexity was not extended with more follow-ups. An example of a simulated patient can be seen in Figure 1.

2.3 Virtual Reader

Each patient was evaluated by 100 independent virtual readers. A virtual reader represents a radiologist that selects 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 “large lesions”. This consists of all lesions with a 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 (between baseline and first follow-up) of the set 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), below -30% to partial response (PR), equal to 100% to complete response (CR) and otherwise, stable disease (SD).

If a patient received at least one different RECIST classification from one of the virtual readers, we classify the patient outcome as disagreeing. The disagreement level is dictated by the percentage of disagreeing patients in the virtual cohort, for each combination of parameters. We record the disagreement level for different combinations of model parameters, describing a specific set of cohort characteristics and treatment effect. For reproducibility purposes, each combination of the parameters is run 100 times.

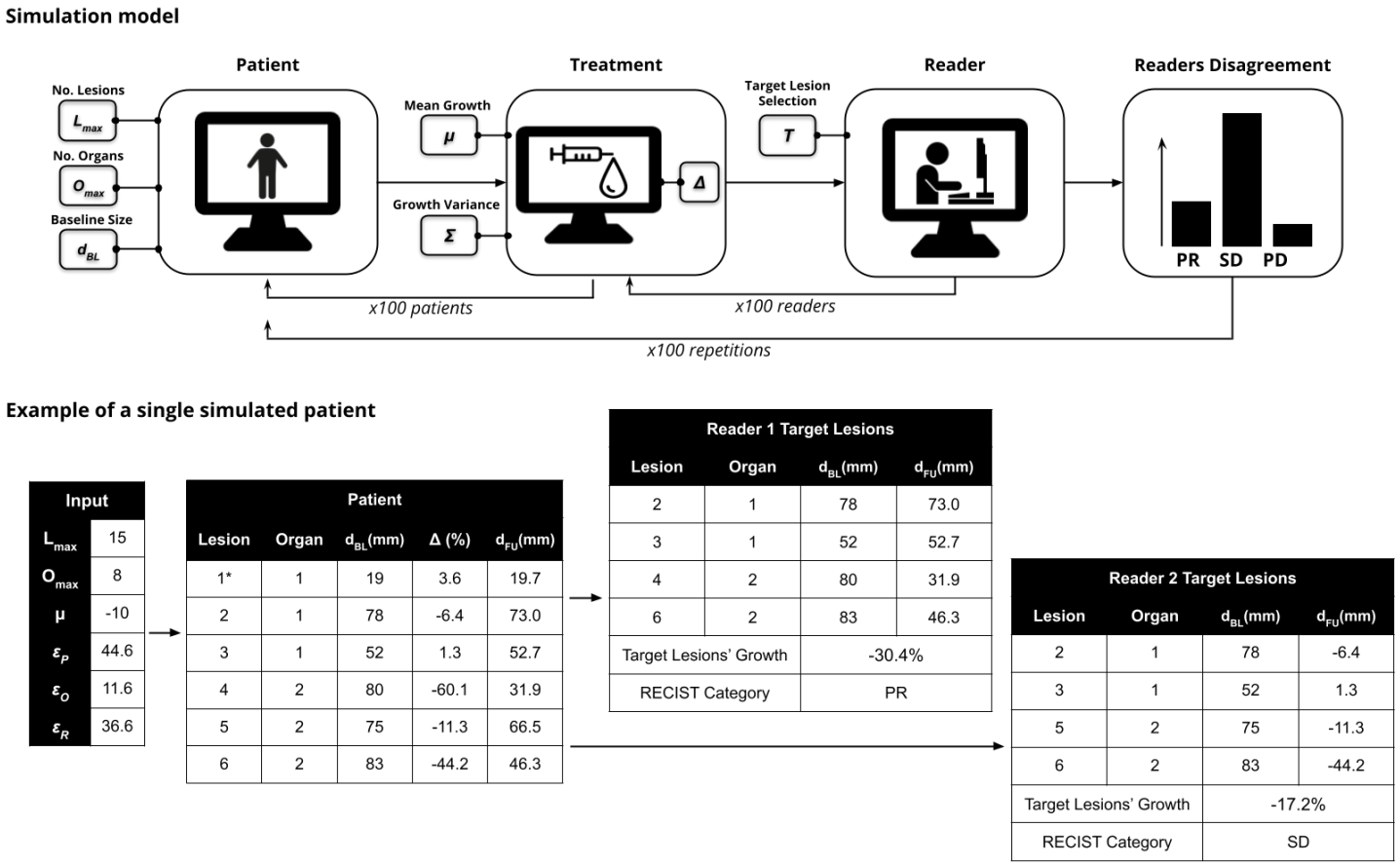

Figure 1: Top: schematic representation of the simulation model. L and O are, for each patient, a random number between L_{max} and O_{max}, respectively, while each lesion's percent growth (Δ) is sampled from a truncated multivariate normal distribution with mean μ and covariance matrix Σ (a combination of $\mathcal{E}_P$, $\mathcal{E}_O$ and $\mathcal{E}_R$). Bottom: example of a simulated patient with L=6 lesions distributed among O=2 organs, and two virtual readers. Only large lesions are eligible for selection as target lesions, so lesion 1(*) is excluded. The two readers disagree on the choice of target lesions, resulting in different RECIST categories.

2.4 Total tumor burden (TTB) estimation and number of readers

A second analysis was carried out, aiming at determining how accurate readers' true TTB estimation is and the required number of readers to obtain a confident estimation. We hypothesize that since different readers expectedly select different target lesions, a combination of the readings of multiple readers might lead to better estimations of true TTB. First, we simulate multiple cohorts of 100 patients, each with a fixed (L_{max} = L; default O_{max} = 8) number of lesions (from 3 to 20), and 100 readers, each selecting target lesions according to RECIST criteria and computing a target lesions percent change between baseline and follow-up. We compute the 'true' TTB percent change (i.e. considering all lesions at follow-up and baseline), ΔTTB%, and we verify if this is correctly estimated by the readers. From the standard deviation and mean of the readings for each patient, we compute its 95% confidence interval (CI) and determine for how many patients was ΔTTB% contained inside the interval, as a function of the number of lesions, a quantity we refer to as "correct TTB estimations". Next, based on the readings' mean standard deviation per number of lesions, we compute how many readers would be required to get 95% CIs with widths of 10 and 20%. The smaller the width of the interval, the smaller the disagreement between readers. Figure 2

shows a schematic representation of the analysis.

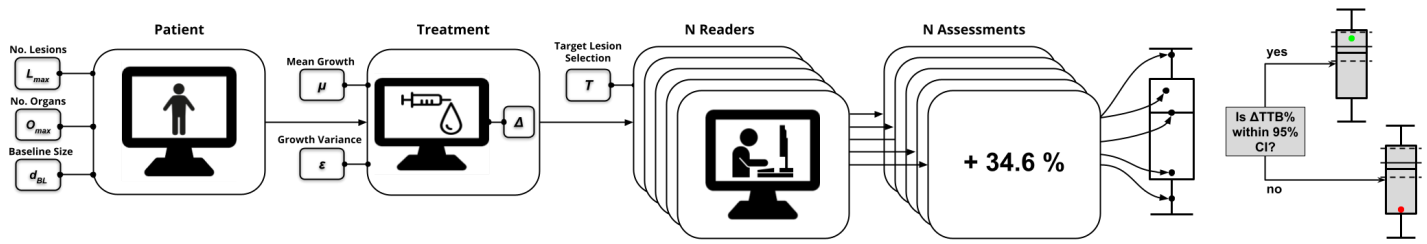

Figure 2: Schematic representation of the analysis performed to estimate whether the ΔTTB% would fall into the 95% confidence intervals of RECIST assessment, indicated by the dashed lines. For this analysis, patients have a fixed number of lesions (L_{max} = L).

The code of the simulation model is publicly available on the department repository[1].

3 RESULTS

3.1 Simulation

Figure 3 depicts RECIST variability across different cohort characteristics. We can observe a linear relation between the maximum number of lesions per patient (L_{max}) and disagreement, which reaches the same level independently from the disease stage. Even when the maximum number of lesions is below 5, we still observe some disagreement, 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, we observe high disagreement initially, since 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 due to lesions becoming more spread out, but increases again due to the maximum of 2 target lesions per organ imposition forcing a choice of target organs. In the low disease stage group (L_{max} = 5), disagreement plateaus due to the limited number of lesions.

When the average tumor growth (μ) borders the thresholds of PD and PR (+20% and -30% respectively), disagreement is higher, and decreases when μ increases, since only PD can be attributed. A nonlinear relation exists between average tumor growth and disagreement which, looking across the different groups of disease stages, seems to be linearly amplified by the stage or spread of the disease.

The impact of the variance parameters in the simulation model varies greatly. When the variance between patients ($\mathcal{E}_P$) increases the disagreement between the virtual readers diminishes as most of the variations between tumor growth is due to between patient variation and not within patients. The disagreement between readers seems robust to variance between-organs ($\mathcal{E}_O$) and only decreases for large values, which could again be due to the use of a truncated multivariate distribution. Finally, when the residual variance ($\mathcal{E}_R$) increases we see that the disagreement

[1] https://github.com/nki-radiology/recist-variability

between readers increases. This is expected as the residual variance represents variations between lesions independent of the fact that these lesions are from the same patient or the same organs. This means that with a large residual variance, lesions within the same organ of the same patient will tend to respond completely differently to the treatment.

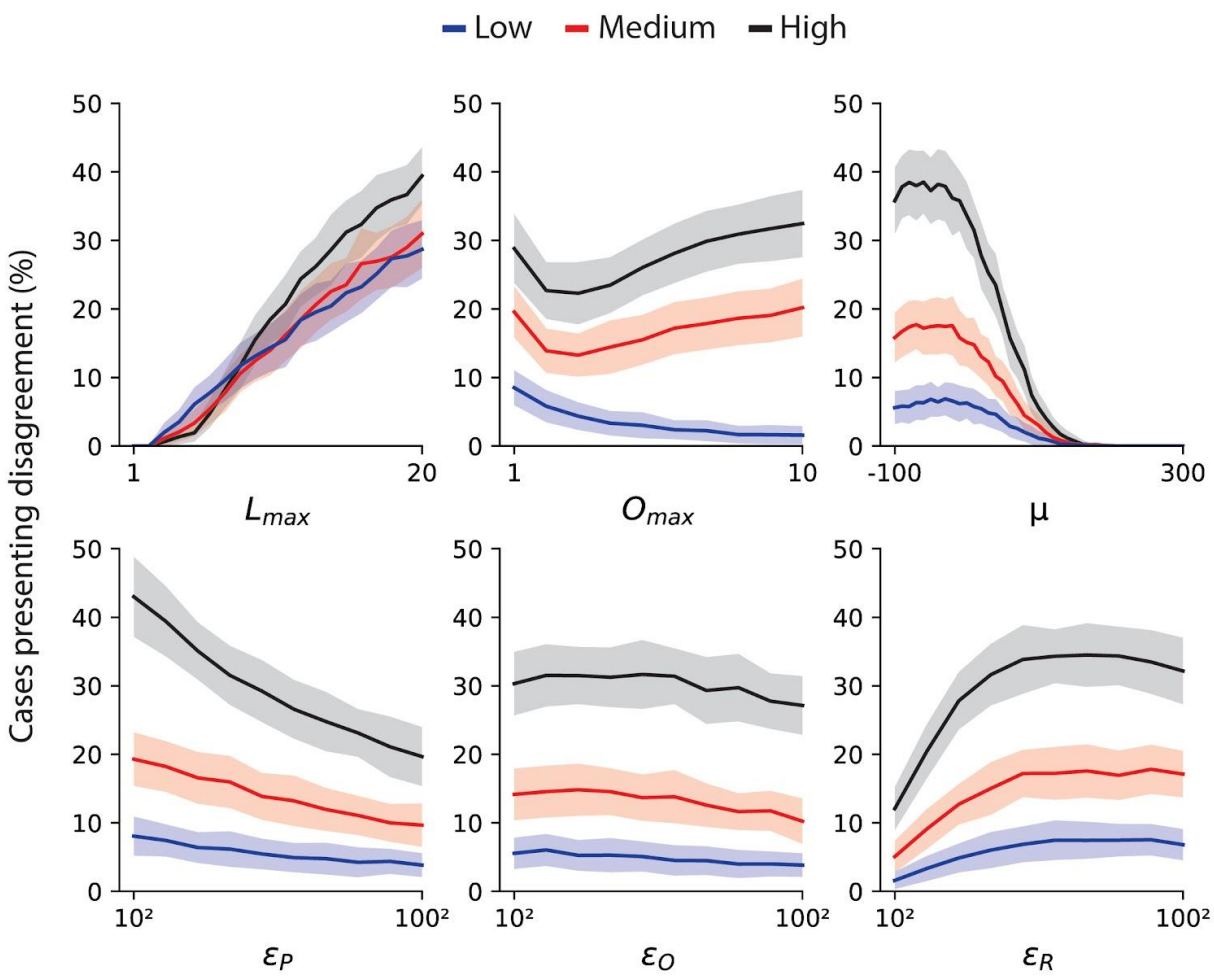

Figure 3: Mean disagreement levels as a function of the number of lesions, number of affected organs, mean lesion growth and within patient, within organ and residual growth variances and confidence interval (± standard deviation). The different colors represent different combinations of the default values of L_{max} and O_{max}, respectively: blue: (5, 2); red: (10, 4); and black (15, 8), representing different stages of disease (low, medium and high, respectively).

3.2 RECIST as surrogate for TTB

Figure 4 depicts the second analysis, concerning the relation between RECIST readings and percentage change in TTB (ΔTTB%), as a function of the number of lesions. A “correct” ΔTTB% estimate is defined when the ΔTTB% falls within the 95% CI of the average RECIST assessments obtained by the different readers. This also provides us with an idea of how much response assessment can be trusted. We observe that the higher the number of lesions (Figure 4A), the lower the percentage of correct estimations. Interestingly, when readers are not restricted to a pool of only “large lesions” to select target lesions from (i.e. RECIST with random lesions), we observe

higher percentages of correct estimates. The number of readers necessary to have a 95% CI with widths of 10% and 20% to estimate the ΔTTB% increases with the number of lesions (figure 4B). If the pool of lesions is restricted to only large lesions, then fewer readers are required to obtain the same CI width. This can be explained by the limited number of tumors to choose from, i.e., by reducing the number of tumors we reduce the variability existing between readers. Due to the smaller variation around the average TTB estimated by the readers, fewer readers are necessary to obtain the same 95% confidence interval widths.

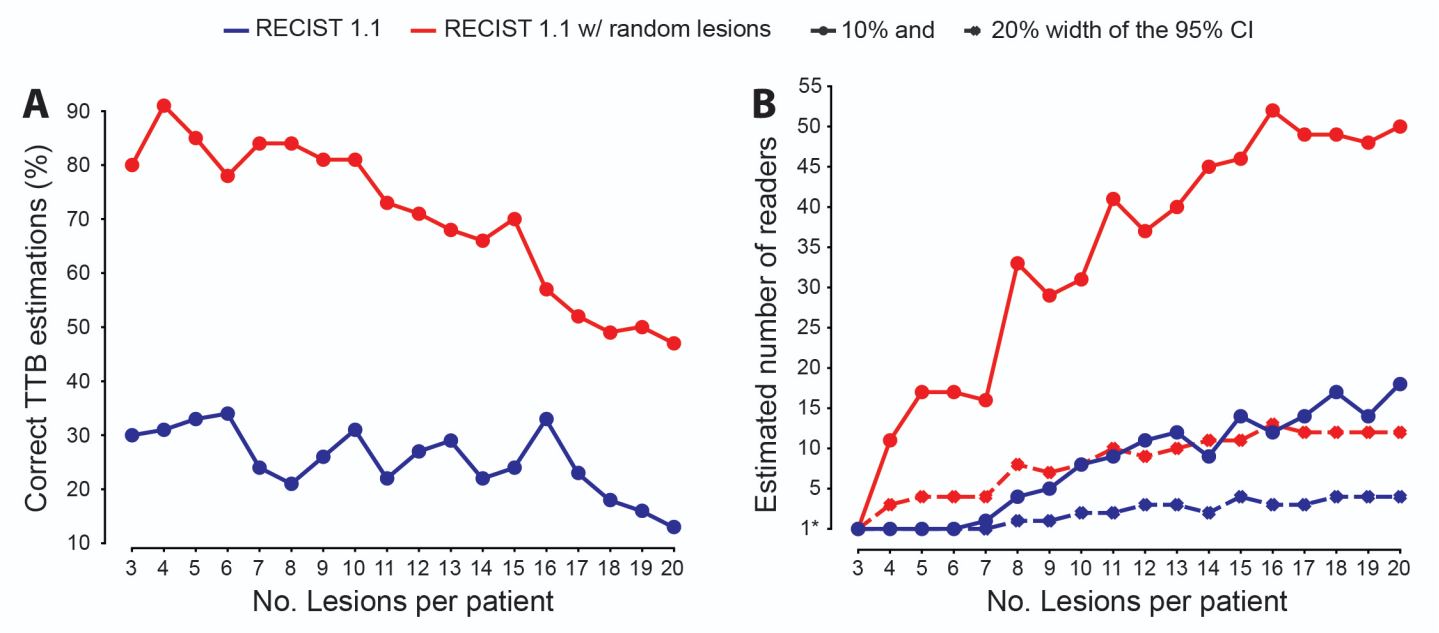

Figure 4: Percentage of correct TTB estimations (A) and (B) estimated number of readers required to obtain a 95% CI with 10% or 20% width, as a function of the number of lesions. The blue lines correspond to RECIST 1.1, where readers should select the largest lesions, and the red lines correspond to a scenario where readers can randomly choose lesions, independently of size. 1*: due to small standard deviations when the number of lesions is small, the estimated sample size is 0. Therefore only 1 reader would be required.

4 DISCUSSION

The aim of this study was to investigate target lesion selection as a source of variability in RECIST, in a controlled experiment, by means of simulated models. 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 employed. This helps us understand where the focus of future developments in tumor size-based response assessment should be: if in polishing the already existing RECIST guidelines or in abandoning them entirely for better alternatives. Overall, our findings suggest a complex function linking patient characteristics with RECIST variability. In other words, while it is possible to explain the behavior of RECIST variability in relation to different characteristics of the patient cohorts, it is not possible to draw a general rule that defines the expected level or behavior of RECIST variability in a simple fashion.

We found that when the total number of lesions per patient increases, so does disagreement between readers, linearly, since the chances of different readers selecting the exact same target lesions decreases, with no difference among cohort characteristics. 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 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 (Moskowitz et al. 2009). In our simulation, we allowed readers to select lesions from a pool of “large lesions”, 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 discernment. Readers need to decide if a lesion is “reproducible enough” or even of malignant nature (Schwartz et al. 2003; Iannessi et al. 2021) and if it should be chosen in detriment 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 (Kuhl 2019), while overlooking the (non-target) lesions that could be as or more determinant for treatment response assessment (Coy et al. 2019). Despite the criteria guiding the choice of target lesions, this is one of the main contributors to RECIST’s inconsistency (Tovoli et al. 2018; Keil et al. 2014; Kuhl et al. 2019). As noted in Kuhl 2019, while additional recommendations would indeed help reproducibility, these might also just be masking the fundamental problems of RECIST and constraining readers to select 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 is not a sufficient number to assure an objective response assessment. Including more, if not all, lesions in the quantitative analysis reduced variability and would help capture the true TTB, likely leading to a more representative evaluation of response to therapy (Keil et al. 2014; Kuhl 2019). In the simulation study in Moskowitz et al. 2009, 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 overview of TTB and the labor intensive task of assessing many lesions. In Schwartz et al. 2003, 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 (Bianchi et al. 2020; Lu et al. 2019; Schmid et al. 2018). As patients usually have more than one site of measurable disease, spreading the choice of lesions across the body (by imposing a maximum of 2 target lesions per organ) allows the selection of a complete set of lesions with varied 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 non-linearly 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 progressive disease (20%) and partial response (-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 four categories has been the target of critique. It has been proposed that a system that describes lesion growth as a continuous variable (Moskowitz et al. 2009; Kuhl 2019; Mercier et al. 2019), would be more appropriate for assessing response to therapy (Jain et al. 2012; Wang et al. 2019). Furthermore, while percent change allows us to easily compare patients with different levels of baseline tumor burden, one could also question the appropriateness of this metric. For example, should the doubling in size of lesion with a very small diameter at baseline be considered as relevant for response assessment as the doubling in size of a lesion with a much larger baseline size? Our results agree with those of studies involving real cohorts of patients and multiple readers (Keil et al. 2014; Tovoli et al. 2018; Kuhl et al. 2019), corresponding to certain combinations of the input parameters of our simulation model.

Nevertheless, 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. If patient, organ and residual default variances were very small, we would expect to see two disagreement peaks centered around the exact cutoff percentages for PD and PR. Due to the nature of the simulation, where lesion growth is sampled from a truncated multivariate normal distribution with non-small variance, we do not observe CR, and we still observe some disagreement when μ = -100%. In our study, we did not take into account selection variability due to disagreements arising from identifying “reproducible” lesions.

Volumetric measures have been described as a better discriminator of tumor size changes than RECIST’s linear measurements, related to the capacity of tumor volumetry better describing the size of irregular lesions (Fenerty et al. 2016; Oubel et al. 2015) and to the reduced interobserver variability (Rothe et al. 2013; Wulff et al. 2013; Zimmermann et al. 2021). While manual tumor segmentation is still very labor demanding, advances in artificial intelligence for automatic volumetric tumor segmentation of whole-body scans (Jemaa et al. 2020; Tang et al. 2021) might soon be a solution in clinical practice. Such a development could change the landscape of response assessment by replacing target lesion selection with TTB estimation and rethinking the coarse threshold-based compartmentalization of response.

We hypothesized that combining the RECIST assessments of multiple independent readers selecting different target lesions could lead to a more representative estimation of TTB. However, as we observed in our second analysis, the number of patients whose TTB was correctly estimated by multiple readers (i.e. whose TTB was contained in the readings 95% CI) decreases with the number of lesions, when readers are allowed to randomly select target lesions and even more so when small lesions are excluded, since not all lesions (that constitute the TTB) can be selected. To obtain narrow CIs for the estimation of TTB an unrealistic number of readers would be required, especially when the number of lesions is high. When excluding small lesions, the number of readers required would be lower (since variance between readings decreases), however it would likely not be representative of TTB. Adding more readers narrows down their estimation of TTB, but it is not assured that this estimation is representative.

In conclusion, RECIST’s response assessment suffers from 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. Including more, if not all, measurable lesions in the quantitative assessment is desirable, possibly aided by automatic segmentation models.

REFERENCES

Abramson, Richard G., Carrie R. McGhee, Nikita Lakomkin, and Carlos L. Arteaga. 2015. “Pitfalls in RECIST Data Extraction for Clinical Trials: Beyond the Basics.” *Academic Radiology* 22 (6): 779–86.

Beaumont, Hubert, Tracey L. Evans, Catherine Klifa, Ali Guermazi, Sae Rom Hong, Mustapha Chadjaa, and Zsuzsanna Monostori. 2018. “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*. https://doi.org/10.1186/s40644-018-0186-0.

Bianchi, Joy J., Xin Zhao, Joseph C. Mays, and Teresa Davoli. 2020. “Not All Cancers Are Created Equal: Tissue Specificity in Cancer Genes and Pathways.” *Current Opinion in Cell Biology* 63 (April): 135–43.

Coy, Heidi J., Michael L. Douek, Kathleen Ruchalski, Hyun J. Kim, Antonio Gutierrez, Maitrya Patel, Victor Sai, et al. 2019. “Components of Radiologic Progressive Disease Defined by RECIST 1.1 in Patients with Metastatic Clear Cell Renal Cell Carcinoma.” *Radiology* 292 (1): 103–9.

Eisenhauer, E. A., P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, et al. 2009. “New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (version 1.1).” *European Journal of Cancer* 45 (2): 228–47.

Fenerty, Kathleen E., Les R. Folio, Nicholas J. Patronas, Jennifer L. Marté, James L. Gulley, and Christopher R. Heery. 2016. “Predicting Clinical Outcomes in Chordoma Patients Receiving Immunotherapy: A Comparison between Volumetric Segmentation and RECIST.” *BMC Cancer* 16 (1): 672.

Iannessi, Antoine, Hubert Beaumont, Yan Liu, and Anne-Sophie Bertrand. 2021. “RECIST 1.1 and Lesion Selection: How to Deal with Ambiguity at Baseline?” *Insights into Imaging*. https://doi.org/10.1186/s13244-021-00976-w.

Jain, Rajul K., J. Jack Lee, Chaan Ng, David Hong, Jing Gong, Aung Naing, Jennifer Wheler, and Razelle Kurzrock. 2012. “Change in Tumor Size by RECIST Correlates Linearly with Overall Survival in Phase I Oncology Studies.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 30 (21): 2684–90.

Jemaa, Skander, Jill Fredrickson, Richard A. D. Carano, Tina Nielsen, Alex de Crespigny, and Thomas Bengtsson. 2020. “Tumor Segmentation and Feature Extraction from Whole-Body FDG-PET/CT Using Cascaded 2D and 3D Convolutional Neural Networks.” *Journal of Digital Imaging* 33 (4): 888–94.

Keil, Sebastian, Alexandra Barabasch, Timm Dirrichs, Philipp Bruners, Nienke Lynn Hansen, Heribert B. Bieling, Tim H. Brümmendorf, and Christiane K. Kuhl. 2014. “Target Lesion Selection: An Important Factor Causing Variability of Response Classification in the Response Evaluation Criteria for Solid Tumors 1.1.” *Investigative Radiology* 49 (8): 509–17.

Kuhl, Christiane K. 2019. “RECIST Needs Revision: A Wake-up Call for Radiologists.” *Radiology*.

Kuhl, Christiane K., Yunus Alparslan, Jonas Schmoee, Bruno Sequeira, Annika Keulers, Tim H. Brümmendorf, and Sebastian Keil. 2019. “Validity of RECIST Version 1.1 for Response Assessment in Metastatic Cancer: A Prospective, Multireader Study.” *Radiology* 290 (2): 349–56.

Lu, Li-Chun, Chiun Hsu, Yu-Yun Shao, Yee Chao, Chia-Jui Yen, I-Lun Shih, Yi-Ping Hung, et al. 2019. “Differential Organ-Specific Tumor Response to Immune Checkpoint Inhibitors in Hepatocellular Carcinoma.” *Liver Cancer* 8 (6): 480–90.

Mercier, Francois, Nicola Consalvo, Nicolas Frey, Alex Phipps, and Benjamin Ribba. 2019. “From Waterfall Plots to Spaghetti Plots in Early Oncology Clinical Development.” *Pharmaceutical Statistics* 18 (5): 526–32.

Moskowitz, Chaya S., Xiaoyu Jia, Lawrence H. Schwartz, and Mithat Gönen. 2009. “A Simulation Study to Evaluate the Impact of the Number of Lesions Measured on Response Assessment.” *European Journal of Cancer* 45 (2): 300–310.

Mushti, Sirisha L., Flora Mulkey, and Rajeshwari Sridhara. 2018. “Evaluation of Overall Response Rate and Progression-Free Survival as Potential Surrogate Endpoints for Overall Survival in Immunotherapy Trials.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 24 (10): 2268–75.

Oubel, Estanislao, Eric Bonnard, Naoko Sueoka-Aragane, Naomi Kobayashi, Colette Charbonnier, Junta Yamamichi, Hideaki Mizobe, and Shinya Kimura. 2015. “Volume-Based Response Evaluation with Consensual Lesion Selection: A Pilot Study by Using Cloud Solutions and Comparison to RECIST 1.1.” *Academic Radiology* 22 (2): 217–25.

Rothe, Jan Holger, Christian Grieser, Lukas Lehmkuhl, Dirk Schnapauff, Carmen Perez Fernandez, Martin H.

Maurer, Axel Mussler, Bernd Hamm, Timm Denecke, and Ingo G. Steffen. 2013. “Size Determination and Response Assessment of Liver Metastases with Computed tomography—Comparison of RECIST and Volumetric Algorithms.” *European Journal of Radiology* 82 (11): 1831–39.

Schmid, Sabine, Stefan Diem, Qiyu Li, Mirjam Krapf, Lukas Flatz, Sebastian Leschka, Lotus Desbiolles, Dirk Klingbiel, Wolfram Jochum, and Martin Früh. 2018. “Organ-Specific Response to Nivolumab in Patients with Non-Small Cell Lung Cancer (NSCLC).” *Cancer Immunology, Immunotherapy: CII* 67 (12): 1825–32.

Schwartz, Lawrence H., Madhu Mazumdar, Wendy Brown, Alex Smith, and David M. Panicek. 2003. “Variability in Response Assessment in Solid Tumors: Effect of Number of Lesions Chosen for Measurement.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 9 (12): 4318–23.

Seymour, Lesley, S. Percy Ivy, Daniel Sargent, David Spriggs, Laurence Baker, Larry Rubinstein, Mark J. Ratain, et al. 2010. “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.” *Clinical Cancer Research*. https://doi.org/10.1158/1078-0432.ccr-09-3287.

Skougaard, Kristin, Mark James Dusgaard McCullagh, Dorte Nielsen, Helle Westergren Hendel, Benny Vittrup Jensen, and Helle Hjorth Johannesen. 2012. “Observer Variability in a Phase II Trial – Assessing Consistency in RECIST Application.” *Acta Oncologica*. https://doi.org/10.3109/0284186x.2012.667149.

Sridhara, Rajeshwari, Sumithra J. Mandrekar, and Lori E. Dodd. 2013. “Missing Data and Measurement Variability in Assessing Progression-Free Survival Endpoint in Randomized Clinical Trials.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 19 (10): 2613–20.

Tang, Youbao, Jinzheng Cai, Ke Yan, Lingyun Huang, Guotong Xie, Jing Xiao, Jingjing Lu, Gigin Lin, and Le Lu. 2021. “Weakly-Supervised Universal Lesion Segmentation with Regional Level Set Loss.” In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 515–25. Springer International Publishing.

Tovoli, Francesco, Matteo Renzulli, Giulia Negrini, Stefano Brocchi, Alessia Ferrarini, Andrea Andreone, Francesca Benevento, et al. 2018. “Inter-Operator Variability and Source of Errors in Tumour Response Assessment for Hepatocellular Carcinoma Treated with Sorafenib.” *European Radiology* 28 (9): 3611–20.

Trebeschi, S., S. G. Drago, N. J. Birkbak, I. Kurilova, A. M. Călin, A. Delli Pizzi, F. Lalezari, et al. 2019. “Predicting Response to Cancer Immunotherapy Using Noninvasive Radiomic Biomarkers.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 30 (6): 998–1004.

Trebeschi, Stefano, Zuhir Bodalal, Nick van Dijk, Thierry N. Boellaard, Paul Apfaltrer, Teresa M. Tareco Bucho, Thi Dan Linh Nguyen-Kim, Michiel S. van der Heijden, Hugo Jwl Aerts, and Regina G. H. Beets-Tan. n.d. “Development of a Prognostic AI-Monitor for Metastatic Urothelial Cancer Patients Receiving Immunotherapy.” *Frontiers in Oncology*. https://doi.org/10.3389/fonc.2021.637804.

Villaruz, Liza C., and Mark A. Socinski. 2013. “The Clinical Viewpoint: Definitions, Limitations of RECIST, Practical Considerations of Measurement.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 19 (10): 2629–36.

Wang, Meihua, Cong Chen, Thomas Jemielita, James Anderson, Xiaoyun Nicole Li, Chen Hu, S. Peter Kang, Nageatte Ibrahim, and Scot Ebbinghaus. 2019. “Are Tumor Size Changes Predictive of Survival for Checkpoint Blockade Based Immunotherapy in Metastatic Melanoma?” *Journal for Immunotherapy of Cancer* 7 (1): 39.

Wulff, A. M., M. Fabel, S. Freitag-Wolf, M. Tepper, H. M. Knabe, J. P. Schäfer, O. Jansen, and H. Bolte. 2013. “Volumetric Response Classification in Metastatic Solid Tumors on MSCT: Initial Results in a Whole-Body Setting.” *European Journal of Radiology* 82 (10): e567–73.

Zimmermann, Markus, Christiane Kuhl, Hanna Engelke, Gerhard Bettermann, and Sebastian Keil. 2021. “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?” *Polish Journal of Radiology / Polish Medical Society of Radiology* 86 (October): e594–600.