Iannessi *et al. Insights into Imaging* (2024)15:199
https://doi.org/10.1186/s13244-024-01774-w

EUROPEAN SOCIETY OF RADIOLOGY
Insights into Imaging

EDUCATIONAL REVIEW

Open Access

RECIST 1.1 assessments variability: a systematic pictorial review of blinded double reads

Antoine Iannessi[1,2], Hubert Beaumont[2*], Christine Ojango[2], Anne-Sophie Bertrand[3] and Yan Liu[2]

Abstract

Reader variability is intrinsic to radiologic oncology assessments, necessitating measures to enhance consistency and accuracy. RECIST 1.1 criteria play a crucial role in mitigating this variability by standardizing evaluations, aiming to establish an accepted "truth" confirmed by histology or patient survival. Clinical trials utilize Blind Independent Centralized Review (BICR) techniques to manage variability, employing double reads and adjudicators to address inter-observer discordance effectively.

It is essential to dissect the root causes of variability in response assessments, with a specific focus on the factors influencing RECIST evaluations. We propose proactive measures for radiologists to address variability sources such as radiologist expertise, image quality, and accessibility of contextual information, which significantly impact interpretation and assessment precision. Adherence to standardization and RECIST guidelines is pivotal in diminishing variability and ensuring uniform results across studies.

Variability factors, including lesion selection, new lesion appearance, and confirmation bias, can have profound implications on assessment accuracy and interpretation, underscoring the importance of identifying and addressing these factors. Delving into the causes of variability aids in enhancing the accuracy and consistency of response assessments in oncology, underscoring the role of standardized evaluation protocols and mitigating risk factors that contribute to variability. Access to contextual information is crucial.

Critical relevance statement By understanding the causes of diagnosis variability, we can enhance the accuracy and consistency of response assessments in oncology, ultimately improving patient care and clinical outcomes.

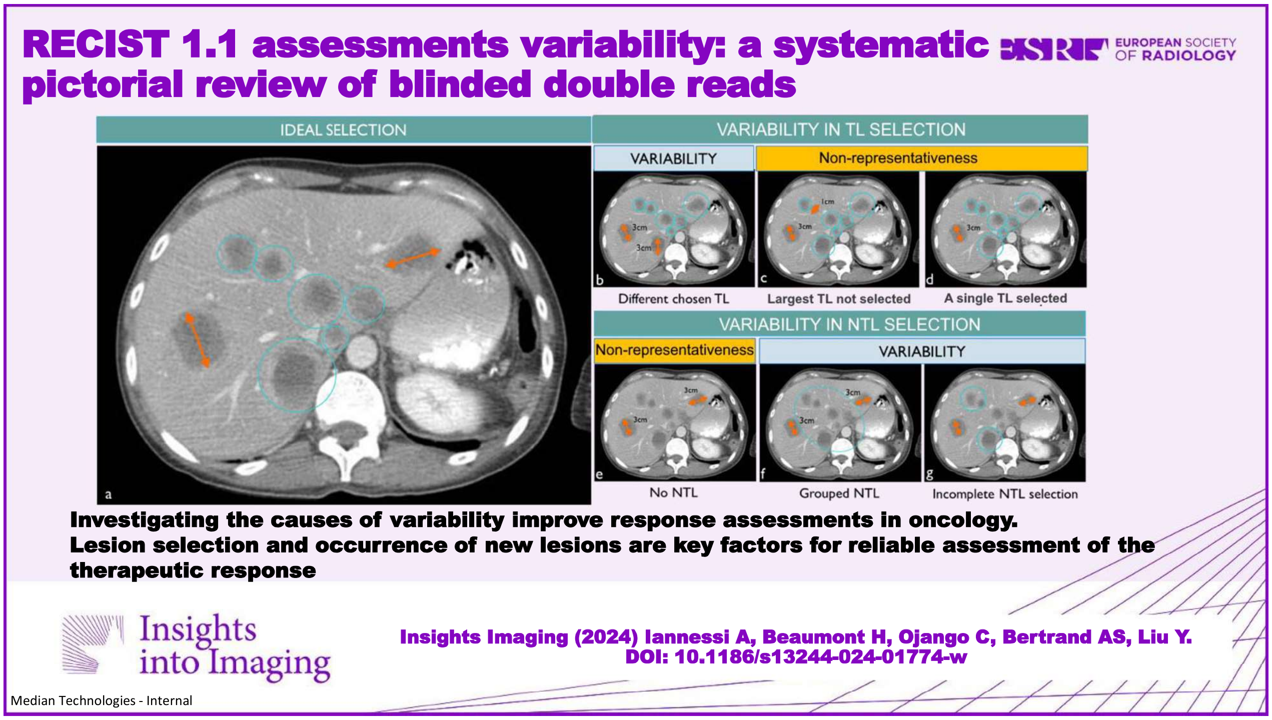
Key Points

- Baseline lesion selection and detection of new lesions play a major role in the occurrence of discordance.
- Image interpretation is influenced by contextual information, the lack of which can lead to diagnostic uncertainty.
- Radiologists must be trained in RECIST criteria to reduce errors and variability.

Keywords Diagnostic errors, Statistics & numerical data, Quality improvement, Oncology, RECIST 1.1

*Correspondence:
Hubert Beaumont
hubertbeaumont@hotmail.com
Full list of author information is available at the end of the article

Graphical Abstract



RECIST 1.1 assessments variability: a systematic pictorial review of blinded double reads

Investigating the causes of variability improve response assessments in oncology. Lesion selection and occurrence of new lesions are key factors for reliable assessment of the therapeutic response

Insights Imaging (2024) Iannessi A, Beaumont H, Ojango C, Bertrand AS, Liu Y. DOI: 10.1186/s13244-024-01774-w

Median Technologies - Internal

Introduction

In the field of radiology, the absence of an indisputable "truth" in images can lead to different diagnoses based on the same image, resulting in inter-observer discordance of opinions.

In oncology, imaging-based evaluations are used to predict an accepted "truth" which is generally confirmed by histology or patient survival. Since this "truth" is not available at the time of analysis, errors can only be determined retrospectively. However, oncology evaluations are standardized using quantified criteria designed to decrease observers' variability. Non-compliance with standardized reading instructions can be considered as avoidable errors, not just differences in opinions.

In clinical trials of cancer treatment efficacy, Blind Independent Centralized Review (BICR) with double reads is recommended by the Food and Drug Administration to control variability and avoid presumed investigator bias when the subject's treatment assignment is known [1]. When discrepancies are detected between those radiologists, a third reader, known as an Adjudicator, is typically consulted to resolve the case and decide which of the two readers made the most accurate interpretation [2]. This reading paradigm also provides an opportunity to study the causes of inter-observer discordance during patient follow-up [3, 4].

In this analysis, we aim to systematize the causes of inter-observer variability in response assessments. We first describe the general construction of a quantified tumor burden follow-up criterion to identify the sources of variability in response assessments. Then, we focus on the four most significant variability factors identified during Response Evaluation Criteria in Solid Tumor (RECIST 1.1) evaluations in several centralized reviews [5]. By investigating the causes of variability, we aim to improve the accuracy and consistency of response assessments in oncology.

Response evaluation criteria and RECIST 1.1 variability factors

In oncology, response evaluation criteria supply standardized reading rules to capture the evolution of cancer disease. These criteria have been shown to correlate with the gold standard of cancer treatment efficacy, which is ultimately "overall survival" [6]. Because Overall Survival is often a costly gold standard to refer to, in terms of patient benefits and trial resources, surrogate endpoints are accepted by drug approval agencies as primary endpoints in clinical trials [1]. The aim of Phase 2 studies is to evaluate treatment efficacy through a response objective, i.e., the "overall response rate", hypothesizing that a decrease in tumor size theoretically correlated with a benefit on

Table 1 Categorical variability factors for "quantified" tumor assessment criteria

Lesion	Assessment (PD, PR, CR)	Variability factors
Target lesions	SOD↗ ≥ 20%/	• *Initial selection: number and location and organ belonging*
	SOD↘ ≥ 30%/	• Measurement
	SOD = 0	• *Qualitative assessment of the "disappearance" of small lesions*
Non-target lesions	Significant increase/completely disappeared	• *Initial selection: number and location and organ belonging*
		• *Qualitative assessment of the "significance" of the progression*
		• *Qualitative assessment of "non-pathological" status of nodes*
		• *Qualitative assessment of "disappearance" of small lesions*
New lesions	Absence/presence	• Detection (False Positive/False Negative)
		• *Qualitative assessment of "indisputability" (characterization)*

SOD sum of diameters, *PD* progressive disease, *PR* partial response, *CR* complete response. In italics, the subjective items of evaluation mostly related to a "characterization" task affect each lesion category and type of outcome

survival. Phase 3 trials confirm the benefit of a large population through a progression-free survival objective. A response evaluation criterion is designed to support those clinical trial aims. Commonly, response criteria define an initial baseline status which is then compared at fixed intervals over time [7]. The definition of the reference disease is quantified for one part and qualified for the other part. The semi-quantitative nature of baseline assessment lies in the fact that part of the disease may not be reliably measured, or measurement is too demanding. During follow-up, evaluations must be consistent with the baseline-defined disease and additionally, the appearance of new tumor lesions is scrutinized.

Considering how RECIST has been designed around three lesion categories, i.e., quantified target lesions (TLs), qualified non-target lesions (NTLs), and new lesions (NL), it is possible to break down its general causes of discordance related to the dimension of quantification and qualification [7].

Target lesions are all measured and their measured summed. The change of their measured sum is compared to the corresponding baseline and depending on a threshold of change, patients are diagnosed as stable, partially or completely responding, or progressing. The variability in TL selection at baseline or their measurements all along the process impacts the patient evaluation into the categories. Furthermore, the near-complete disappearance of lesions is subject to qualitative variability in evaluating the persistence of residual tumors or tissue considered scarred.

Non-target lesions are evaluated qualitatively. As for TL, their initial selection will impact the subsequent response and the evaluation of their complete disappearance is subject to divergent opinions, when display may correspond to scar tissue. In addition, lymph node sites that never disappear must be subjectively qualified as "non-pathological" to allow for a complete response. Regarding progression, in the absence of a formal quantified threshold, the subjective nature of the evaluation also leads to greater variability in the "significance" of progression.

New lesions are prone to observers' variability in terms of detection and characterization due to their occasionally equivocal display.

Even if the oncological evaluation criteria are said to "quantify" the response, they remain subjective from several standpoints that impact the three main categories determining the overall response.

To summarize, we can consider three main criteria-derived factors of discordance between two radiologists which are: The initial lesions selection, the measurement of TLs, the qualitative evaluation of NTLs, and the detection of NLs (Table 1) [5].

RECIST variability in initial disease evaluation

The literature suggests that a large contribution to the response discrepancies is due to a difference in the definition of the disease at the baseline [8–10].

We illustrate hereafter the 3 causes of this baseline-derived variability.

Ill-defined tumor selection criteria (Fig. 1)

Whereas RECIST gives clear selection rules for analyzing the disease at baseline, the cancer presentation however may be multi-metastatic so the choice of TL/NTLs, number, and distribution may vary between radiologists without infringing these rules [11]. Consequently, the "insufficient" whole disease representativity of the selection leads to response variability [12, 13].

Reader non-compliance with criteria (Fig. 2)

A non-compliance to the criteria's rules of selection or measurability may be explained by a poor knowledge of the guidelines (Fig. 1), or occasionally by a flawed selection due to a lack of expertise in cancer imaging (Fig. 2) [9].

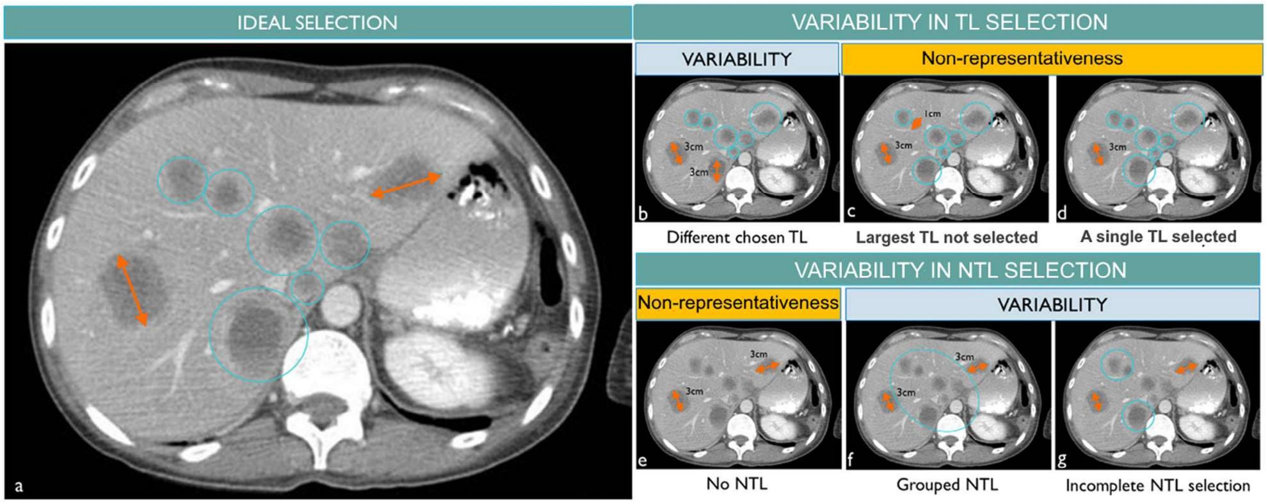


Fig. 1 Acceptable reader variability of baseline selection considering RECIST 1.1 guidelines. The purpose of the baseline selection guideline is to standardize a fashion to best represent the disease by selecting the maximum and largest lesions (**a**). However, for practical use, the RECIST 1.1 criteria include room for variability in baseline assessment by limiting the number of target lesions (in red) to 2 per organ and 5 in total. Then, for patients with multiple lesions, there is a compliant inter-reader variability in the assessment (**b**, **f**, **g**). Besides the maximum number of lesions a justifiable reason such as equivocality of the finding, robustness of the measurement, or information about previously irradiated lesion must explain the variability but at the same time challenge the representativity goal of the selection task (**c**, **d**, **e**)

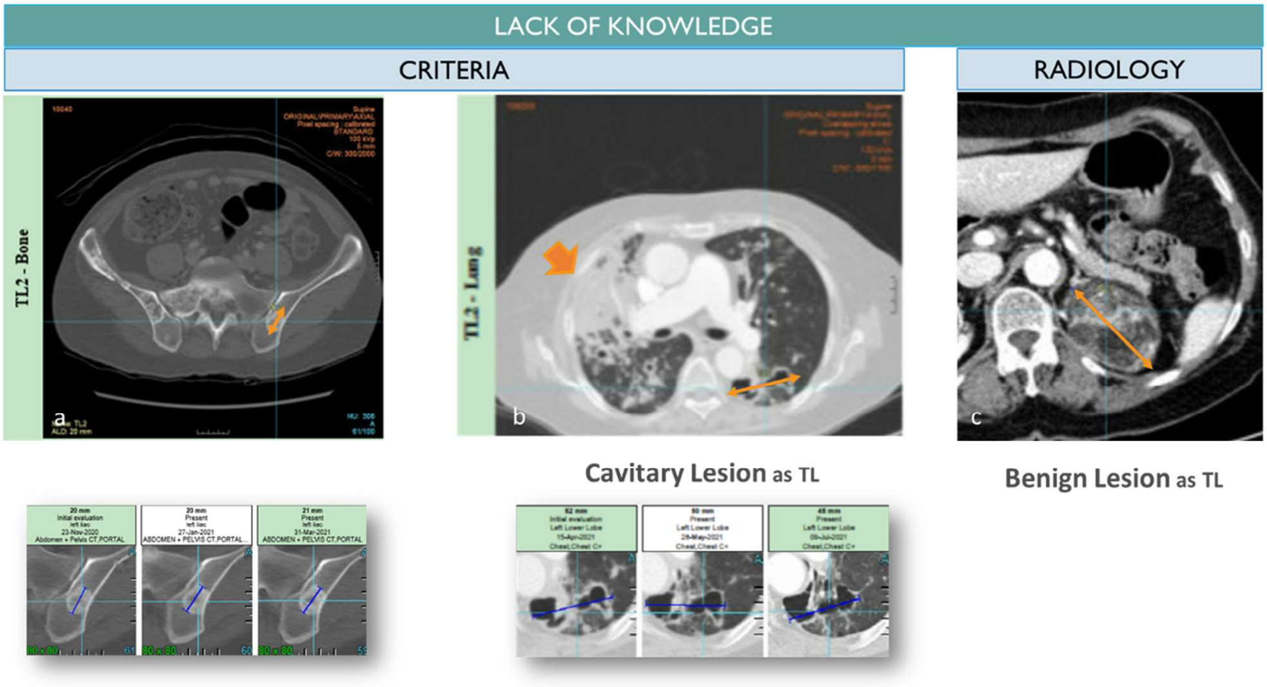


Fig. 2 Baseline discrepancy caused by a lack of training. The lack of knowledge is not a frequent provider of variability between readers however often it results in errors of guidance when the reader does not master the criteria and selects bone blastic lesions as measurable lesions (**a**) or cavitary lesions while some more robust lesions for measurement is available (**b**). Rarely, the selection wrongly includes an unequivocally benign lesion due to a radiologist's lack of knowledge such as the adrenal myelolipoma containing fat (**c**)

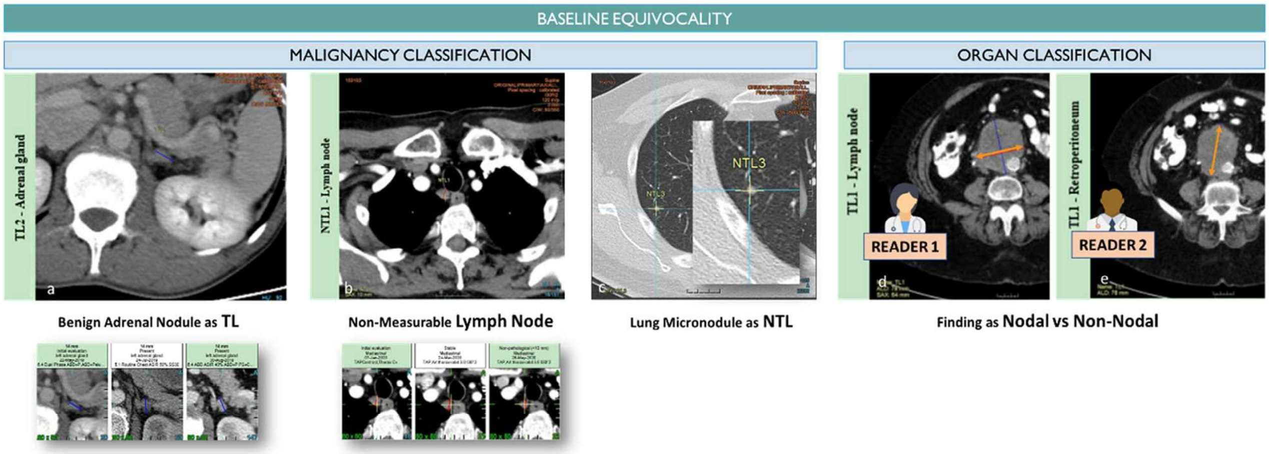


Fig. 3 Baseline discrepancy caused by equivocality on the disease malignancy or the finding class. Equivocality at baseline is frequent due to the lack of previous examination or censored information for a blind evaluation. In such conditions, small lesions, e.g., adrenal nodules, juxta centimetric lymph node, and lung micronodules, might be wrongly considered as disease-related findings at baseline while they will remain stable over time and prove their benign characteristic (**a**, **b**, **c**). In some cases, the reader might select the same target but with a different organ classification (**d**, **e**). For RECIST 1.1, depending on how the same lesion is considered to be a lymph node or a mass, the specific method of measurement and threshold of the nodal lesion can impact the overall assessment

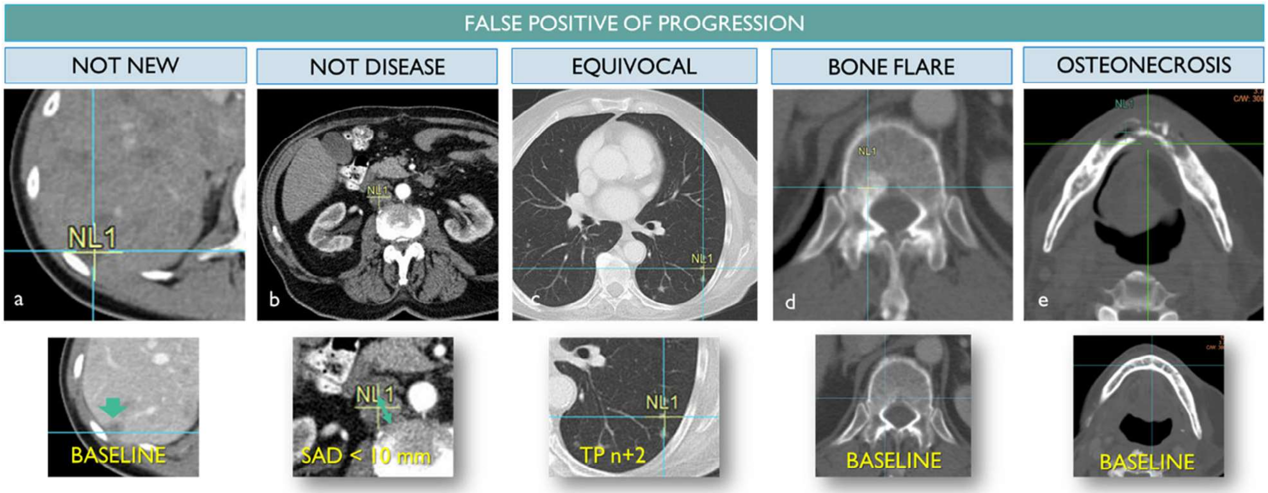


Fig. 4 False alarm for progression. From left to right, to define a new lesion, the finding should be new and not present in the baseline (**a**), considered pathologic so ≥ 1 cm for lymph nodes (**b**), unequivocally related to the cancer disease spreading unlike a micronodule possibly linked to intercurrent inflammation (**c**), not related to a blastic bone healing phenomenon (**d**) or a treatment-induced adverse event such as osteonecrosis of the jaw (**e**)

Equivocal disease presentation (Fig. 3)

Due to healthy and diseased overlapped imaging presentation, sole imaging information can be insufficient to discriminate a malignant lesion from a benign finding leading to antagonist selections in trial with double reading. The selection of such equivocal findings by one of the readers as early as at baseline can lead to including in his tumor burden lesions that will remain stable over time and will impair the sensitivity of the response thresholds, Fig. 2 [14]. The impact of equivocal disease presentation is worse in the BICR setting with respect to routine-based RECIST evaluations where part of the patient's clinical information is censored. Censored information that would have been helpful to clear a doubtful image such as historical data. Literature also demonstrates that the reader concordance of selection is lower for specific organ locations such as lymph nodes and peritoneal disease [15].

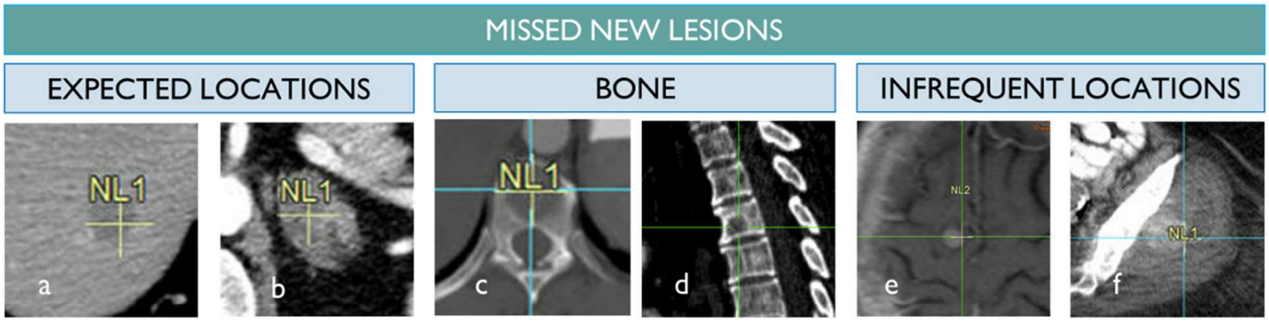


Fig. 5 False Negative for Progression. The most frequent metastatic sites for lung cancer depend on the primary site, such as the liver and adrenal gland (**a**, **b**). While these lesions can be missed, more often, new lesions are missed when they are located in infrequent locations like the bone due to their low conspicuity in axial view (**c**) compared to sagittal reformats (**d**). Infrequent locations, such as the brain or soft tissue, are unexpected and more susceptible to attention bias (**e**, **f**)

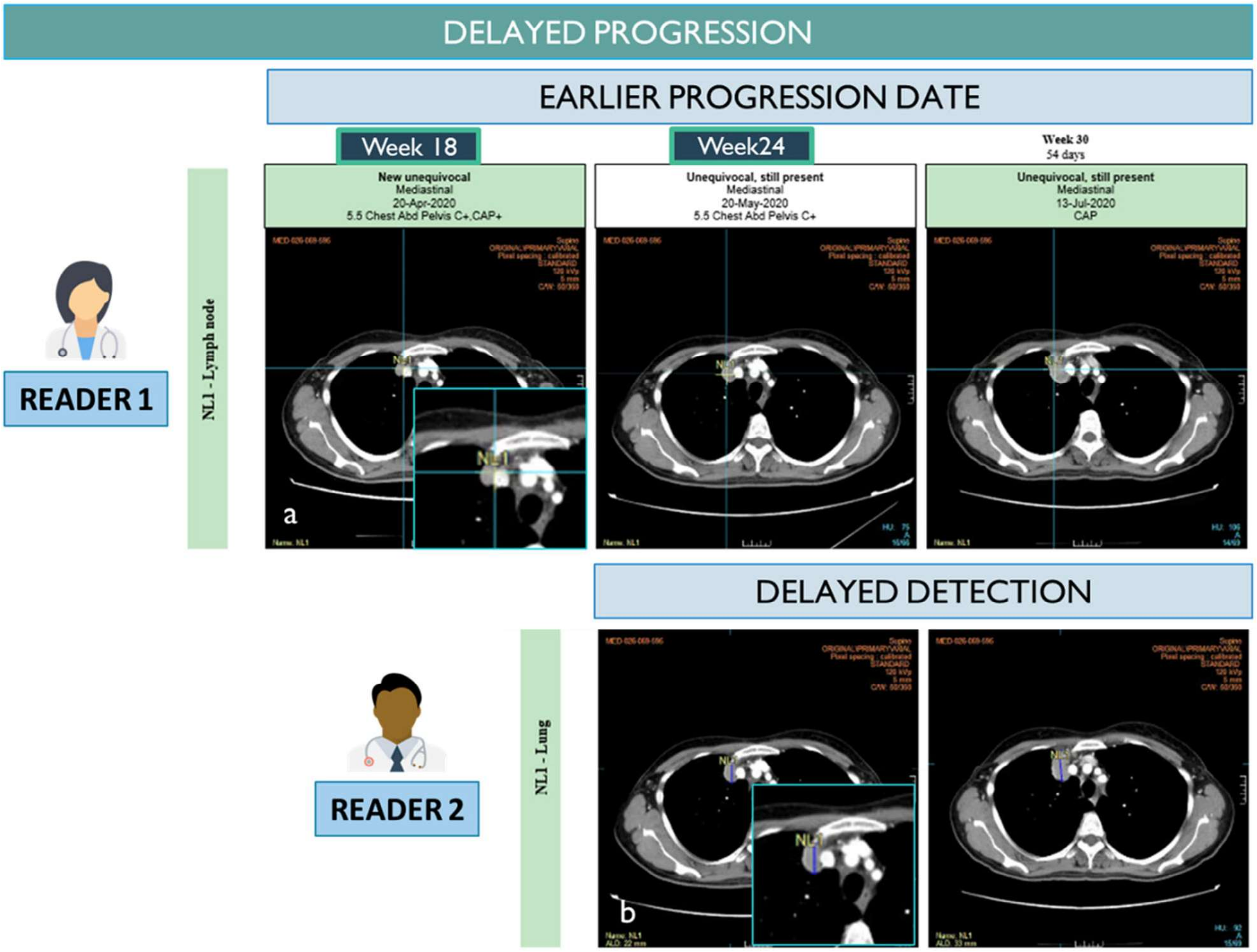


Fig. 6 Delayed detection. The two radiologists identified the same new lesion (**a**, **b**) but radiologist 2 detected it one visit later leading to a 6 weeks discrepant date of progression

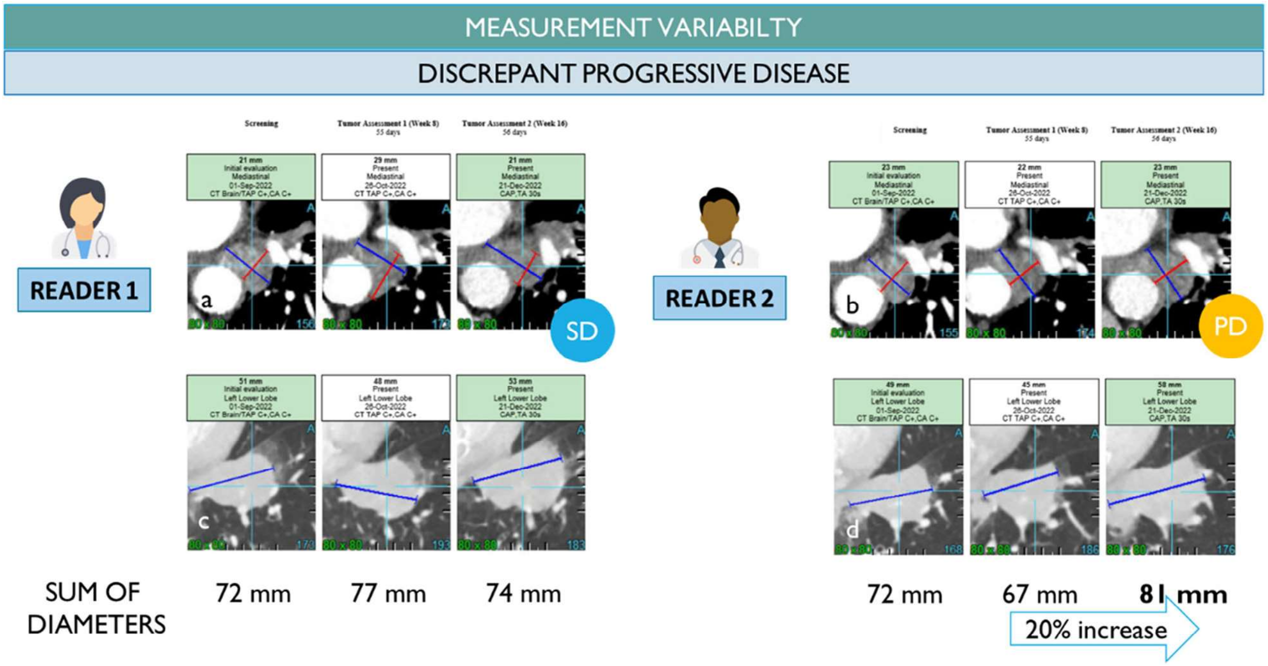

MEASUREMENT VARIABILTY
DISCREPANT PROGRESSIVE DISEASE
READER 1
READER 2
Screening
Tumor Assessment 1 (Week 8)
55 days
Tumor Assessment 2 (Week 16)
56 days
21 mm Initial evaluation Mediastinal 01-Sep-2022 CT Brain/TAP C+,CA C+
29 mm Present Mediastinal 26-Oct-2022 CT TAP C+,CA C+
21 mm Present Mediastinal 21-Dec-2022 CAP,TA 30s
23 mm Initial evaluation Mediastinal 01-Sep-2022 CT Brain/TAP C+,CA C+
22 mm Present Mediastinal 26-Oct-2022 CT TAP C+,CA C+
23 mm Present Mediastinal 21-Dec-2022 CAP,TA 30s
51 mm Initial evaluation Left Lower Lobe 01-Sep-2022 CT Brain/TAP C+,CA C+
48 mm Present Left Lower Lobe 26-Oct-2022 CT TAP C+,CA C+
53 mm Present Left Lower Lobe 21-Dec-2022 CAP,TA 30s
49 mm Initial evaluation Left Lower Lobe 01-Sep-2022 CT Brain/TAP C+,CA C+
45 mm Present Left Lower Lobe 26-Oct-2022 CT TAP C+,CA C+
58 mm Present Left Lower Lobe 21-Dec-2022 CAP,TA 30s
a
b
c
d
SD
PD
SUM OF DIAMETERS
72 mm
77 mm
74 mm
72 mm
67 mm
81 mm
20% increase

Fig. 7 Inter-observer measurement variability. At baseline, the two radiologists selected the same targets as mediastinal node (**a**, **b**) and lung nodule (**c**, **d**) with variable measurement at the lesion level but compensated when summing the index lesions (identical sum of diameters). Then, during the follow-up, the measurement variability leads to a patient response discrepancy with progressive disease (20% increase) assessed by reader 2 versus non-progressive disease for reader 1 while the difference of the sum of diameter at the visit 1 is 10 mm and at visit 2 is 7 mm

In addition to the uncertainty of malignancy, even for a lesion that is obviously cancerous, its belonging to a type of organ can be equivocal. It is frequently observed for intra-abdominal tissue nodules or neck nodules after surgery, which are reported variably as lymph nodes or carcinosis nodules. This classification affects the measurement method (short axis or long axis) and evaluation (pathological threshold for lymph nodes).

RECIST variability in the recording of new lesions

The literature reports that 1/3 of discrepant detection of NLs during follow-up [16]. Four causes of discrepancy are illustrated below.

False detections (Fig. 4)

The radiologist might wrongly consider a newly detected finding as an NL.

To qualify for an NL, a finding must be unequivocally "new", "present", and "cancerous".

Indeed, it should be noted that an NL can only be defined as such if it is not present at baseline because otherwise, it belongs either to the TL or to the NTL group and benefits from qualitative or quantitative follow-up.

Also, if an NL is small, it could be a result of an artifact or a higher image definition compared to the baseline. Ultimately, a new and indisputable anomaly is not necessarily cancerous during RECIST follow-up:

a. An intercurrent lung consolidation may mimic an NL.
b. The malignant and reactive lymph nodes are a differential diagnosis that can lead to variability. Therefore, RECIST specifies a threshold of 10 mm short axis to be considered lymph node disease [17].
c. Treatment-related changes must be known. The "bone flare" phenomenon has been documented for a long time and it leads to the condensation of sometimes subtle bone lesions initially mimicking the appearance of new condensing lesions [18]. Only the appearance of a lytic or extraosseous tissue lesion should be considered as an NL.

The reporting of an NL during follow-up leads to classifying the corresponding patient as progressing, triggering early discontinuation of his treatment. In this respect, the RECIST 1.1 criterion is conservative. In ambiguous situations, the guidelines propose to be decided through later follow-up. This follow-up allows us to confirm the progressive nature of the finding with more certitude.

Non-detections (Fig. 5)

Non-detection during one or more time points is observed and involves usual sites of metastasis such as the

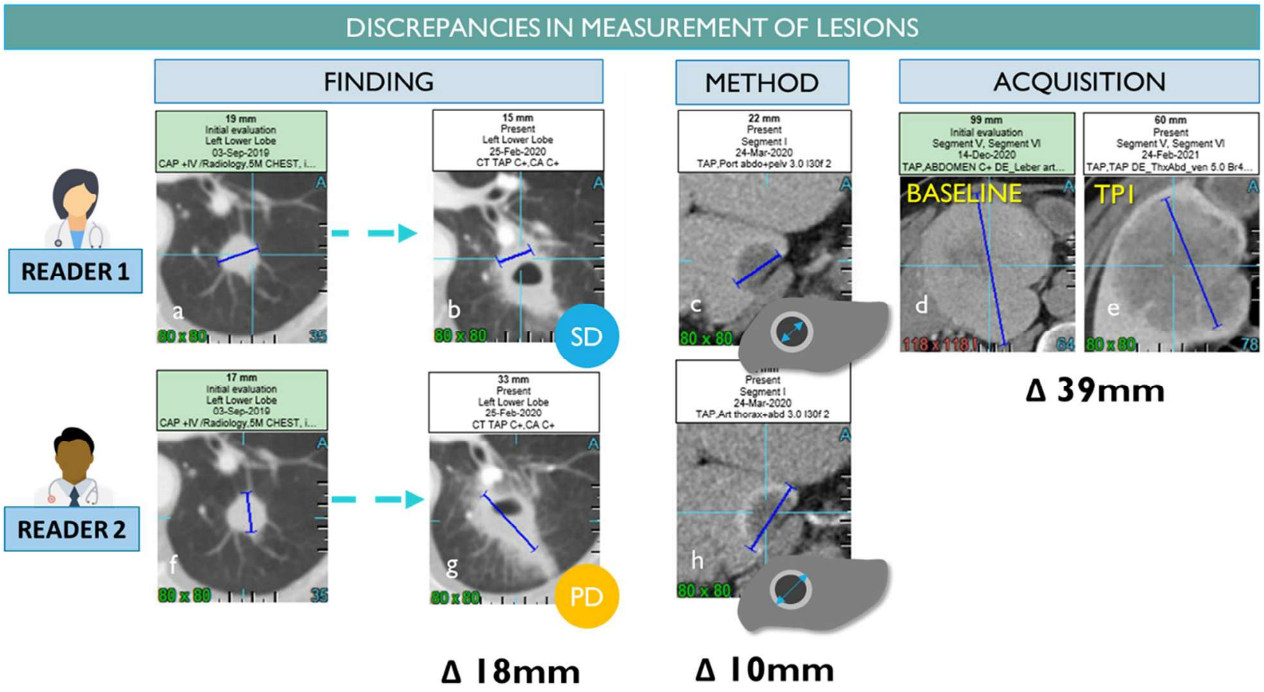


Fig. 8 Measurement variability factors (inter and intra-observers). The discrepancy in measurements can lie inside the finding such as cavitation or speculation and difficulty of measurement (**a**, **b**, **f**, **g**), the method of measurement such as rim included or not in the measurement (**c**, **h**), the selection of a different phase of injection even for a same reader during follow-up (**d**, **e**)

liver or the lung. On the other hand, more often the discordant sites are either in the bone or in less frequent metastatic sites [19]. It is possible that a tunneling bias of attention plays an important role in these occurrences [20]. Indeed, bone lesions mostly are low conspicuity because they are lytic or sparsely dense in a dense environment, so they require an active search by reformatting and adapted windowing because otherwise, the radiologist cannot see them. Also, unexpected sites such as first slices or soft tissue are responsible for non-detection [21].

Under-detections

Under-detection corresponds to incomplete detection of several NLs. This phenomenon has theoretically little impact on a discordant response according to RECIST since a single NL is sufficient to trigger progression. This has been well documented and corresponds to 20% of evaluation errors in oncology. It reflects a premature loss of attention by the radiologist after the first phase of successful research and is frequently described as “satisfaction of search” bias [22].

Delayed-detections (Fig. 6)

It is common to observe discordance in the date of progression between two readers, leading to different progression-free survival rates (around 30% in the literature) [19]. Part of the explanation lies in the size of the NL, which becomes more visible over time, reducing the risk of perception discordance. Another part of the explanation lies in the unacknowledged subjectivity of the RECIST criterion about the significance of disease progression. It is possible that the radiologist does not identify a new anomaly because, in their opinion, it does not represent sufficient progression based on the rest of the evolution and is therefore somewhat equivocal. Delayed detections frequently involve isolated appearances of equivocal findings, such as lung micronodules and centimeter-sized lymph nodes.

RECIST variability in measurements of the lesion (Figs. 8–11)

The baseline introduces a measurability cutoff of 1 cm for solid tumors to qualify as TL versus NTLs. This rule is supported by a reduction in reader variability and the expected shrinkage of tumors after treatment, requiring a minimal initial tumor burden to interpret changes versus variability [23]. It should be noted that the “measurability” criterion applies per lesion, but no measurable criteria apply to the patient. For this reason, patients are eligible for quantification from at least one lesion of 1 cm, which raises questions about the variability of response

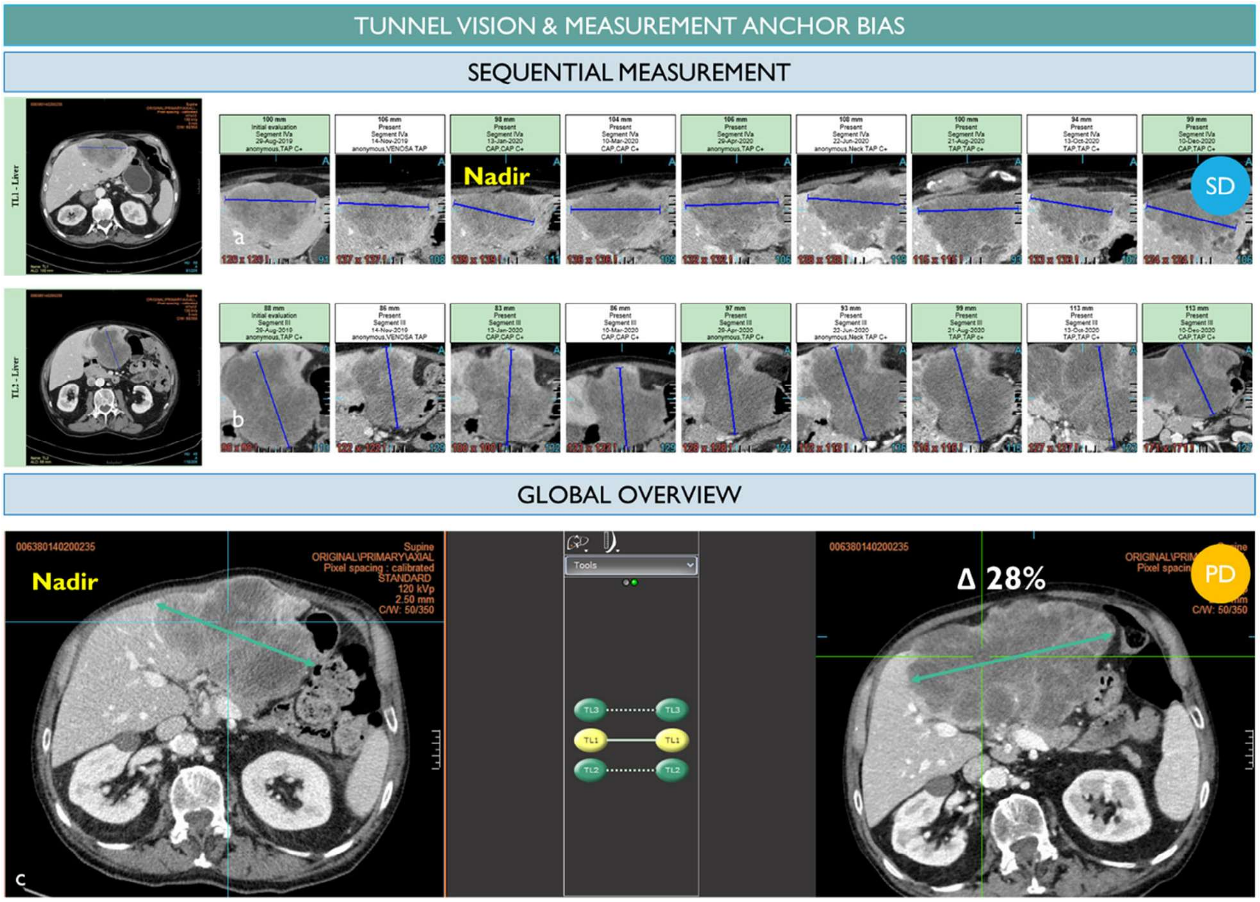


Fig. 9 Anchor bias in measurement. The radiologist chose 2 targets in the liver and his assessment concluded to be a stable disease after 9 evaluations (**a**, **b**). However, with a retrospective look, one can visually observe a significant increase in the same target measured (**c**). We suspect a confirmation bias in the measurement workflow timepoint after timepoint

assessment in patients with smaller tumor burdens compared to those with larger tumor burdens.

During follow-up, the same use of threshold explains that even a millimetric difference between two radiologists can be responsible for a discrepancy between progression and stable disease, and even progression and complete response, especially after a response when tumor burden includes lymph node elements (Fig. 7).

Measurement is repeated for each timepoint and the sources of the sum of diameter variability are illustrated in Fig. 8 and can be broken down as follows [24]:

a. Disease distribution, location, finding appearance (e.g., single or multiple lesions, type of organ such as head and neck are difficult to delimitate with respect to lung nodules) [25, 26]
b. Non-standardized measurement method (e.g., measurement rules for cavitary lesions)
c. Inappropriate windowing
d. Inherent reader variability (i.e., inter-observer variability higher than intra-observer) [15]
e. Inconsistent choice of series for annotations (e.g., inconsistent acquisition times) [16]
f. Poor image quality (cf. discussion)

Furthermore, in RECIST analysis, those repeated measurements are typically performed sequentially in a workflow, and this can make radiologists vulnerable to confirmation bias, leading to a tendency to reconfirm previous measurements [27]. As a result, the overall change in lesion size may be minimized (as shown in Fig. 9). If a radiologist works differently and systematically with respect to the baseline or nadir, the measured lesion size may differ accordingly.

Finally, the discordance of responses is seen when lesions have become very small without complete disappearance. The measurement of a small residual lesion for which the measurement can be continued indefinitely without obtaining a complete response. In this case, the

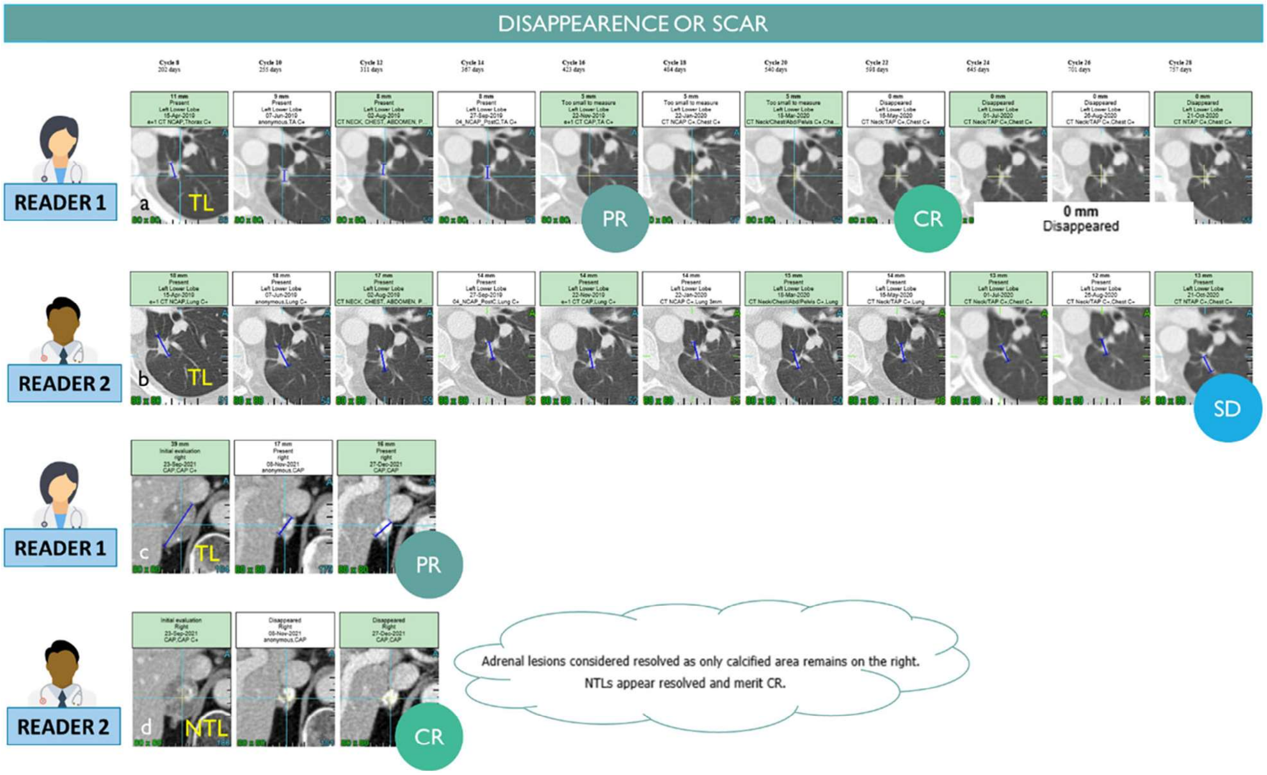

DISAPPEARENCE OR SCAR
READER 1
a
TL
PR
CR
0 mm
Disappeared
READER 2
b
TL
SD
READER 1
c
TL
PR
READER 2
d
NTL
CR
Adrenal lesions considered resolved as only calcified area remains on the right.
NTLs appear resolved and merit CR.

Fig. 10 Discrepancy on remaining visible findings after treatment. The same lung nodule target was chosen by two radiologists (**a**, **b**). Radiologist 1 considers the remaining visible finding as a scar while radiologist 2 continues to measure the remaining disease (18 mm) preventing a complete response for this patient. The same adrenal gland finding was chosen as TL by reader 1 and NTL by reader 2 (**c**, **d**). Unlike reader 1, reader 2 considers that the lesion disappears with remaining calcified scars although it is still visible and measured by reader 2

lesions no longer decrease but no longer grow and could correspond to a scar (Fig. 10). In characterizing whether a complete response has been achieved, the radiologist's judgment is therefore necessary to consider the overall level of response. In practice, the discrepancy between PR and CR is less important from a clinical trial perspective since the overall response rate is calculated on the whole PR + CR, and the date of the first response is not influenced.

RECIST variability in the assessment of the not-measured disease (Fig. 11)

First, the evaluation of the NTL is subjective as it concerns the truly non-measurable diseases such as blastic bone metastasis and the not-measured baseline residual disease after TL selection. Also, NTLs are sometimes grouped together to ease the follow-up record of diffuse disease (Fig. 1).

Dissociated response is defined as the coexistence of responding and non-responding lesions within the same patient. It is particularly observed with immunotherapy where such a paradoxical response between TL and NTL (Fig. 11) can be a difficulty [28, 29]. Unlike TL, the regular workflow that consists of comparing lesions to their Nadir is not well suited for monitoring NTLs as it is tricky to identify the Nadir of NTL without plotting a quantified measure of their changes. Therefore, in practice, NTL evaluations are compared to the last available examination.

Second, the evaluation of each NTL is aggregated subjectively into an NTL category response and RECIST states that NTL assessment is a function of the overall disease and therefore should consider the evolution of TLs and the presence of NLs [30].

Because of these above different factors, there is interobserver variability in the evaluation of NTL categories responsible for discordance in the overall response. As explained, this occurs even more when the TL/NTL response is dissociated.

Previously mentioned difficulties in assessing complete response versus non-complete response for small remaining lesions versus scars also concern the NTL category (Fig. 11). However, even larger lesions are concerned for non-measurable bone lesions when a lytic lesion heals as a blastic lesion (bone flare) or blastic

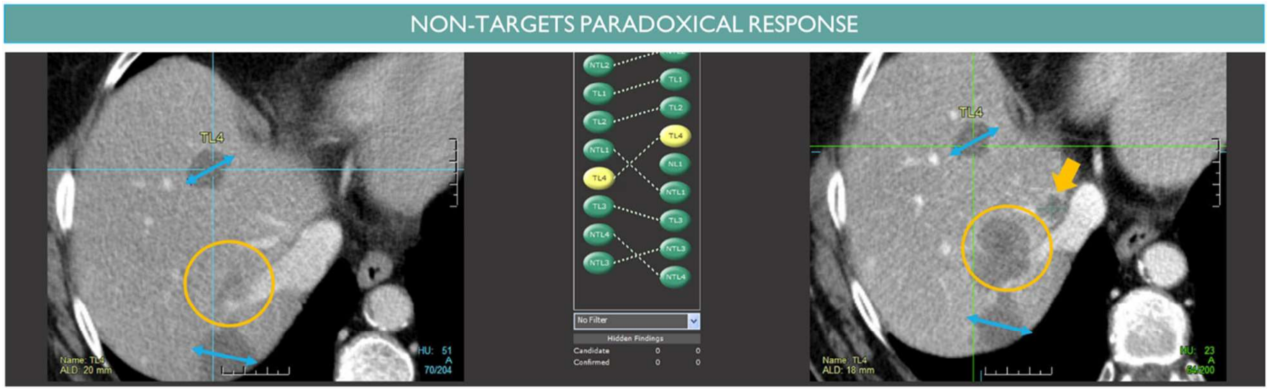


Fig. 11 Paradoxical response of non-targeted disease. The measured disease is stable (double arrowhead) while the non-targeted disease (circle) increases unequivocally. A small new lesion can also be identified (arrow)

Table 2 Radiologist RECIST 1.1 toolbox of preventive action

Selection of lesions	• Expertise and training on the criteria's read rules for follow-up.
	• Comprehensiveness and Representativeness of the choice
	• Ensuring non-equivocality of lesions, particularly for target lesions
Detection of new lesions	• Always checking the baseline to confirm the absence of new lesions
	• Measuring juxta centimetric lymph nodes
	• Confirming equivocal and micro-lesions using follow-up imaging
	• Using a checklist to avoid in attentional bias and actively search in certain areas not expected
	• Including bones on the checklist:
	○ Beware of the appearance of new condensed bone lesions and recognize the bone flare-up effect
	○ Using sagittal and coronal reformatting with bone windowing
	• After detecting a new lesion:
	○ Always looking for a second metastasis finding the first one
	○ Checking previous examinations to precisely date the progression of the disease
Assessment of non-target lesions	• Performing a comparison of the current examination with the baseline and an intermediate timepoint to verify the absence of dissociated response to the targets and to get a more global view of the evolution.
Measurement of targets	• Measuring with the same method throughout the analysis
	• Using the same injection time for the measurements
	• Performing a comparison of the current examination to the baseline and an intermediate nadir to get a global perspective and avoid confirmation bias.
	• Utilizing semi-automatic measurement tools when available
	• Not measuring a target when the quality for robust measurement is not met (non-evaluable)

List of items to control variability and discordance rate

lesions persists throughout the evaluation with no increase (Figs. 2a and 5d). in such case, the assessment of complete response versus remaining bone metastasis is subjective. The same difficulty in distinguishing CR versus non-CR for cystic or cavitary lesions after treatments (Fig. 2b).

Discussion

In our pictorial overview, we highlight that each response category in the RECIST criteria is subject to significant sources of reader variability. The expertise of the radiologist certainly plays an important part [31] such as the patient assessment performed by a single observer [32]. Therefore, for this observer, we propose to summarize individual preventive actions that radiologists may find valuable into a "toolbox," as detailed in Table 2 [33].

The literature provides figures on the causes of RECIST variability, which can help prioritize actions using a Pareto strategy based on the context of the evaluation [34, 35]. In phase 3 or adjuvant settings of clinical trials, progression is the key event of interest, and strategies to minimize variability may focus on assessing NLs. Conversely, in phase 2 trials, the primary event of interest is the first response, and thus, the measurement of the disease and choice of lesions are critical components of the assessment that require close monitoring.

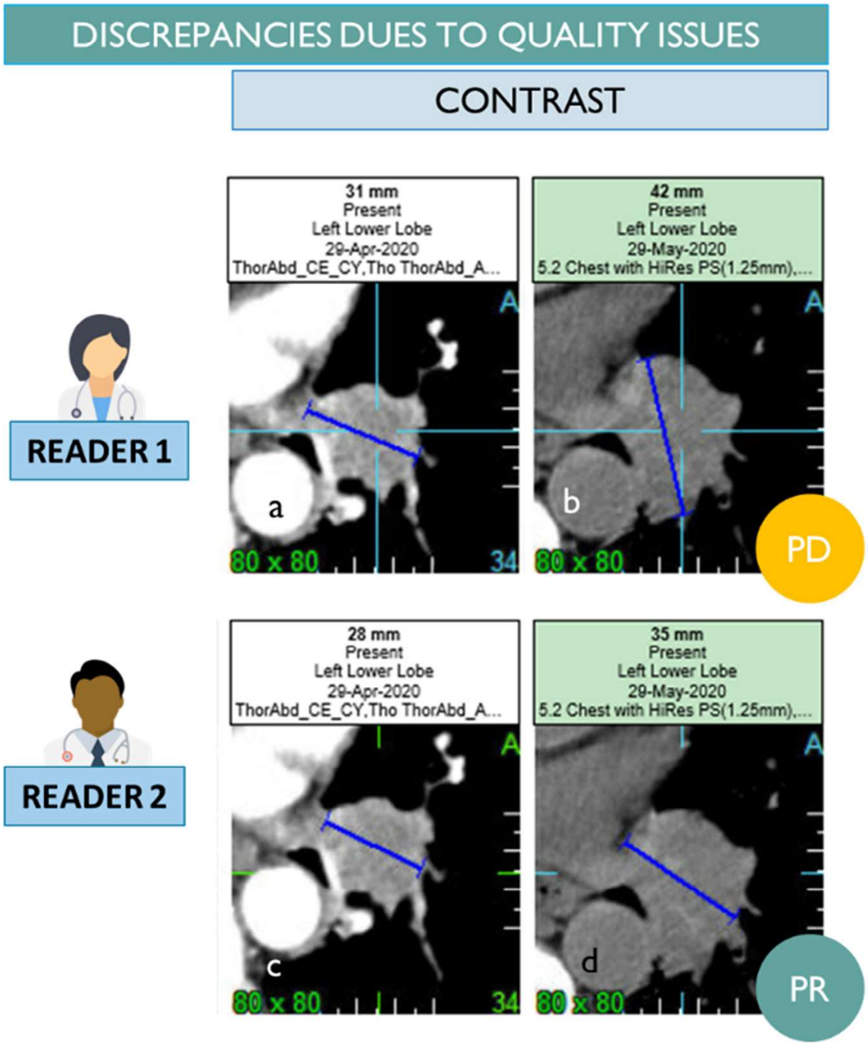


Fig. 12 Image quality-related discrepancy. One lung para-mediastinal target is measured with mediastinal windowing and good visualization of adjacent vascular structures (**a**, **c**). During the follow-up, one CT evaluation is performed without contrast injection (intercurrent contra-indication). The higher inter-reader variability of measurement due to the lack of contrast in the image resulted in a progression discrepancy (**b**, **d**)

In addition to the radiologist-related causes of variability, we should consider two other sources of variability related to the information provided to those observers.

The first source is related to the quality of the images provided for assessment [36]. To comply with the RECIST 1.1 assessment, the images should meet specific minimal requirements for acquisition, such as slice thickness. Standardization is also necessary to ensure that images from the same patient are acquired under the same conditions. In a multi-site study, standardization between sites can be challenging, even though the evaluation using a chest abdomen pelvis used for size measurement is less of a problem than quantitative metabolic assessments. To address this issue, acquisition charters are provided to sites taking part in clinical trials to help standardize the acquisition process. The quality of the image can also be deteriorated by patient-specific factors such as contraindications that may prevent the injection of contrast agents or prosthetic implants that can cause image artifacts (Fig. 12), making a robust analysis difficult. Despite these challenges, it is the responsibility of the radiologist to provide an assessment based on the images provided, considering the level of trust they have in the images.

The second source of variability is related to contextual information that may or may not be accessible to the radiologist. During central analysis, radiologists are often blinded to clinical information and historical examinations, which can introduce uncertainty into radiologic findings and result in variability in the assessment. This lack of information can make the selection of liver findings at baseline challenging, as the radiologist may not have the knowledge of their evolutive character to relate them to metastasis in the absence of a typical benign appearance. In contrast, when assessing patients at the hospital, the radiologist may be influenced by contextual information, leading to an optimistic bias, particularly for response assessments [16, 37].

Conclusion

This pictorial provides a comprehensive overview of the variability in RECIST assessments, with two significant impacts: it affects study endpoints during drug development and influences clinical decisions in patient management. Discrepancies in the detection of NLs can lead to inconsistent assessments of disease progression, potentially influencing the overall outcome of the study. Moreover, the accuracy of assessments relies heavily on adequate additional clinical information, emphasizing the importance of considering the reader variability when making clinical decisions based on RECIST assessments. Reader variability is inherent in radiologic oncology assessment. RECIST 1.1 criteria help to control this variability by standardizing evaluation rules to produce comparable results between different studies. Being compliant, the radiologist's primary objective is to minimize the risk factors contributing to variability. During the process of double reading for central assessments, this variability is materialized into discordance rates which are informative of the origin of the reader's variability. The baseline and target selection plays a major role in the occurrence of discordance. The appearance of NLs is also a determining factor in terms of both detection and characterization discrepancies. It seems important to always reconsider the presence of an equivocal lesion at baseline or during follow-up in order not to distort the calculation of the initial lesion sum or to declare a progression too early. The evaluation of serial radiological examinations two by two may favor the confirmation bias while intermediate and global retrospective reviews help to remove this bias for a better understanding of the overall patient response. The interpretation of images can be influenced by contextual information, and the lack of such information can lead to uncertainty in the assessment. Conversely, the presence of specific information can result in a biased assessment.

Abbreviations

BICR	Blinded Independent Central Review
NL	New lesion
NTL	Non-target lesion
RECIST	Response Evaluation Criteria in Solid Tumor
TL	Target lesion

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13244-024-01774-w.

ELECTRONIC SUPPLEMENTARY MATERIAL

Authors contributions

All authors contributed equally to the final manuscript. A.I.: conceptualization, initial draft, editing. H.B.: editing, literature review. C.O.: editing, literature review. A.S.B.: initial draft, editing. Y.L.: editing, literature review. All authors read and approved the final manuscript.

Funding

No funding to declare.

Data availability

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Antoine Iannessi & Hubert Beaumont are part-time employees of Median Technologies. Christine Ojango & Yan Liu are full-time employees of Median Technologies. Anne-Sophie Bertrand declares having no competing interests.

Author details

[1]Cancer Center Antoine Lacassagne 33 Av. de Valombrose, 06100 Nice, France. [2]Median Technologies SA 1800 Route des Crêtes, 06560 Valbonne, France. [3]Imaging Center Beaulieu-sur-mer 18 Bd Eugène Gauthier, 06310 Beaulieu-sur-Mer, France.

Received: 28 March 2024 Accepted: 7 July 2024
Published online: 07 August 2024

References

1. (FDA) FaDA (2018) Clinical trial imaging endpoint process standards guidance for industry. In: Research CfDEaRCfBEa (ed.). FDA. 26 Apr 2018
2. Ellingson BM, Brown MS, Boxerman JL et al (2021) Radiographic read paradigms and the roles of the central imaging laboratory in neuro-oncology clinical trials. Neuro Oncol 23:189–198
3. Ford R, O'Neal M, Moskowitz S, Fraunberger J (2016) Adjudication rates between readers in Blinded Independent Central Review of Oncology Studies. J Clin Trials 6:289
4. Schmid AM, Raunig DL, Miller CG et al (2021) Radiologists and clinical trials: Part 1 the truth about reader disagreements. Ther Innov Regul Sci 55:1111–1121
5. Abramson RG, McGhee CR, Lakomkin N, Arteaga CL (2015) Pitfalls in RECIST Data extraction for clinical trials: beyond the basics. Acad Radiol 22:779–786
6. Blumenthal GM, Kluetz PG, Schneider J, Goldberg KB, McKee AE, Pazdur R (2017) Oncology drug approvals: evaluating endpoints and evidence in an era of breakthrough therapies. Oncologist 22:762–767
7. Eisenhauer EA, Therasse P, Bogaerts J et al (2009) New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 45:228–247
8. Fournier L, de Geus-Oei LF, Regge D et al (2021) Twenty years on: RECIST as a biomarker of response in solid tumours an EORTC imaging group—ESOI Joint Paper. Front Oncol 11:800547

9. Tovoli F, Renzulli M, Negrini G et al (2018) Inter-operator variability and source of errors in tumour response assessment for hepatocellular carcinoma treated with sorafenib. Eur Radiol 28:3611–3620
10. Keil S, Barabasch A, Dirrichs T et al (2014) Target lesion selection: an important factor causing variability of response classification in the response evaluation criteria for solid tumors 1.1. Invest Radiol 49:509–517
11. Zimmermann M, Kuhl CK, Engelke H, Bettermann G, Keil S (2021) CT-based whole-body tumor volumetry versus RECIST 1.1: feasibility and implications for inter-reader variability. Eur J Radiol 135:109514
12. Zimmermann M, Kuhl C, Engelke H, Bettermann G, Keil S (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? Pol J Radiol 86:e594–e600
13. Darkeh MH, Suzuki C, Torkzad MR (2009) The minimum number of target lesions that need to be measured to be representative of the total number of target lesions (according to RECIST). Br J Radiol 82:681–686
14. Iannessi A, Beaumont H, Liu Y, Bertrand AS (2021) RECIST 1.1 and lesion selection: How to deal with ambiguity at baseline? Insights Imaging 12:36
15. Krasovitsky M, Lee YC, Sim HW et al (2022) Interobserver and intraobserver variability of RECIST assessment in ovarian cancer. Int J Gynecol Cancer 32:656–661
16. Beaumont H, Evans TL, Klifa C et al (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 18:50
17. Schwartz LH, Bogaerts J, Ford R et al (2009) Evaluation of lymph nodes with RECIST 1.1. Eur J Cancer 45:261–267
18. Coleman RE, Mashiter G, Whitaker KB, Moss DW, Rubens RD, Fogelman I (1988) Bone scan flare predicts successful systemic therapy for bone metastases. J Nucl Med 29:1354–1359
19. Beaumont H, Iannessi A, Wang Y, Voyton CM, Cillario J, Liu Y (2021) Blinded Independent Central Review (BICR) in new therapeutic lung cancer trials. Cancers 13:4533
20. Busby LP, Courtier JL, Glastonbury CM (2018) Bias in radiology: the how and why of misses and misinterpretations. Radiographics 38:236–247
21. Williams L, Carrigan A, Auffermann W et al (2021) The invisible breast cancer: Experience does not protect against inattentional blindness to clinically relevant findings in radiology. Psychon Bull Rev 28:503–511
22. Berbaum KS, Franken EA Jr, Dorfman DD et al (1990) Satisfaction of search in diagnostic radiology. Invest Radiol 25:133–140
23. Watanabe H, Kunitoh H, Yamamoto S et al (2006) Effect of the introduction of minimum lesion size on interobserver reproducibility using RECIST guidelines in non-small cell lung cancer patients. Cancer Sci 97:214–218
24. Yoon SH, Kim KW, Goo JM, Kim DW, Hahn S (2016) Observer variability in RECIST-based tumour burden measurements: a meta-analysis. Eur J Cancer 53:5–15
25. Krajewski KM, Nishino M, Franchetti Y, Ramaiya NH, Van den Abbeele AD, Choueiri TK (2014) Intraobserver and interobserver variability in computed tomography size and attenuation measurements in patients with renal cell carcinoma receiving antiangiogenic therapy: implications for alternative response criteria. Cancer 120:711–721
26. Patil V, Noronha V, Joshi A et al (2013) Is there a limitation of RECIST criteria in prediction of pathological response, in head and neck cancers, to postinduction chemotherapy? ISRN Oncol 2013:259154
27. Richie M, Josephson SA (2018) Quantifying heuristic bias: anchoring, availability, and representativeness. Teach Learn Med 30:67–75
28. Humbert O, Chardin D (2020) Dissociated response in metastatic cancer: an atypical pattern brought into the spotlight with immunotherapy. Front Oncol 10:566297
29. Tazdait M, Mezquita L, Lahmar J et al (2018) Patterns of responses in metastatic NSCLC during PD-1 or PDL-1 inhibitor therapy: comparison of RECIST 1.1, irRECIST and iRECIST criteria. Eur J Cancer 88:38–47
30. Tuma RS (2006) Sometimes size doesn't matter: reevaluating RECIST and tumor response rate endpoints. J Natl Cancer Inst 98:1272–1274
31. Bellomi M, De Piano F, Ancona E et al (2017) Evaluation of inter-observer variability according to RECIST 1.1 and its influence on response classification in CT measurement of liver metastases. Eur J Radiol 95:96–101
32. Muenzel D, Engels HP, Bruegel M, Kehl V, Rummeny EJ, Metz S (2012) Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. Radio Oncol 46:8–18
33. Karmakar A, Kumtakar A, Sehgal H, Kumar S, Kalyanpur A (2019) Inter-observer variation in response evaluation criteria in solid tumors 1.1. Acad Radiol 26:489–501
34. Raunig DL, Schmid AM, Miller CG et al (2021) Radiologists and clinical trials: Part 2: practical statistical methods for understanding and monitoring independent reader performance. Ther Innov Regul Sci 55:1122–1138
35. Borradaile K, Ford R, O'Neal JM, Byrne K (2010) Analysis of the cause of discordance between two radiologists on the assessment of radiographic response and progression for subjects enrolled in breast cancer clinical trials employing Blinded Independent Central Review. J Clin Oncol 28:6044–6044
36. Henschke CI, Yankelevitz DF, Yip R et al (2016) Tumor volume measurement error using computed tomography imaging in a phase II clinical trial in lung cancer. J Med Imaging 3:035505
37. Dello Russo C, Cappoli N, Pilunni D, Navarra P (2021) Local investigators significantly overestimate overall response rates compared to Blinded Independent Central Reviews in phase 2 oncology trials. J Clin Pharm 61:810–819

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.