© Pol J Radiol 2021; 86: e594-e600
DOI: https://doi.org/10.5114/pjr.2021.111048

Received: 08.09.2020
Accepted: 23.11.2020
Published: 22.10.2021

POLISH JOURNAL of RADIOLOGY

http://www.polradiol.com

Original paper

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?

Markus Zimmermann[A,B,C,D,E,F], Christiane Kuhl[A,D,E], Hanna Engelke[B,F], Gerhard Bettermann[B,C], Sebastian Keil[A,B,D,E]

Department of Diagnostic and Interventional Radiology, University Hospital, RWTH Aachen University, Germany

Abstract

Purpose: Target lesion selection is known to be a major factor for inter-reader discordance in RECIST 1.1. The purpose of this study was to assess whether volumetric measurements of target lesions result in different response categorization, as opposed to standard unidimensional measurements, and to evaluate the impact on inter-reader agreement for response categorization when different readers select different sets of target lesions.

Material and methods: Fifty patients with measurable disease from solid tumours, in which 3 readers had blindly and independently selected different sets of target lesions and subsequently reached clinically significant discordant response categorizations (progressive disease [PD] vs. non-progressive disease [non-PD]) based on RECIST 1.1 analyses were included in this study. Additional volumetric measurements of all target lesions were performed by the same readers in a second read. Intra-reader agreement between standard unidimensional measurements (uRECIST) and volumetric measurements (vRECIST) was assessed using Cohen's κ statistics. Fleiss κ statistics was used to analyse the inter-reader agreement for uRECIST and vRECIST results.

Results: The 3 readers assigned the same response classifications based on uRECIST and vRECIST in 33/50 (66%), 42/50 patients (84%), and 44/50 patients (88%), respectively. Inter-reader agreement improved from 0% when using uRECIST to 36% when using vRECIST.

Conclusions: Volumetric measurement of target lesions may improve inter-reader variability for response assessment as opposed to standard unidimensional measurements. However, in about two-thirds of patients, readers disagreed regardless of the measurement method, indicating that a limited set of target lesions may not be sufficiently representative of the whole-body tumour burden.

Key words: RECIST 1.1, inter-reader variability, target lesion, volumetric measurement.

Introduction

The Response Evaluation Criteria in Solid Tumours (RECIST) criteria are the most commonly used guidelines for standardized response assessment in cancer patients undergoing systemic treatment. The current version, RECIST 1.1, was published in 2009 and allows radiologists to designate a maximum of 5 metastases per patient (maximum 2 metastases per organ) as target lesions, which are subsequently used as "surrogates" to evaluate the response to treatment in a patient [1].

Correspondence address:

Dr. Markus Zimmermann and Dr. Sebastian Keil, Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Pauwelsstraße 30, 52074 Aachen, Germany, e-mails: mzimmermann@ukaachen.de; skeil@ukaachen.de

Authors' contribution:

A Study design · **B** Data collection · **C** Statistical analysis · **D** Data interpretation · **E** Manuscript preparation · **F** Literature search · **G** Funds collection

Previous studies, however, have shown that the selection of target lesions significantly impacts the inter-reader agreement for RECIST analyses [2,3]. When readers choose the same set of target lesions, inter-reader agreement for response categorization according to RECIST is almost perfect, whereas when readers choose different target lesions, the inter-reader agreement for response categorization drops significantly. One of the most likely explanations for this observation is clonal diversification of different cancer metastases in a patient, which may lead to differences in treatment response of these metastases. Depending on which subset of metastases is selected as target lesions and used for response evaluation, the response categorization according to RECIST 1.1 may therefore vary.

However, one other possible explanation for the inter-reader variability could be that uni-dimensional measurements are insufficiently representative of the 3-dimensional tumour load. Several previous studies have demonstrated that response classification according to RECIST may differ significantly depending on whether unidimensional or volumetric measurements are used [4,5].

The purpose of this study was therefore to assess whether volumetric measurements of target lesions result in a different response classification compared to normal RECIST measurements and whether these volumetric measurements improve the inter-reader variability in a subset of patients in which 3 readers chose different target lesions for analysis and reached discordant RECIST results.

Material and methods

Approval for this retrospective study was granted by the institutional review board (EK 028/19).

Between July 2015 and July 2018, 355 patients with measurable disease from solid tumours, who underwent contrast-enhanced multi-detector computed tomography according to a standardized protocol at our institution, were included in a prospective multi-reader study. All study participants underwent a baseline and at least 1 follow-up CT examination, up to a maximum of 3 follow-up CTs. The CT examinations were interpreted according to RECIST 1.1 guidelines by 3 radiologists with 5-12 years of experience in oncological imaging.

Parts of this patient cohort were from a previously published study [3]. For the present study, all patients in which readers chose the same target lesions, in which they reached the same RECIST results, and those patients in which the readers disagreed in their RECIST results purely based on changes to non-target lesions were excluded (n = 288). Of the remaining 67 patients, in which all 3 readers had chosen different sets of target lesions and disagreed with their assessments regarding progressive versus non-progressive disease based purely on the analyses of target lesions, we randomly selected 50 patients to be included in the present study.

Image interpretation and response analysis with standard unidimensional measurements

The basic study design, including image interpretation by the 3 readers, has been described in detail elsewhere [3]. In brief, CT images of all patients were transferred to a workstation with dedicated oncology software (Mint-Lesion, 2.6.4; MintMedical, Heidelberg, Germany), where 3 radiologists with 5-12 years of experience in oncological imaging prospectively and independently read the scans. All readers were blinded to the study objectives to avoid biased readings. On the baseline examination the readers defined a maximum of 5 target lesions per patient (max of 2 per organ), as stipulated by the RECIST 1.1 guidelines, and strove to choose the largest measurable lesions that appeared "best reproducible" as recommended by Schwarz *et al.* [6]. The longest diameter of each target lesion was automatically measured by MintLesion, and the software also calculated the sum of diameters (SOD) of all target lesions after the read was finished. For the follow-up scans, readers re-identified their previously defined target lesions, and the longest diameter of each lesion was again measured automatically by the software. Readers also searched and evaluated non-target disease and other findings on the baseline and follow-up scans. Response classes were then automatically assigned by the software based on the change in SOD of the target lesions and based on other relevant findings (unequivocal progression of non-target lesions or occurrence of new target lesions) according to the RECIST guidelines.

Response analysis with volumetric measurements

For the present study, all readers performed a second read of selected patients and performed volumetric measurements of all their previously selected target lesions. Readers re-opened the examinations of the respective patients in MintMedical, reviewed which target lesions they had previously chosen for analyses, and used the volume measurement tool of the software to determine the volume of each target lesion on the baseline and all follow-up scans. All volume measurements were performed on axial slices with 1mm slice thickness in order to assure that the measurements were as accurate as possible. The "interpolated volume measurement tool" requires the user to draw region of interests (ROI) along the borders of a lesion on the uppermost slice, on the slice of the largest diameter, and the lowermost slice. The software then automatically interpolates the ROI on all slices in-between, which the user can then revise and correct as required. Once the lesion has been adequately defined by ROIs on all contiguous slices, the software automatically calculates the volume of the lesion.

After the volumes of all lesions in a patient were determined using the abovementioned method, and the software automatically calculated the sum of all target lesion volumes. Response classes were assigned based on the RECIST guidelines for volumetric measurements [7]: an increase in the sum of volumes of all target lesions of ≥ 73% was considered progressive disease, a decrease of the sum of volumes of ≥ 65% was classified as partial remission, complete disappearance of all target lesions was classified as complete remission, and everything that did not meet the aforementioned criteria was classified as stable disease.

Data analysis

First, the response assessment results using standard unidimensional RECIST measurements (uRECIST) were compared to the results with use of volumetric measurements (vRECIST) for each reader individually (intra-reader agreement). Classification of a patient as "responder" or "non-responder" is the main objective of a response assessment tool and usually directly impacts patient management; therefore, we dichotomized response class assignments into PD (progressive disease) vs. non-PD (stable disease, partial response, or complete response). This analysis was done on a patient-based level, meaning that agreement between uRECIST and vRECIST results in a patient by a specific reader was present when the reader assigned the same response class (PD vs. non-PD) based on both uRECIST as well as vRECIST over all follow-up occasions of a patient.

Secondly, we analysed whether using vRECIST improves inter-reader agreement. Again, this was done using dichotomized response classes on a patient-based level. Agreement was present when all readers agreed regarding the distinction of PD vs. non-PD over all follow-up occasions of a patient.

Statistical analysis

Cohen's κ statistics were used to assess the intra-reader agreement for response assessment based on uRECIST and vRECIST for each of the 3 readers individually. Fleiss κ statistics were performed to analyse the inter-reader agreement for response assessment results based on volumetric measurements (vRECIST). Kappa values (κ) were interpreted as follows according to Landis *et al.* [8] : < 0.2 as slight, 0.21-0.4 as fair, 0.41-0.6 as moderate, 0.61-0.8 as substantial, and 0.81-1.00 as almost perfect.

Continuous variables were summarized using proportions, mean, and standard deviation. Distributions were analysed using Clopper-Pearson 95% confidence intervals (95% CI). All statistical analyses were performed with SPSS statistical software (version 25; IBM, Armonk, NY, USA).

Table 1. Patient characteristics

Total number of patients included		*N* = 50
Mean age, years		60.9 ± 12.3
Male/Female, *n*		29/21
Mean number of target lesions per patient		2.6 ± 1.1
Mean baseline sum of target lesion diameter (mm)		78 ± 48
Mean baseline sum of volume of target lesions (mm^3)		76 ± 140
Type of treatment		
	Targeted cancer agent	9
	Conventional chemotherapy	41

Results

The 50 patients in this study had undergone a total of 147 CT examinations (baseline and up to three follow-up CTs). Reader 1 chose a total of 134 target lesions, reader 2 chose 125 target lesions, and reader 3 chose 129 target lesions. Further patient demographics are summarized in Table 1, and a summary of primary tumour types in the patient cohort is shown in Table 2.

Intra-reader agreement for response categorization using uRECIST and vRECIST

Reader 1 assigned the same response classifications based on uRECIST and vRECIST in 33/50 patients (66%). Reader 2 reached identical response results using uRECIST and vRECIST in 42/50 patients (84%). Reader 3 assigned the same response classifications in 44/50 patients (88%). Intra-reader agreement for reader 1 (κ = 0.298, $p = 0.035$) was only fair whereas for readers 2 (κ = 0.682, $p < 0.001$) and 3 (κ = 0.754, $p < 0.001$) it was substantial.

Inter-reader disagreement for response categorization using uRECIST and vRECIST

Readers disagreed in 100% of patients regarding the response classification (PD vs. non-PD) when using the standard unidimensional measurements of target lesions

Table 2. Types of primary tumours in the patient cohort (*N* = 50)

Type of primary tumour	*n*
Non-small cell lung cancer	10
Head and neck cancer	8
Breast cancer	5
Colorectal cancer	4
Prostate cancer	3
Malignant melanoma	3
Pancreatic cancer	2
Renal cell carcinoma	2
Small cell lung cancer	2
Oesophageal cancer	2
Other (e.g. sarcomas, cancer of unknown primary, neuroendocrine tumours)	9

according to the RECIST 1.1 guidelines. Using the volumetric measurements, readers disagreed in 32/50 patients (64%) and agreed in 18/50 patients (36%). The improved inter-reader agreement among the 3 readers in those 18 patients resulted from upgrading a "non-PD" response classification with the use of uRECIST to "PD" when using volumetric measurements of target lesions (vRECIST) in 7 patients (7/18: 39%). In the remaining 11/18 (61%) patients, there was a downgrading of a "PD" classification when using unidimensional measurements (uRECIST) to "non-PD" when using vRECIST. A sample case illustrating inter-reader disagreement for response categorization is shown in Figures 1 and 2.

Fleiss κ statistics for the inter-reader agreement on response classification using vRECIST showed only a slight agreement ($\kappa = 0.137$, $p = 0.094$), which was nevertheless obviously better than the inter-reader agreement for uRECIST ($\kappa = -0.281$, $p = 0.001$).

Discussion

The selection of different target lesions is a known factor for inter-reader disagreement for response assessment according to RECIST 1.1 guidelines [2,3]; however, the underlying reason for this observation remains to be identified. Unidimensional measurements, as stipulated by the current RECIST 1.1 guidelines, have been previously shown to lead to higher intra- [9] and inter-reader variability [10] compared to volumetric measurements. Therefore, we tried to assess in this study whether such 3-dimensional, volumetric measurements change the response categorization of individual readers and whether this results in improved between-reader agreement in patients in whom radiologists selected different sets of target lesions. We found that in 12-34% of patients, volumetric measurements resulted in different response categorization as responder or non-responder compared to when using unidimensional measurements. Accordingly, the between-reader agreement for response categorization (PD vs. non-PD) improved from 0% with unidimensional measurements to 36% with 3-dimensional measurements.

One possible explanation for the intra-individual variability of response categorization as responder or non-responder when using uni-dimensional or 3-dimensional measurements with the same set of target lesions could be differences in threshold values for response categorization. Progressive disease according to RECIST 1.1 guidelines requires an increase of the sum of diameters of all target lesions for > 20% and at least 5 mm in absolute values. Therefore, for patients in whom the absolute sum of target lesion diameters is small, very small size changes of just a few millimetres of these target lesions may push the SOD over the threshold to progressive disease, or the size changes may be just short of the threshold and lead to a classification as stable disease. This becomes even more problematic when considering that small lesions are known to have a lower reproducibility of unidimensional lesion measurements compared to large ones [11]. For volumetric measurements, an increase in the sum of volumes of all target lesions of ≥ 73% according to RECIST guidelines indicates progressive disease, and this threshold may be just different enough from the threshold for standard unidimensional measurements to yield different response categorizations when the change of the SOD is close to the thresholds. At least 1 previous study has made similar observations and reported discordant classification of response to treatment between RECIST and volumetric measurements in 10-21% of patients [9]. This study also observed that volumetric measurements have an intra-observer reproducibility and are thus superior to unidimensional measurements.

However, even with volumetric measurements, in two-thirds of patients in this study the readers still reached discordant response categorizations of individual patients as responder or non-responder. In these cases, the most likely explanation is that different metastatic lesions responded differently to treatment, and this heterogeneity of response is reflected by different response classifications depending on which metastases are selected as target lesions and used for response assessment. In other words, in these patients, a limited subset of metastases may not be representative of the whole body tumour burden. However, the RECIST 1.1 guidelines are based on this exact assumption – that a maximum of 5 lesions are indeed sufficient for response assessment of the whole body tumour load assumes – which means that RECIST may just not lead to valid response assessment results in a significant number of patients. Further studies will need to evaluate whether volumetric measurements of the whole body tumour load are actually necessary for accurate response assessment results or whether a smaller number – although not as small as in the current RECIST guidelines – also yields acceptable results.

There are several limitations of this study, most notably the small cohort of patients with different types of primary tumours. However, manual volumetric measurements of target lesions are very time-consuming; hence, limiting the number of patients was necessary. Furthermore, patients received different types of treatment, including targeted cancer agents in a minority of patients, which usually requires the use of iRECIST instead of RECIST 1.1 for response assessment. However, the use of iRECIST would not have changed the results of this study because the aim was to compare the inter-reader variability instead of determining the "ground truth" for response categorization for each patient, and it was impossible to diagnose a possible pseudoprogression for the readers because they were all equally blinded regarding the type of treatment a patient received. Finally, we did not include patients in which readers reached concordant response assessment results using uRECIST, in which volumetric measurements could also lead to individual changes

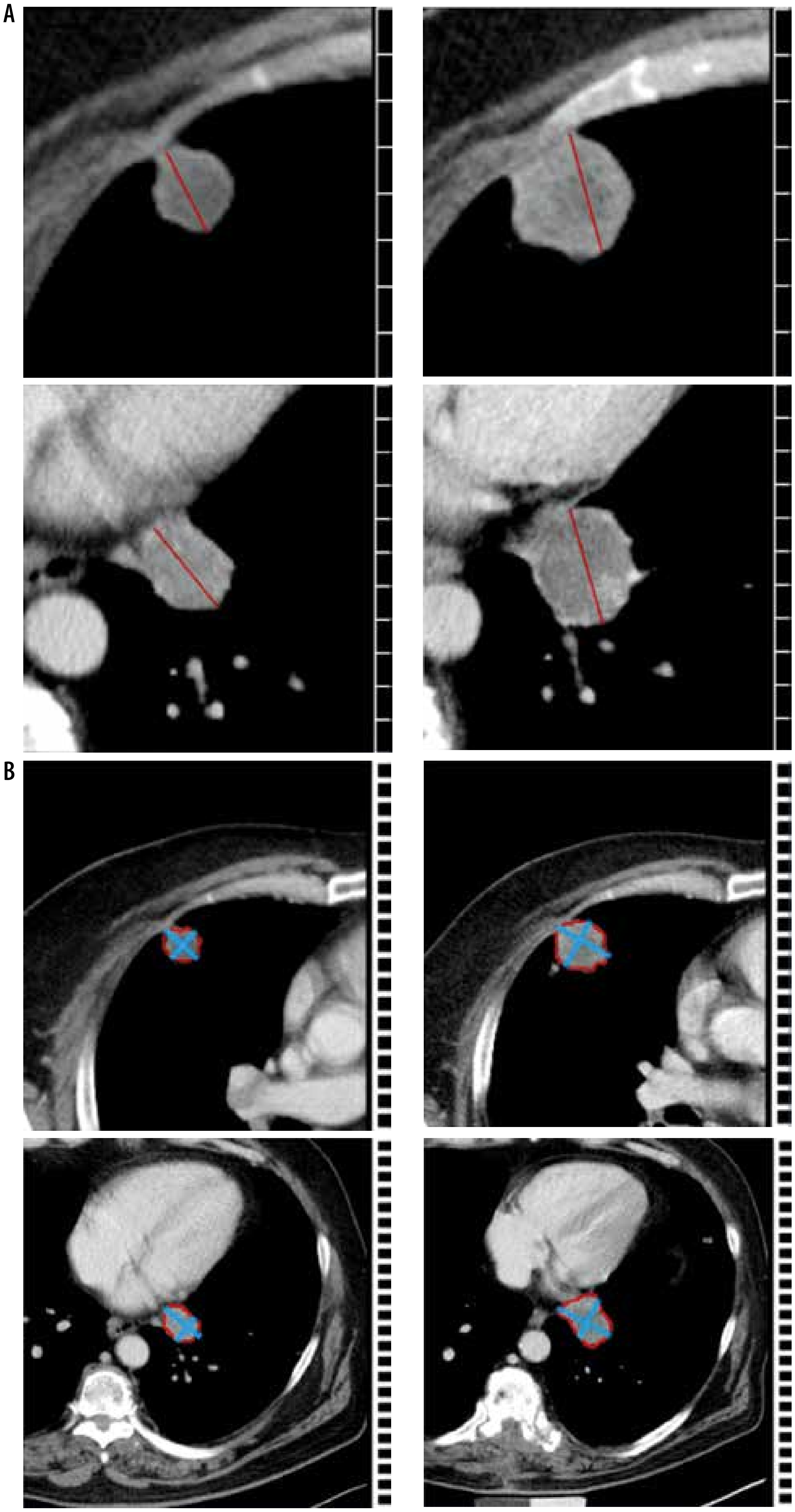

Figure 1. **A)** Reader 1 selected 2 pulmonary metastases as target lesions, which yielded a 25% growth of the target lesions based on conventional unidimensional RECIST measurements between baseline (left) and follow-up examination (right). **B)** Based on volumetric measurements, reader 1 found an increase of 175% of these 2 target lesions between baseline (left) and follow-up examination (right). Reader 1 therefore assigned progressive disease (PD) as response category for both types of measurements

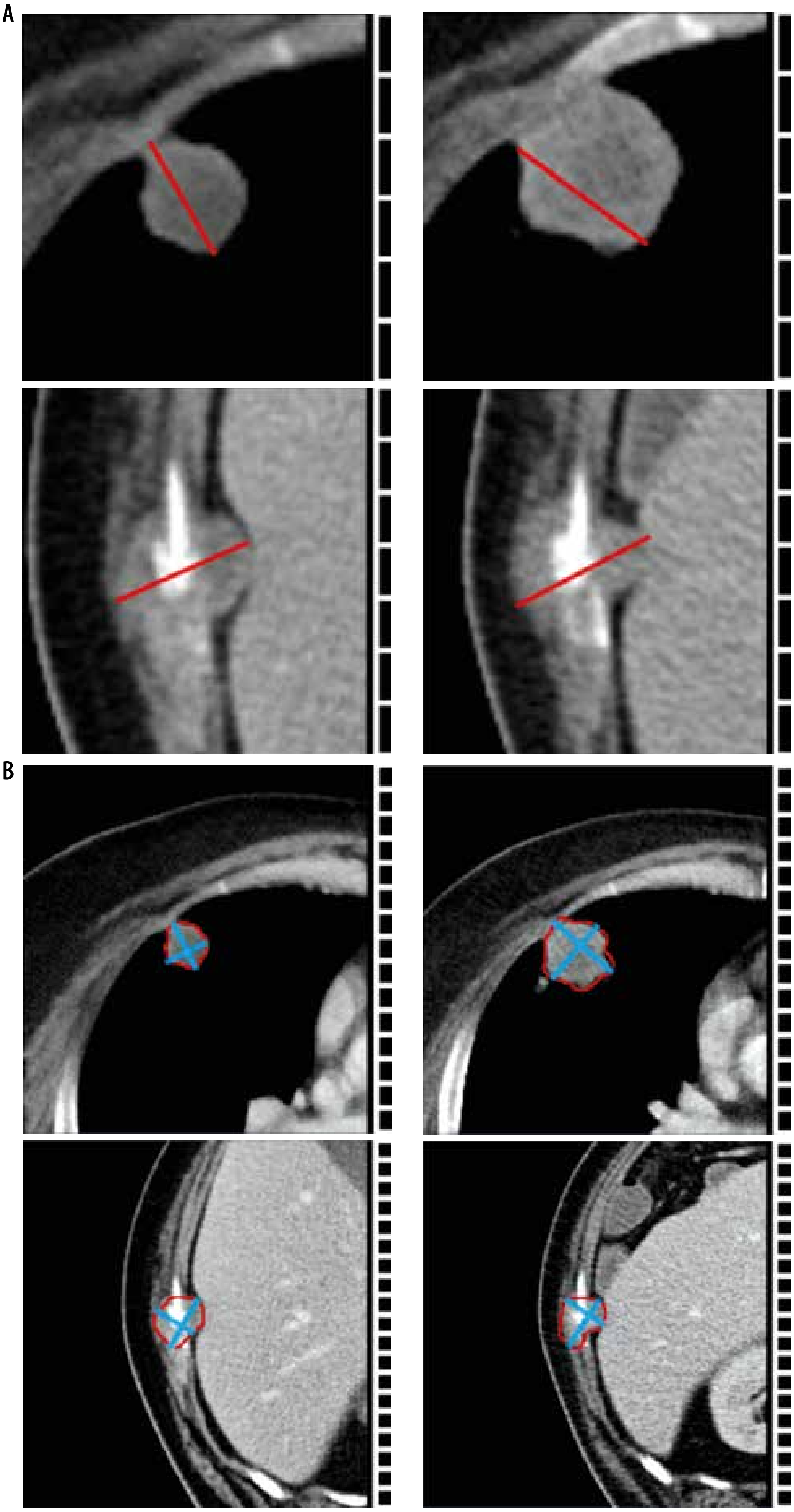

Figure 2. A) Reader 2 selected 1 pulmonary and 1 osseous metastasis as target lesions in the same patient, which yielded a 13% increase on the sum of the longest axial diameters based on unidimensional RECIST measurements between baseline (left) and follow-up examination (right). **B)** Based on volumetric RECIST measurements, a 49% increase in the sum of target lesion volumes was noted between baseline (left) and follow-up examination (right) by reader 2. This resulted in the assignment of SD as response classification for both types of measurements by reader 2

in response categorization and possibly increase the inter-reader variability.

Conclusions

Volumetric measurements may slightly improve the inter-reader variability compared to standard unidimensional measurements for response assessment according to RECIST 1.1 guidelines. However, in about two-thirds of patients, readers disagreed about response categorization as PD vs. non-PD regardless of the way of measurement, meaning that changing the way of measuring does not solve the problem. Most likely, different metastatic lesions may respond differently to treatment; therefore, a limited subset of a maximum of 5 metastases may not be sufficiently representative of the whole body tumour load.

Conflicts of interest

The authors report no conflict of interest.

References

1. 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.
2. 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.
3. 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.
4. Prasad SR, Jhaveri KS, Saini S, et al. CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques initial observations. Radiology 2002; 225: 416-419.
5. Tran LN, Brown MS, Goldin JG, et al. Comparison of treatment response classifications between unidimensional, bidimensional, and volumetric measurements of metastatic lung lesions on chest computed tomography. Acad Radiol 2004; 11: 1355-1360.
6. Schwartz LH, Litiere S, de Vries E, et al. RECIST 1.1-Update and clarification: from the RECIST committee. Eur J Cancer 2016; 62: 132-137.
7. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 2000; 92: 205-216.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.
9. 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: 1831-1839.
10. Marten K, Auer F, Schmidt S, et al. Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria. Eur Radiol 2006; 16: 781-790.
11. Van Hoe L, Van Cutsem E, Vergote I, et al. Size quantification of liver metastases in patients undergoing cancer treatment: reproducibility of one-, two-, and three-dimensional measurements determined with spiral CT. Radiology 1997; 202: 671-675.