Reduction in growth threshold for pulmonary metastases: an opportunity for volumetry and its impact on treatment decisions

^{1,2}M N VOGEL, MD, ¹S SCHMÜCKER, MD, ³O MAKSIMOVIC, MD, ⁴J HARTMANN, MD,
¹C D CLAUSSEN, MD and ¹M HORGER, MD

¹Department of Diagnostic and Interventional Radiology, Eberhard-Karls-University, Tübingen, Germany, ²Department of Neuroradiology, University of Heidelberg, Heidelberg, Germany, ³Department of Internal Medicine II, Eberhard-Karls-University, Tübingen, Germany, and ⁴Internistische Onkologie University Medical Center Schleswig-Holstein, Kiel Campus, Kiel, Germany

Objectives: This study compares tumour response assessment by automated CT volumetry and standard manual measurements regarding the impact on treatment decisions and patient outcome.

Methods: 58 consecutive patients with 203 pulmonary metastases undergoing baseline and follow-up multirow detector CT (MDCT) under chemotherapy were assessed for response to chemotherapy. Tumour burden of pulmonary target lesions was quantified in three ways: (1) following response evaluation criteria in solid tumours (RECIST); (2) following the volume equivalents of RECIST (*i.e.* with a threshold of $-65/+73\%$); and (3) using calculated limits for stable disease (SD). For volumetry, calculated limits had been set at $\pm 38\%$ prior to the study by repeated quantification of nodules scanned twice. Results were compared using non-weighted κ -values and were evaluated for their impact on treatment decisions and patient outcome.

Results: In 15 (17%) of the 58 patients, the results of response assessment were inconsistent with 1 of the 3 methods, which would have had an impact on treatment decisions in 8 (13%). Patient outcome regarding therapy response could be verified in 5 (33%) of the 15 patients with inconsistent measurement results and was consistent with both RECIST and volumetry in 1, with calculated limits in 3 and with none in 1. Diagnosis as to the overall response was consistent with RECIST in six patients, with volumetry in six and with calculated limits in eight cases. There is an impact of different methods for therapy response assessment on treatment decisions.

Conclusion: A reduction of threshold for SD to ± 30 – 40% of volume change seems reasonable when using volumetry.

Received 24 October 2010
Revised 19 February 2011
Accepted 4 May 2011

DOI: 10.1259/bjr/87835487

© 2012 The British Institute of
Radiology

Advances in CT technology, specifically volumetric data acquisition and image processing, permit volumetric tumour burden quantification. Volumetric measurement techniques have been shown to provide a more reliable and precise assessment of tumour size and evolution than linear measurements [1–3], and are increasingly integrated in image-processing software.

Accurate assessment of therapeutic tumour response is critical for the evaluation of chemotherapy results in patients enrolled on Phase II and III clinical trials, as well as in clinical routine. The response evaluation criteria in solid tumors (RECIST), including the recently revised RECIST version 1.1, are mainly designed for manual measurements and calculation of the sum of the longest target lesion diameters on axial planes [4, 5]. Although recent evidence suggests that the value of RECIST may be limited in various circumstances in which one-dimensional measurements carry high variability, they are widely considered the methodology

of choice for assessment of tumour response to treatment [6]. However, it could be shown that the recommended diameter thresholds (20% tumour growth; 30% size diminution) and their converted volume equivalents (73% and 65%, respectively) can be effectively reduced without sacrifice of reproducibility in order to permit a more precise or even earlier diagnosis of progressive disease (PD) or partial response (PR). The limits for stable disease (SD) can be narrowed and, as a consequence, diagnosis of SD, PD or PR can be made at an earlier stage [2]. This raises the question of whether it is “time” to move from one-dimensional assessment of tumour burden to volumetric assessment, or at least to standardise the latter. The revised RECIST guideline (version 1.1) does not recommend volumetry, nor are the limits for SD narrowed in volumetric response evaluation, which is mainly because there was not sufficient standardisation or widespread availability to recommend adoption of this alternative assessment method [5].

In clinical routine, however, and especially in critical cases (*e.g.* patients with SD suffering from side effects of chemotherapy), the question about any tendency to growth or shrinkage of tumour metastases is posed regularly by clinicians.

Address correspondence to: Dr Monika Vogel, Department of Neuroradiology, University of Heidelberg, INF 400, 69120 Heidelberg, Germany. E-mail: monika.vogel@med.uni-heidelberg.de

The results of several studies have shown that volumetric quantification would result in a different response assessment to that of one-dimensional and two-dimensional techniques [7, 8], but as there cannot possibly be a gold standard to verify real tumour growth or shrinkage under chemotherapy, more research is needed on this subject. The purpose of our study was to compare manual one-dimensional and automated volumetric techniques, including the reduction of growth thresholds for the evaluation of treatment response in patients, with pulmonary metastases of solid malignant tumours in a clinical routine setting, to evaluate the impact on treatment decisions and to correlate with patient outcome.

Methods and patients

Approval of the University Hospital Tübingen review board was obtained. 58 consecutive patients (36 males and 22 females; median age 55.7 years; range 20–93 years) with a total of 203 pulmonary metastases under chemotherapy were evaluated over a median follow-up interval of 3 months (range 77–98 days).

Pulmonary metastases originated from solid tumours, including colorectal cancer and sarcoma (each $n=15$), renal cell carcinoma ($n=5$), breast cancer, germ cell cancer, lung cancer (each $n=4$), urinary bladder and gastric cancer (each $n=2$), and seven others (each $n=1$). All individuals received chemotherapy according to various regimens between the initial and the follow-up multirow detector CT (MDCT) scans, which were initiated as part of routine drug response assessment.

CT technique

All CT examinations of the thorax were obtained as part of a standard protocol for tumour staging after the intravenous application of 120 ml of Ultravist 300 (Schering AG, Berlin, Germany). A 64-slice MDCT scanner (Sensation 64; Siemens Medical Solutions, Forchheim, Germany) was used. Scanning parameters were 120 kVp, 140 mAs, 0.6 mm collimation, 0.5 s^{-1} rotation. Scans were obtained during full inspiration. The table speed rotation was 6 mm and the rotation time was 0.75 s. Axial reconstructions were made with 3 mm slice thickness and overlapping slices (increment 2.5 mm).

Quantification

Tumour size was measured on digitised images in the lung window using both electronic callipers and segmentation software (OncoTREAT; MeVis, Bremen, Germany). The software ensured that the same lesion was evaluated on pre-treatment and follow-up scans [9]. For all patients, the five largest pulmonary nodules with a minimum diameter of 4 mm were chosen in consensus by two experienced radiologists. The post-therapy follow-up scans were evaluated without the opportunity to review data of the first measurements. A stopwatch was used to measure the time for both manual linear and automated volumetric measurements.

Linear quantification was made according to (modified) RECIST, as the minimum diameter of target lesions was 4 mm instead of 10 mm. Volumetric evaluation was made in two ways: (1) following a volume equivalent RECIST (*i.e.* PR is defined as a decrease in volume by 65% and PD is defined as an increase in volume by 73%); and (2) with a reduced threshold using calculated limits. For volumetry, calculated limits (limits for SD) at our institution had been set at a preliminary *in vivo* test with repeated volumetric and one-dimensional measurements. For this purpose a method similar to the one applied by Rampinelli et al [10] was used. Technical parameters were the same as in the main study. 101 nodules, which were located in the basal parts of 28 patients' lungs and were consecutively imaged twice on separate breath-holds during the thoracic and abdominal scans, were imaged and quantified. Calculated limits were set using the method published by Bland and Altman [11, 12] for measuring repeatability of methods of clinical measurement. This method calculates limits of agreement. As the true nodule volumes were unknown, the relative measurement error was calculated as the percentage difference of the two measurements compared with the mean of the two measurements. In the preliminary test, the mean of differences as a percentage of the mean values (MDV) and standard deviation were respectively 0.97 and 5.88 for linear measurements and 1.0 and 7.6 for volumetry. For linear measurements the limits calculated (*i.e.* limits for SD) were $-28.44\%/+30.34\%$ when 5 standard deviations were added to the MDV of linear measurements, which is close to the RECIST threshold. For this reason, in the main study, limits of volumetrically measured SD were defined by $\text{MDV} \pm 5$ standard deviations of the percentage difference in the mean value. Adding 5 standard deviations to the MDV of repeated volumetric measurements calculated limits for volumetry were $-37\%/+39\%$ volume change, rounded to a volume change threshold of $\pm 38\%$. These limits are also close to the optimum threshold for volumetry ($\pm 35\%$) calculated by Marten et al [2] and also resemble limits calculated by Rampinelli et al [10].

Diagnoses as to the overall response were based on the evolution of both tumour marker concentrations and CT morphology as quantified in the routine setting separated from the study and using manual linear measurements only. Additionally to pulmonary metastases, primary tumours and metastases as assessed in routine follow-up on other locations were taken into account if present.

Statistics

Differences of response were quantified using non-weighted κ -values and were evaluated for their impact on treatment decisions. Differences between the results of the two groups of patients (*i.e.* those with consistent and those with diverging measurement) were evaluated using analysis of variance *t*-test. Differences were regarded as significant for $p < 0.05$. Correlation with patient outcome and diagnosis as to the overall response was performed and was evaluated using descriptive statistics. All statistics were calculated with JMP IN 5.1 (SAS Institute, Cary, NC).

Results

Nodule characteristics and automated volumetry

The 203 metastases followed over the study period had a mean volume of 4.24 mm³ (range 0.10–73.40 mm³) and a mean maximum diameter of 16.06 mm (range 4.20–75.83 mm) at initial assessment, corresponding to a mean volume of 3.72 mm³ (range 0–86.06 mm³) and a mean maximum diameter of 15.59 mm (range 0–113.89 mm) at follow-up. Automated volumetry was accomplished in all nodules without the necessity of manual post-processing of the segmentation result. Both manual linear and automated volumetric measurements took an average of 8 s nodule⁻¹.

For the 58 study patients, a total number of 174 (*i.e.* 58 × 3) tumour response classifications were given in consensus by two readers. Four patients had complete response. Following RECIST (using the maximum diameter change threshold of −30%/+20%), 14 patients were classified as having PD, 38 patients as having SD and 2 patients as having PR.

Using volumetry (following RECIST equivalents with a volume change threshold of +73%/−65%), 12 patients were classified as having PD, 40 patients as having SD and 2 patients as having PR.

With the reduced volume threshold, the number of patients classified as having PD or PR increased while the number of patients with SD was reduced. As a result, at a volume change threshold of ±38% there were only 28 patients classified as having SD (reduction of 30%), 18 patients as having PD (increase of 50%) and 8 patients as having PR (increase of 300%; Figure 1).

In 15 (17%) of the 58 patients, results of response assessment were inconsistent with 1 of the 3 methods, which would have had an impact on treatment decisions in 8 (13%) of them.

Between the two groups of patients (those with consistent and those with diverging results) there were no significant differences in age, size or location of

pulmonary metastases (parapleural, paravascular or free in the lung parenchyma), quality of measurements and segmentation, degree of changes or number of nodules per patient.

Following RECIST, diverging results between manual and volumetric measurements were seen in six patients ($\kappa=0.79$) with a theoretical impact on treatment decisions in five of them. Following calculated limits, results differed from those of manual measurements in 13 patients ($\kappa=0.62$), with a theoretical impact on treatment decisions in 7; results differed from those of volumetry in 11 patients ($\kappa=0.67$), with a theoretical impact on treatment decisions in 4; and results differed from those of both manual measurements and volumetry in 9 patients, with a theoretical impact on treatment decisions in 3. Patients' outcome regarding therapy response could be retrospectively verified in 5 (33%) of the 15 patients with inconsistent measurement results and was consistent with both RECIST and volumetry in 1 patient, with calculated limits in 3 and with none in 1. Diagnosis as to the overall response was consistent with linear measurements in 7 (47%) of the 15 patients, with volumetry in 5 (33%) of the 15 patients and with calculated limits in 9 (53%) of the 15 patients (Table 1).

Manual and volumetric measurements were consistent as to the question of overall behaviour of pulmonary tumour burden (*i.e.* tumour growth, stable state, shrinkage or disappearance of metastases) in 52 (90%) of the 58 patients. However, the degree of changes varied, which led to differences in response evaluation, depending on the criteria, method or limits applied. Overall behaviour of tumour burden differed in seven patients classified as having SD by all three methods.

Impact on treatment decisions and patient outcome

Chemotherapy was continued in patients with SD and PR and was modified in patients with PD or in the case of side effects. Differences of response evaluation would have had an impact on treatment decisions in cases with

Figure 1. Evolution of pulmonary metastases as quantified with the different methods at follow-up. White, response evaluation criteria in solid tumours; black, volumetry (+73%/−65%); grey, volumetry with previously calculated reduced threshold (±38%). CR, complete response; PD, progressive disease; PR, partial response; SD, stable disease.

Table 1. Patients with discordant results of follow-up as quantified with the three different methods: correlation with the overall diagnosis and procedure at University Hospital Tübingen, theoretical impact on treatment decisions and outcome

Patient number	RECIST	Volumetry (+73%/–65)	Volumetry (+/–38%)	Diagnosis at our institution	Theoretical impact on treatment decisions	Procedure at our institution	Outcome
1	SD	SD	PD	PD	No	Changed (SE)	–
10	SD	SD	PD	PD	Yes	Changed	–
23	SD	SD	PR	PR	No	Continued	SD
24	SD	SD	PD	SD	Yes	Continued	PD
27	SD	SD	PR	SD	No	Changed (SE)	–
29	SD	SD	PR	PR	No	Continued	PD
41	SD	SD	PR	PR	No	Continued	PR
57	SD	SD	PD	PD	Yes	Continued (dosage elevation)	PD
58	SD	SD	PR	SD	No	Died	–
18	SD	PD	PD	PD	Yes	Changed	–
25	PD	SD	SD	PD	Yes	Changed	–
46	PD	SD	SD	PD	Yes	Changed	–
47	SD	PD	PD	PD	Yes	Changed	–
9	PD	SD	PD	PD	No	Changed	–
32	PD	SD	PD	PD	Yes	Changed	–

PD, progressive disease; PR, partial response; RECIST, response evaluation criteria in solid tumours; SD, stable disease; SE, side effects.

PD shown by one of the modalities in a patient without side effects.

Patient outcome as to the size evolution of pulmonary metastases was only possible to obtain in those for whom the same chemotherapy was continued until another CT scan of the thorax was done. This was the case for 5 (33%) of the 15 patients with discordant results. In 3 (60%) of these 5 patients, size evolution was the same as diagnosed before by using calculated limits (Figure 2).

Discussion

RECIST is considered the methodology of choice for the morphological assessment of tumour response [4, 5, 8]. RECIST attempts to compensate for measurement inaccuracies using high thresholds for diagnosis of PD or PR (+20% and –30% diameter change, respectively), but

it has been shown that volumetrically quantified changes in tumour size may have the potential to be an earlier or better marker of regression or progression. Such a marker, by enabling a lack of tumour response to be recognised earlier or more reliably than with one-dimensional measurements, could help to limit the amount of ineffective chemotherapy patients receive [13]. Moreover, both growth threshold and disagreement on tumour response can be substantially reduced by using automated volumetry [2, 14, 15].

In our study, volumetry had been “calibrated” by calculation of limits for SD using Bland and Altman’s approach [11, 12]. Limits for SD were calculated as the range in which intra-observer and interscan variability occurs with a probability of less than 0.1% (*i.e.* MDV ± 5 standard deviations). For the maximum diameter, this resulted in a threshold of –28.44/+30.34, which is

Figure 2. Positive correlation of the different methods and thresholds for response evaluation with overall tumour evolution (white) and true outcome of pulmonary metastases (black) in the 15 patients with discordant results. RECIST, response evaluation criteria in solid tumours.

relatively close to RECIST. For volumetry, the so calculated threshold was $-37\%/+39\%$ volume change (rounded to a volume change threshold of $\pm 38\%$), which is similar to the optimum threshold calculated by Marten et al [2]. These data also corroborate the results of a study by Rampinelli et al [10], which could show that a measured volume variation of $<30\%$ was in many cases exclusively due to interscan variability.

Fair to poor agreement of volumetric and one-dimensional measurement is reported in the literature [7, 13, 16]. The κ -value of 0.79 in our study corresponds well with the results published by Tran et al [7], who found a κ -value of 0.74 when comparing one-dimensional and volumetric response classifications, although they exclusively used a threshold of $+40\%/-65\%$ for volumetry. Differences in the volumetric and one-dimensional methods for response classification partly reflect the fact that tumours do not necessarily have a spherical shape or change in a spherical manner [13, 16, 17].

The results of a previous retrospective analysis have shown that the discrepancies between various measurement methods may theoretically influence treatment decisions in up to 51.4% of patients [8]. In our evaluation, not only the nodules but also the occurrence of side effects of chemotherapy were considered for the retrospective evaluation of the theoretical impact of measurement methods on therapeutic decisions. This can only partly explain the smaller percentage of possible treatment changes in our study, however, because we found discordant results in only 17% of patients and a possible impact on treatment decisions in 13%.

In our study, the treatment would possibly have been different with the application of volumetry instead of linear measurements in 5 (33%) of the 15 patients and with the application of calculated limits instead of RECIST or volumetry in 8 (53%) of the 15 patients with discordant results. In clinical routine, however, diagnoses were based on many factors, such as tumour markers, evolution of primary tumours and metastases on other sites. Treatment decisions were additionally dependent on the tolerance of chemotherapy. This overall tumour evolution was consistent with the results of calculated limits in 8 (53%) of the 15 patients, compared with 6 (40%) of 15 for each volumetric and linear RECIST measurements.

To our knowledge, there are no data in the literature about the correlation of RECIST, volumetry and calculated limits with patient outcome. Although overall disease behaviour could be retrospectively verified in all patients, the correlation of outcome and size evolution of pulmonary metastases was only possible in those in whom the same chemotherapy was continued until the patient received another CT scan of the thorax. This was the case in 5 (33%) out of 15 patients with discordant results. In 3 (60%) of these 5 patients, size evolution was the same as diagnosed before by using calculated limits.

The robustness of a decreased volume change threshold against increasing intra-individual and interobserver variance has been proven by the stability of observer confidence in data sets based on five, four and three metastases in the study by Marten et al [2].

In our patients, the reduction in the volume change threshold from $-65\%/+73\%$ to $\pm 38\%$ was reasonable because of the prior calibration and corroboration with the results of other independent working groups [2, 10].

The relatively small number of patients with discordant results ($n=15$), however, does not allow for reasonable significance testing. With a reduced threshold for volumetry, the diagnosis of PD, SD and PR correlated better with overall disease behaviour and patient outcome than the RECIST and volumetry values. In a retrospective *in vivo* setting there can be no gold standard for the verification of true tumour evolution, and, although postulated by many authors, correlation with patient outcome to retrospectively verify tumour evolution and to replace the missing gold standard is only possible (if at all) in a very small number of patients. Yet, our data suggest, in corroboration with the results of previous studies, that it seems reasonable to reduce the threshold of SD for up-to-date automated volumetry from $-65\%/+73\%$ to values between 30% and 40% volume change. The repeatability of excellent interscan and interobserver agreement with a reduced relative volume change, as found by independent working groups (using different applications) [2, 10, 15], suggests that our results can be extended to other software. However, a prior calibration of the particular application (either by the producer or by the user) seems sensible. In a study by Honda et al [15], the interobserver reproducibility of automated volumetry was shown to be superior to that of manual volumetry based on two-dimensional axial planes. A transfer of our results to manual multidimensional measurements with volume calculation would be possible only if time was an unlimited resource. To achieve the same precision in manual volume determination, a disproportionate amount of time would have to be invested not only in multidimensional image reformations, but also in multiple measurements and calculations. In our study, both manual linear and automated volumetric measurements took an average of 8 s per nodule, a speed that does not appear achievable in manual volume determination.

Our study has several limitations, the foremost of which are the inhomogeneous group of patients, the modest sample size, the retrospective setting and (last but not least) the difficulty to reasonably correlate the result of a morphological follow-up interval with patient outcome. In theory, a larger data sampling or prospective study design could permit more comprehensive testing of the hypothesis.

Conclusion

The impact of different quantification methods for therapy response assessment on treatment decisions should not be underestimated. The use of volumetry enables more accurate response assessments, and the reduction of growth threshold to 30–40% of volume change seems reasonable in many respects when using volumetry. This is especially important in patients suffering from side effects.

References

1. Heussel CP, Meier S, Wittelsberger S, Gotte H, Mildenerberger P, Kauczor HU. Follow-up CT measurement of liver malignoma according to RECIST and WHO vs. volumetry. [In German.] *Rofo* 2007;179:958–64.

2. Marten K, Auer F, Schmidt S, Rummeny E, Engelke C. Automated CT volumetry of pulmonary metastases: the effect of a reduced growth threshold and target lesion number on the reliability of therapy response assessment using RECIST criteria. *Eur Radiol* 2007;17:2561–71.
3. Vogel MN, Vonthein R, Schmücker S, Maksimovic O, Bethge W, Dicken V, et al. Automated pulmonary nodule volumetry with an optimized algorithm. Accuracy at different slice thicknesses compared to unidimensional and bidimensional measurements. [In German.] *Rofo* 2008;180:791–7.
4. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 2000;92:205–16.
5. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;45:228–47.
6. Therasse P, Eisenhauer EA, Verweij J. RECIST revisited: a review of validation studies on tumour assessment. *Eur J Cancer* 2006;42:1031–9.
7. Tran LN, Brown MS, Goldin JG, Yan X, Pais RC, Nitt-Gray MF, et al. Comparison of treatment response classifications between unidimensional, bidimensional, and volumetric measurements of metastatic lung lesions on chest computed tomography. *Acad Radiol* 2004;11:1355–60.
8. Pauls S, Kürschner C, Dharaiya E, Muche R, Schmidt SA, Krüger S, et al. Comparison of manual and automated size measurements of lung metastases on MDCT images: potential influence on therapeutic decisions. *Eur J Radiol* 2008;66:19–26.
9. Bornemann L, Kuhnigk JM, Dicken V, Zidowitz S, Kuemmerlen B, Krass S, et al. Informatics in radiology (infoRAD): new tools for computer assistance in thoracic CT part 2. Therapy monitoring of pulmonary metastases. *Radiographics* 2005;25:841–8.
10. Rampinelli C, De Fiori E, Raimondi S, Veronesi G, Bellomi M. In vivo repeatability of automated volume calculations of small pulmonary nodules with CT. *AJR Am J Roentgenol* 2009;192:1657–61.
11. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
12. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–60.
13. Zhao B, Schwartz LH, Moskowitz CS, Wang L, Ginsberg MS, Cooper CA, et al. Lung cancer: computerized quantification of tumor response: initial result. *Radiology* 2006;241:892–8.
14. Marten K, Auer F, Schmidt S, Kohl G, Rummeny E, Engelke C. Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria. *Eur Radiol* 2006;16:781–90.
15. Honda O, Kawai M, Gyobu T, Kawata Y, Johkoh T, Sekiguchi J, et al. Reproducibility of temporal volume change in CT of lung cancer: comparison of computer software and manual assessment. *Br J Radiol* 2009;82:742–7.
16. Prasad SR, Jhaveri KS, Saini S, Hahn PF, Halpern EF, Sumner JE. CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques initial observations. *Radiology* 2002;225:416–19.
17. Husband JE, Schwartz LH, Spencer J, Ollivier L, King DM, Johnson R, et al. Evaluation of the response to treatment of solid tumours—a consensus statement of the International Cancer Imaging Society. *Br J Cancer* 2004;90:2256–60.