

Katharina Marten
Florian Auer
Stefan Schmidt
Ernst J. Rummeny
Christoph Engelke

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

Received: 14 September 2006
Revised: 13 March 2007
Accepted: 22 March 2007
Published online: 10 May 2007
© Springer-Verlag 2007

K. Marten (✉) · F. Auer · S. Schmidt ·
E. J. Rummeny · C. Engelke
Department of Radiology,
University Hospital Wuerzburg,
Josef-Schneider-Str. 2,
97080 Wuerzburg, Germany
e-mail: katharina.marten@roe.med.
tu-muenchen.de

Abstract The purpose of this study was to evaluate the reproducibility of CT-volumetric tumour response assessment of pulmonary metastasis using variable volume change thresholds (VCT) and target lesions with response evaluation criteria in solid tumours (RECIST). Fifty consecutive patients with pulmonary metastases undergoing follow-up multislice CT under chemotherapy were assessed for response to chemotherapy with modifications to RECIST: (1) decreasing the percentual VCT for diagnosis of tumour response (range=70%–20%), (2) reducing the number of target lesions (range=1–5). Continuous and categorical observer agreements were tested by Bland and Altman and extended (κ_e) or non-weighted kappa (κ) and correlated with percentual VCT to predict observer agreement. A total of 202 metastases were evaluated (average volume=522.4 mm³±

902.4 mm³). General agreement on treatment response was very high ($\kappa_e=0.93$ –1), but was reduced with VCT<35% ($\kappa_e<0.95$). Kappa correlation with VCT values was strong ($r=0.94$ –0.96; $p\leq 0.0002$). Average confidence decreased significantly at VCT<45% ($p<0.01$) and agreement on stable disease at VCT<35% ($\kappa_e<0.95$; $p<0.01$). Reduction of target lesions ($n<3$; VCT=35%) resulted in decreased reader confidence (for $n=1$: $\kappa=0.49$; $p<0.05$). Agreement for evaluation of treatment response was robust using VCT \geq 35% and ≥ 3 metastases. This may translate into shortening of follow-up intervals or enable for response assessment with tumours displaying minimal volume change.

Keywords Lung · Nodule · Lung neoplasms · Computed tomography (CT) · Multi-detector row

Introduction

Accurate assessment of therapeutic tumour response is critical for evaluation of chemotherapy results in patients enrolled on phase I and II clinical trials. The “Response Evaluation Criteria in Solid Tumours” (RECIST) using unidimensional manual measurements of the sum of the longest marker lesion diameters [1] are widely considered as methodology of choice for assessment of tumour response to treatment [2]. However recent evidence suggests that the value of RECIST may be limited in various circumstances in which uni-dimensional measurements carry high variability. Moreover, in the evaluation of

targeted drugs that do not yield frank disease remission, the RECIST classification, while compensating for errors of manual measurements, is insensitive to such minimal tumour responses [3]. In a consensus paper the International Cancer Imaging Society advocated modifications to RECIST to compensate for such shortcomings and opted for incorporation of multislice CT scanning options and automated measurement approaches [4]. Fully automated volumetry of pulmonary nodules was shown to decrease observer variation with very high confidence in tumour response assessment using RECIST [5–8]. However it remains unknown if the recommended diameter thresholds (20% tumour growth; 30% size diminution) and their

converted volume equivalents (65% and 73%, respectively) can be effectively reduced without sacrifice of reproducibility in order to permit an earlier diagnosis of tumour progression or partial response. It is uncertain if the number of target lesions may be safely reduced to decrease the amount of time required for volumetric assessment.

This study in patients with pulmonary metastases was performed to longitudinally investigate: (1) the minimal tumour volume change allowing to maintain high reproducibility when used for classification of progressive disease or partial tumour response and (2) the minimum number of target lesions that is required to maintain the high reproducibility of this reduced volume change threshold.

Materials and methods

Patients and CT scans

This is a study evaluating the observer confidence for RECIST response assessment of pulmonary metastases in a prospective patient cohort with solid malignant tumours in the setting of a tertiary referral cancer centre. We included clinical routine multidetector-row CT (MDCT) scans of 50 consecutive patients (mean age: 59 years; range: 23–79 years; 39 men, 11 women) being evaluated for treatment response of pulmonary metastatic disease. Patient informed consent for the CT examinations was obtained in all cases, and the study was approved by our institutional review board. Intra- and inter-observer variability of manual calibre and volumetric tumour measurements in this patient collective has been reported previously [6].

All examinations were performed at our centre using a 16-row MDCT unit (Sensation 16, Siemens Medical Solutions, Forchheim, Germany). The scan parameters used were tube voltage of 120 kV, tube current of 80 mAs at 0.5/s gantry revolution, 0.75 mm slice collimation and table feed of 19.2 mm/rot (pitch=1.6). Single phase peripheral intravenous power injection was performed using 80 ml non-ionic contrast material of 300 mg/ml iodine concentration (Ultravist 300, Schering AG, Berlin, Germany). Image data were reconstructed using a lung filter kernel (B60f) at a slice thickness setting of 0.75 mm at a 0.7-mm reconstruction increment.

CT evaluation

CT image data were evaluated on a dedicated workstation in random order by two independent radiologists with 5 years of experience in clinical chest CT (FA, KM). Each radiologist was blinded to patient data and to the results of the other observer. Following RECIST criteria requirements, the five largest metastases were assessed as target lesions by automated volumetry using commercially

available software (LungCare, Somaris 5 VB10 AW. 4, Siemens Medical Solutions, Forchheim, Germany) [9].

The post-therapy follow-up scans were evaluated following the same protocol without opportunity to review data of the first measurements. On follow-up the treatment response of lung metastases was defined according to RECIST criteria as disease progression, stable disease, partial remission or total remission [1].

As a modification to traditional RECIST criteria, varied thresholds of volume change (hereinafter referred to as volume change thresholds) were used to define tumour response. For this purpose the diameter changes required for diagnosis of progressive disease and partial tumour response in the original RECIST guidelines were converted to tumour volumes [1, 6]. The evaluation started at a volume change threshold of 70% for diagnosis of progressive disease or partial response. Then the volume change threshold was reduced in steps of 5% down to 20% for calculation of categorical observer agreement. After the lowest acceptable volume change threshold for confident diagnosis of treatment response was identified, the number of included lesions was reduced one by one from five to one, each time eliminating the smallest metastasis, and continuous and categorical observer agreements were repeatedly assessed to define the smallest number of lesions allowing the maintenance of the response assessment result with five included lesions.

Data analysis

In the first analysis, categorical agreement was calculated using kappa statistics with the Fleiss-Nee-Landis extension that is applicable to data of more than two observers or ratings per pulmonary nodule [10, 11]. With this method individual kappa values for each specific tumour response category can be calculated using the Landis-Koch extension [12]. Agreement analysis was followed by linear and polynomial regression analyses to identify various fitted models for prediction of confidence over the entire range of volume change thresholds. The best-fitting regression function (R) was defined by additional regression-embedded analysis of variance-which calculates the variance ratio F , the partial correlation coefficient and r square (R^2) and adjusted r square (Ra^2) values from the regression model, thereby defining the predictive value of the model while compensating for the fact that R^2 is bound to increase with the number of predictors in the model-and was used for prediction of kappa values within 95% confidence intervals. Further, comparison was made between the average predicted agreement of all response categories reflecting the overall observer confidence, and the predicted agreement on specific categories of treatment response, because the latter give more precise information on how an individual patient is classified. Agreement on a specific response category may be significantly decreased

against overall observer confidence and could therefore substantially deteriorate the confidence in individual patients of that response group. Subsequently, the lowest volume change threshold maintaining the overall observer agreement and response category-specific agreement within 95% confidence intervals was identified. This value was used as a basis for subsequent agreement analysis with reduction of the target lesion number.

In the second analysis, individual percentual metastatic volume changes of the sum of all target lesions and the average variance pertaining to all tumour response ratings were calculated for both observers in the five datasets based on five, four, three, two or one metastasis. Continuous inter-observer agreement on individual metastatic volume change was tested using the Bland and Altman method and was compared between datasets including the five, four, three, two and one target lesions [13]. To assess the clinical relevance of inter-observer measurement inaccuracy, the 95% limits of agreement were converted to tumour doubling times on the basis of the average metastasis volume in this study and a standard follow-up interval of 90 days. The regression-predicted standard error of relative volume change comparing the two observers and pairs of the five datasets was calculated. Categorical agreement on treatment response was tested using non-weighted kappa statistics. Because Cohen's kappa statistics are likely to fall into error with small numbers ($\kappa < 0.5$) additional reader disagreement was estimated by using the marginal homogeneity method of Maxwell [14].

Our statistical software (StatsDirect release 2.4.4, CamCode, Herts, England) calculates Bland and Altman statistics with a one-way random effects intra-class correlation coefficient, estimated within-subjects standard deviation and a repeatability coefficient. The intra-class correlation coefficient is calculated as: $r_I = (mSS_b - SS_t) / ((m - 1)SS_t)$ -where m is the number of observations per subject, SS_b is the sum of squared between subjects and SS_t is the total sum of squares (as per one way ANOVA). Within-subjects standard deviation is estimated as the square root of the residual mean square from one way ANOVA. The repeatability coefficient is calculated as: $C_r = (\text{square root of } m)Z \xi_w$ -where m is the number of observations per subject, Z is a quantile from the standard normal distribution (usually taken as the 5% two-tailed quantile of 1.96), and ξ_w is the estimated within-subjects standard deviation (calculated as above). A P-value < 0.05 was accepted to indicate statistical significance.

Results

Patients

In the 50 patients a total of 202 lung metastases (median=5 nodules per patient, range 1–5) were evaluated over a

median follow-up interval of 2 months (range 1–9 months). Pulmonary metastases originated from solid tumours including oesophageal cancer (n=13), non small cell lung cancer (n=10), colorectal cancer (n=7), malignant melanoma (n=3), transitional cell carcinoma of the bladder (n=3), chondrosarcoma (n=2), renal cell carcinoma (n=2), and one patient each with breast cancer, malignant histiocytoma, stomach carcinoma, pharyngeal carcinoma, leiomyosarcoma, chorion carcinoma, sarcoma of the pleura, osteosarcoma, synovial sarcoma and carcinoma of unknown primary. All individuals received chemotherapy according to various regimens between the initial and the follow-up MDCT scans, which were initiated as part of routine drug response assessment after the median follow-up interval.

Nodule characteristics and automated volumetry

The 202 metastases followed over the study period had a median volume of 182.22 mm³ (volume range=3.16–5,195.13 mm³; converted median diameter=9.08 mm) at initial assessment corresponding to a median volume of 124.79 mm³ (range: 0–6,359.23 mm³) at follow-up. Automated volumetry was accomplished in all nodules. Manual correction of automated nodule volume segmentation (resizing of nodule templates) was performed in 7 solid nodules (1.75%) on first or second assessments (7/404 measurements=1.75%) and was limited to cases with obvious over-segmentation into neighbouring structures, such as pleura or vessels. Intra- and interobserver variation of automated volumetry ranged from –17.9 to 13.2% and –11.1 to 4.8% for readers 1 and 2, respectively, and from –16.9 to 13.2% and –21.3 to 3.9% for readings 1 and 2, respectively.

Tumour response classification using varied volume change thresholds

For the 50 study patients, a total number of 200 tumour response classifications were given (2 readers, 2 readings for each reader). Using a volume change threshold of 70%, 13 patients were classified as having progressive disease, 28 patients as having stable disease, 7 patients as having a partial response and 2 as having total response (Table 1). With decreasing volume threshold the number of patients classified as having progressive disease or partial response increased substantially, while the number of patients with stable disease was reduced. As a result, at a volume change threshold of 20% there were only 9 patients classified as having stable disease (reduction of 67.9%), 21 patients as having progressive disease (increase of 61.5%) and 18 as having partial response (increase of 157.1%). The number of patients with complete remission maintained stable over the entire threshold range ($\kappa_c = 1$;

Table 1 Patient treatment responses using modified volume change thresholds

Volume change threshold [%]	Response reviewer 1	Response reviewer 2			
		PD	SD	PR	CR
70	PD	13	–	–	–
	SD	–	28	–	–
	PR	–	–	7	–
	CR	–	–	–	2
65	PD	14	–	–	–
	SD	–	27	–	–
	PR	–	–	7	–
	CR	–	–	–	2
60	PD	14	–	–	–
	SD	–	27	–	–
	PR	–	–	7	–
	CR	–	–	–	2
55	PD	14	–	–	–
	SD	–	25	1	–
	PR	–	–	8	–
	CR	–	–	–	2
50	PD	15*	1*	–	–
	SD	–	22*	–	–
	PR	–	–	10	–
	CR	–	–	–	2
45	PD	16*	–	–	–
	SD	1*	21*	–	–
	PR	–	–	10	–
	CR	–	–	–	2
40	PD	16*	–	–	–
	SD	1*	18*	–	–
	PR	–	1*	12*	–
	CR	–	–	–	2
35	PD	16*	–	–	–
	SD	1*	17*	1*	–
	PR	–	–	13*	–
	CR	–	–	–	2
30	PD	16*	1*	–	–
	SD	1*	14*	1*	–
	PR	–	1*	14*	–
	CR	–	–	–	2
25	PD	18*	–	–	–
	SD	1*	11*	1*	–
	PR	–	–	17*	–
	CR	–	–	–	2
20	PD	20*	–	–	–
	SD	1*	9*	1*	–
	PR	–	–	17*	–
	CR	–	–	–	2

RECIST guidelines with varying volume thresholds for response category change (20–70%)
 PD=progressive disease;
 SD=stable disease; PR=partial remission
 CR=complete remission;
 *intraobserver response discordance

all classifications concordant). When using 65% or 60% volume change thresholds, all response classifications were identical between the reviewers and readings ($\kappa_e=1$). However there was growing discordance between the observers/readings using lower volume change thresholds (Table 1): at volume change thresholds between 55% and 45%, there were classification discordances in one patient, at thresholds between 40% and 20% in two patients and at a threshold of 30% in four patients. The overall agreement using Fleiss-Nee-Landis kappa extension as well as the agreements on specific response categories “progressive disease”, “stable disease”, “partial response” and “complete remission” are listed in Table 2, supplemented by the “minimal agreement” (lowest kappa value) of any of the specific response categories. All agreement categories except the agreement on total response showed a decrease of confidence with falling volume change thresholds. The overall agreement, the agreement on stable disease and the minimal agreement of the specific response categories showed the most severe deterioration with decreasing volume change threshold: for low thresholds between 30%–20% these agreements dropped to values below 0.95 ($\kappa_e=0.93, 0.89$ and 0.89 for combined categories, “stable disease” and “lowest kappa” categories, respectively).

Regression analyses achieved best results when using second degree polynomial fitting (Fig. 1): The linear correlation coefficients were smaller than 0.9 [$R^2=86.91\%$, $R^2=87.47\%$ and $R^2=84.86\%$ for the regression functions R1 (“combined categories”), R2 (“stable disease”) and R3 (“lowest kappa”), respectively]. With polynomial regression the correlation coefficients of the corresponding functions R1-R3 were for R1: $r=0.94$ ($R^2=87.61\%$; $Ra^2=84.52\%$, $F=28.30$; $P=0.0002$), for R2: $r=0.96$ ($R^2=91.99\%$; $Ra^2=89.98\%$; $F=45.91$; $P<0.0001$) and for

R3: $r=0.94$ ($R^2=87.92\%$; $Ra^2=84.90\%$; $F=29.11$; $P=0.0002$), respectively.

The variance ratios, F, for the overall second degree polynomial regression were highly significant: thus we have very little reason to doubt that agreement on either response category is associated with volume change threshold. The corrected r square values (Ra^2) show that over 84% of the variance of agreement is accounted for by the polynomial regression. The intercepts were highly significant for all response categories in the polynomial regression model ($P<0.0001$ for all functions R1-R3).

Therefore the predictive value of this model is slightly higher than that of the linear regression model, and the linear regression model was dropped for prediction of reader confidence. The polynomial regression functions R1-R3 are expressed by the following equations:

Average agreement (R1):

$$\kappa_e = 0.89 + 2.305E - 3(\text{volume change threshold}) - 0.9E - 5(\text{volume change threshold})^2$$

Agreement on stable disease (R2):

$$\kappa_e = 0.76 + 6.35E - 3(\text{volume change threshold}) - 4.2E - 5(\text{volume change threshold})^2$$

Lowest kappa of any response category (R3):

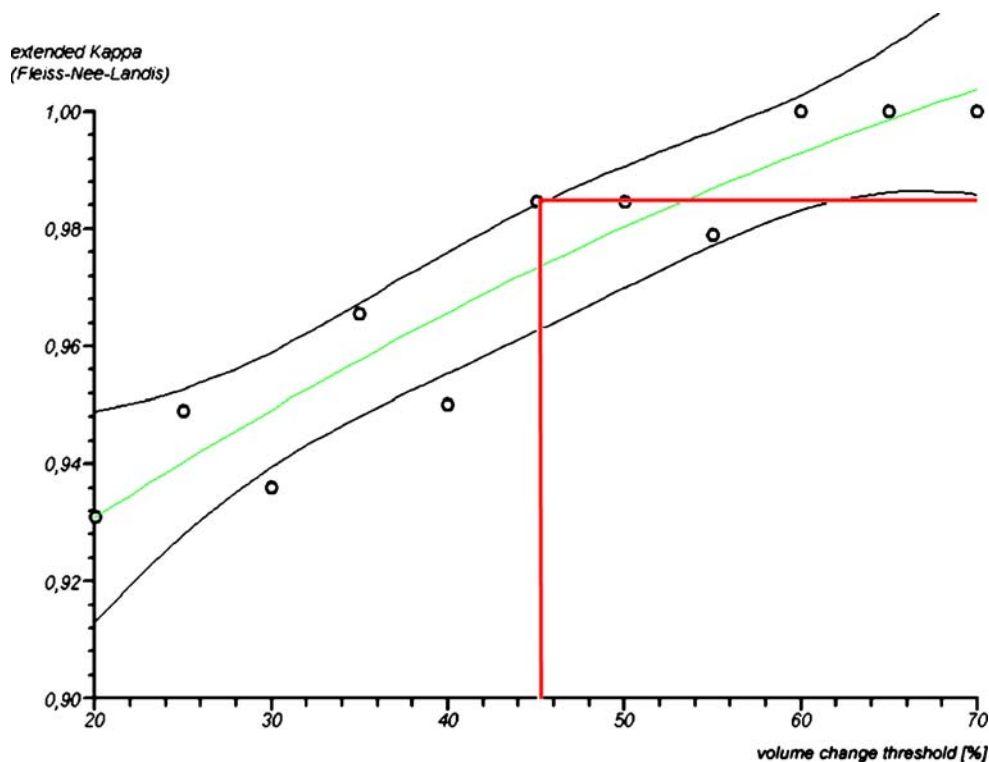
$$\kappa_e = 0.78 + 5.55E - 3(\text{volume change threshold}) - 3.4E - 5(\text{volume change threshold})^2$$

Table 2 Categorical agreement on response using modified volume change thresholds

Volume change threshold [%]	Fleiss-Nee-Landis extended kappa		Landis-Koch extended kappa per response category					SE
	Combined categories	SE	PD	SD	PR	CR	Lowest κ of any response category	
70	1	0.040	1	1	1	1	1	0.058
65	1	0.040	1	1	1	1	1	0.058
60	1	0.040	1	1	1	1	1	0.058
55	0.98	0.040	1	0.97	0.95	1	0.95	0.058
50	0.98	0.039	0.98	0.98	1	1	0.98	0.058
45	0.98	0.039	0.98	0.98	1	1	0.98	0.058
40	0.95	0.039	0.98	0.95	0.95	1	0.95	0.058
35	0.97	0.039	0.98	0.95	0.97	1	0.95	0.058
30	0.94	0.038	0.94	0.89	0.95	1	0.89	0.058
25	0.95	0.039	0.97	0.89	0.98	1	0.89	0.058
20	0.93	0.038	0.97	0.89	0.95	1	0.89	0.058

RECIST guidelines with varying volume thresholds for response category change (20–70%)
 PD=progressive disease; SD=stable disease; PR=partial remission; CR=complete remission
 SE=standard error

Fig. 1 Polynomial regression of extended observer agreement (κ_e) on average tumour response with 95% confidence intervals. Note the significant drop of κ_e values for modification of RECIST criteria using volume change thresholds (VCT) smaller than 45% (solid indicator lines) ($p < 0.05$). Agreement on average tumour response remained excellent ($\kappa \geq 0.95$) for VCT larger or equal 35%



The predicted observer confidence using the fitted polynomial regression curves differed significantly over the volume change threshold range for each of the three indicated agreement categories ($p < 0.01$). The deterioration of average agreement became significant at a volume change threshold of 45%. However the overall agreement at this cut-off was still high ($\kappa_e > 0.95$; Fig. 1). With reduction of the volume change threshold below 35% the predicted agreement in the stable disease category—and of the lowest kappa of all response categories—fell significantly below average agreement ($p < 0.05$), indicating the influence of discordance on patients within this category (Fig. 2). For achievement of excellent predicted observer confidence ($\kappa_e \geq 0.95$) in the “average agreement” (R1), “stable disease” (R2) and “lowest kappa” (R3) categories, the volume change thresholds were $VCT \geq 30.4\%$, 39.8% and 41.1% , respectively. For a defined volume threshold of 35% the estimated agreements were for R1: $\kappa_e = 0.96$ (95% CI=0.95–0.97), for R2: $\kappa_e = 0.94$ (95% CI=0.92–0.95) and for R3: $\kappa_e = 0.93$ (95% CI=0.92–0.95). Because excellent agreement without significant additional deterioration on specific response categories could be achieved at a volume change threshold equal to or larger than 35%, this threshold value was chosen as the basis for subsequent evaluation of reduced numbers of target lesions. With use of this modification to RECIST, 9 of 50 patients (18%) who had had stable disease with application of 70% volume change threshold were concordantly categorized from stable to progressive disease ($n=3$) or to partial response ($n=6$).

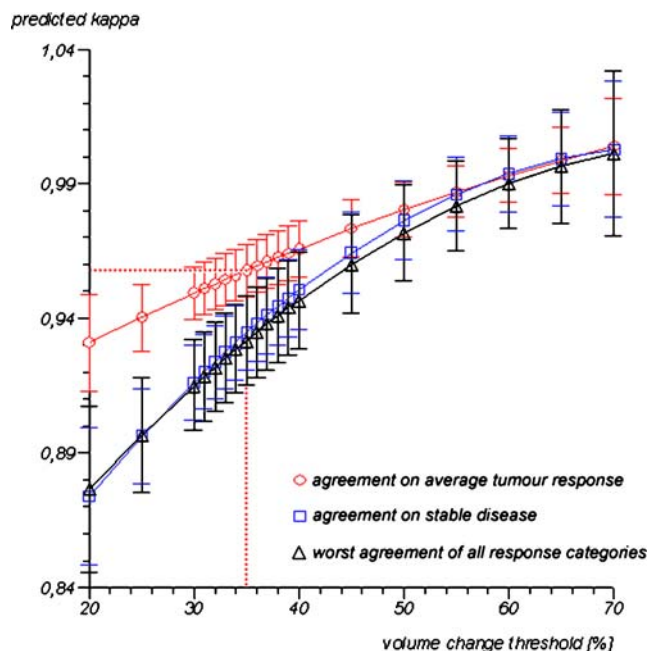


Fig. 2 Predicted extended observer agreement (κ_e) using polynomial regression for average tumour response (circles), stable disease (squares) and worst agreement of all response categories (triangles) with 95% CI interval bars. Note the significant decline of confidence with reduction volume change thresholds ($p < 0.01$). For VCT values smaller than 35% (dotted indicator lines) this deterioration is significantly worse in the “stable disease” and “worst agreement” categories than for average tumour response ($p < 0.05$)

Tumour response classification using variable numbers of target lesions

Of the 50 patients included in this study a subgroup of 33 patients (mean age: 66 years; range: 43 to 79 years; 8 women, 25 men) presented with five or more pulmonary metastases. Data of this patient subgroup form the basis of the tumour response assessment using a predefined volume change threshold of 35% and consisted of 132 response classifications of the two observers. The descriptive statistics of the resulting five tumour response datasets based on five, four, three, two and one target lesions showed growing average variance (increase of 66%) in percentual volume change with falling nodule numbers ($p > 0.05$; Table 3). Consequently, the continuous interobserver agreement was best for the dataset based on five nodules and poorest in the dataset based on one nodule. The limits of agreement increased almost threefold with reduction of the number of target lesions from five to one (95% limits of agreement = -12.84 – 11.28 mm³ and -33.07 – 29.53 mm³ for five and one nodule, respectively, Table 4). When in the dataset pertaining to one target lesion the limits of agreement were converted to volume doubling times, these figures were within the range of malignant growth (VDT = 311.5 – 415.4 days), indicating a clinical relevance of measurement inaccuracy. In the datasets pertaining to five, four and three lesions the volume doubling times were within the range of benign growth (VDT = 693.7 – $1,038.6$ days). When the dataset based on five target lesions was accepted as best reference standard, comparison with datasets based on reduced target lesions revealed an increasing average measurement error (predicted measurement error $e = 7.52$ – 7.8% for four target lesions, and $e = 49$ – 52.1% and 48.7 – 49.9% for datasets with one and two target lesions, respectively), which more than three times exceeded the inter-observer measurement error (Table 5).

As a result, when including five, four or three lesions for assessment of tumour response, all response classifications were identical between the two observers ($\kappa = 1$; Tables 6

and 7). In these datasets there were 9 patients classified as having progressive disease, 14 patients as having stable disease and 10 patients as having partial response. When the number of included metastases was further reduced, there were increasing classification discordances, which affected three patients (9.1%) in the dataset based on two lesions (one under-classification, two over-classifications) and ten patients (30.3%) in the dataset based on one lesion (seven under-classifications, three over-classifications; Table 6). Due to these discordances, the intraobserver agreement dropped substantially to kappa values smaller than 0.87 for response data derived from two lesions and significantly to values smaller than 0.5 for response data derived from only one lesion ($\kappa = 0.86$ and 0.49 , respectively, $p < 0.05$) (Table 5). The marginal disagreement was significant when comparing response between datasets with five and one metastasis ($p = 0.033$), indicating a particularly low reliability of the assessment using only one target lesion.

Discussion

The RECIST criteria are considered as the methodology of choice for the morphological assessment of tumour response [1, 2]. Despite the fact that RECIST attempts to compensate measurement inaccuracies by inclusion of five target lesions per organ and relatively high growth thresholds for diagnosis of progressive disease or partial response (+20% and -30% diameter change, respectively), various studies have demonstrated discordant response rates between primary investigators and secondary review panels [15–17]. This disagreement was chiefly caused by manual measurement inaccuracy [18–20] accounting for 45% of misclassifications [15].

Disagreement on tumour response can be substantially decreased by use of fully automated volumetry [6]. Published data confirming high accuracy (within 3% of 3-mm-diameter nodules) of fully automated volumetry [17,

Table 3 Percentual metastatic volume change per patient with varying number of included metastases

Lesions included	Observer	Minimum	Median	Mean	Maximum	Average variance
5	1	-99.50	11.73	27.48	595.35	16,039.62
5	2	-99.50	11.17	28.04	595.36	16,443.26
4	1	-99.43	9.48	27.58	633.58	17,110.58
4	2	-99.43	8.85	28.23	633.59	17,499.85
3	1	-99.30	5.52	23.27	633.58	15,996.90
3	2	-99.30	4.81	23.88	633.59	16,121.91
2	1	-100	-19.25	20.78	748.46	21,337.66
2	2	-100	-19.25	21.79	748.14	21,476.97
1	1	-100	-17.90	18.55	828.33	26,507.97
1	2	-100	-17.90	20.10	828.33	26,731.48

Table 4 Interobserver volumetric agreement with varying number of included metastases

Lesions included	95% LoA [mm ³]	ICC coeff.	ws-STD [mm ³]	Repeat. coeff. for $\alpha=0.05$	VDT conversion of lower 95% LoA	VDT conversion of upper 95% LoA
5	-12.84–11.28	0.999	4.32	11.97	853.75	1,038.63
4	-14.16–12.39	0.998	4.76	13.19	771.18	948.32
3	-15.67–13.88	0.998	5.29	14.66	693.77	849.79
2	-21.34–18.76	0.998	7.18	19.90	500.85	636.62
1	-33.07–29.53	0.995	11.19	31.02	311.51	415.36

LoA=limits of agreement (Bland and Altman); ICC-coeff=intra-class correlation coefficient
 ws-STD=within subject standard deviation; repeat. coeff.=repeatability coefficient
 VDT=volume doubling time

21] and an overall in vivo standard measurement error of 0.21–1.5% [5, 22] prompted us to assess its potential with use of modified volume change thresholds.

Our findings demonstrate a very high observer concordance and allowed halving the volume change threshold to 35% (converted original RECIST recommendation) at an average predicted observer confidence of $\kappa \geq 0.95$.

However pulmonary metastases growth may be inhomogeneous. In a retrospective cohort of 21 patients with multiple metastases (n=408) Chojniak et al. described a high intra-individual variability of volume doubling times with a tendency of smaller nodules to exhibit more rapid growth [23]. In a second study on 33 patients the authors found divergent intra-individual tumour responses in 35% of metastases after chemotherapy compared to the overall tumour response per patient, suggesting a possible influence of the choice of nodules on the assessment result. Similarly, Schwartz et al. described a reduction of intra-individual standardized average response variance by 90%

Table 5 Standard error of percentual metastatic volume change per patient

Comparison lesions included	Comparison observers	Predicted SE [%]
5	1 vs. 2	5.49
4	1 vs. 2	6.09
3	1 vs. 2	6.88
2	1 vs. 2	9.37
1	1 vs. 2	15.16
5 vs. 4	1	7.80
5 vs. 4	2	7.52
5 vs. 3	1	30.26
5 vs. 3	2	29.26
5 vs. 2	1	49.88
5 vs. 2	2	48.65
5 vs. 1	1	52.14
5 vs. 1	2	49.03

Predicted SE=regression-predicted standard error

in an experimental modelling analysis simulating all possible combinations of lesions for response classification [24]. However individual tumour response depends in a high proportion on volume change of large nodules and to a lower degree on growth of smaller metastases. Neither of these studies explored the effect of a reduction of the number of assessed lesions on the result of response categorisation, and it remains unclear how many pulmonary target lesions can be excluded without sacrificing observer confidence and thereby introducing misclassifications that could substantially reduce the reliability of this method.

Our results support the findings of the groups of Chojniak et al. and Schwartz et al. showing that the average variance of metastatic volume change increased by 66.6% with reduction of the number of included pulmonary lesions from five to one (Table 3), and the 95% limits of inter-observer agreement were elevated almost threefold (Table 4) and within the range of malignant lesions reported by other authors [21, 25]. This implies that the agreement between our observers assessing only one target lesion would suffer from substantial inaccuracy.

When the assessment result of five lesions was accepted as best available reference, the relative standard error for comparison to datasets with reduced target lesion number increased by more than 600%, exceeding the interobserver error by more than 300% (Table 5). As a result, there were discordances in the response classifications of two and one nodules for both observers (Table 6), and the agreement fell significantly after reduction to one nodule (Table 7). However, a reduction of the number of target lesions from five to three was possible with total observer agreement ($\kappa=1$).

To date, the literature on the number of lesions to treat for confident response assessment is scarce. Zacharia et al. reported no significant discordance between measurements of five and one lesions ($\kappa=0.88$, $p>0.05$). However these authors used manual measurements with original RECIST criteria on liver metastases of unknown size. Therefore their data are not directly translatable to our study [26].

Our results are of potential relevance in various clinical settings: first, chemotherapy trials with application of targeted drugs could benefit from monitoring of minimal

Table 6 Comparison of patient responses when reducing the number of included metastases

	Lesions included	Observer	Response	Response pertaining to five included lesions			
				PD	SD	PR	CR
Response data based on RECIST guidelines using modified volume threshold for category change of 35% PD=progressive disease; SD=stable disease; PR=partial remission CR=complete remission; *intraobserver response discordance	4	1	PD	9	–	–	–
			SD	–	14	–	–
			PR	–	–	10*	–
			CR	–	–	–	–
	4	2	PD	9	–	–	–
			SD	–	14	–	–
			PR	–	–	10*	–
			CR	–	–	–	–
	3	1	PD	9	–	–	–
			SD	–	14	–	–
			PR	–	–	10*	–
			CR	–	–	–	–
	3	2	PD	9	–	–	–
			SD	–	14	–	–
			PR	–	–	10*	–
			CR	–	–	–	–
	2	1	PD	8	2	–	–
			SD	–	12*	–	–
			PR	–	–	10*	–
			CR	1	–	–	–
2	2	PD	8	2*	–	–	
		SD	–	12	–	–	
		PR	–	–	10*	–	
		CR	1	–	–	–	
1	1	PD	7	4	–	–	
		SD	1	7	–	–	
		PR	–	3	7*	–	
		CR	1	–	3	–	
1	2	PD	7	4	–	–	
		SD	1	7	–	–	
		PR	–	3	7*	–	
		CR	1	–	3	–	

Table 7 Categorical agreement on response with varying number of included metastases

Observer	Comparison of numbers of lesions	Kappa	95% CI	Disagreement
1	5 vs. 4	1	0.750–1.243	ns
	5 vs. 3	1	0.750–1.243	ns
	5 vs. 2	0.864	0.630–1.099	ns
	5 vs. 1	0.491	0.284–0.698	p=0.033
2	5 vs. 4	1	0.750–1.243	ns
	5 vs. 3	1	0.750–1.243	ns
	5 vs. 2	0.864	0.630–1.099	ns
	5 vs. 1	0.491	0.284–0.698	p=0.033

All response data based on RECIST guidelines using modified volume threshold for category change of 35%
CI=confidence interval; disagreement=marginal homogeneity method of Maxwell

changes in tumour volume, which is likely to enhance the diagnosis of therapy success, undetectable using standard RECIST criteria [27]. This is of particular importance in small pulmonary metastases <10 mm in diameter that do not allow reliable assessment by other imaging techniques such as positron emission tomography (PET) [28]. Second, a higher sensitivity to volume change may allow for earlier detection of treatment response, thereby allowing for adjustment of highly toxic conventional chemotherapy after shorter follow-up periods with a potential to increase clinical patient benefit [22]. Third, a careful reduction of the smallest target lesions could shorten assessment time significantly, which has to be expected to be less than 5 min per lesion, but was not assessed by this study.

A limitation of our study is the heterogeneous patient collective, which included diverse tumour entities and therapeutic regimens, without independent review of the response rates reported by the study investigators. Also, some of the follow-up intervals were relatively short. However this is unlikely to affect measurement errors and observer agreement in the evaluation of tumour response comparing volume change thresholds within one measure-

ment technique. Further, no primary tumours or extrapulmonary metastatic lesions were evaluated, because the volumetry software is specified to application in pulmonary lesions. The study subgroup of 33 patients with five target lesions is relatively small for assessment of reader confidence on tumour response, and before clinical routine evaluation of therapy response should adopt decreased tumour size change thresholds or shorten follow-up intervals, this data should be confirmed in large populations [24]. Last, the importance of additional clinical data for definition of therapy response should be indicated, which was not the focus of this study.

In conclusion, in our patients, the diagnosis of disease progression, stable disease and partial response after reduction of the volume change threshold from 70% to 35% was well reproducible. The stability of this observer confidence in datasets based on five, four and three metastases proves the robustness of the modified volume change threshold against increasing intra-individual and inter-observer variance. Therefore assessment of tumour response in pulmonary metastases using these modified criteria should include a minimum of three lesions.

References

1. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS et al (2000) New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 92:205–216
2. Therasse P, Eisenhauer EA, Verweij J (2006) RECIST revisited: a review of validation studies on tumor assessment. *Eur J Cancer* 42:1031–1039
3. Hayes TG, Falchock GF, Varadhachary GR et al (2006) Phase I trial of oral talactoferrin alfa in refractory solid tumors. *Invest New Drugs* 24:233–240
4. Husband JE, Schwartz LH, Spencer J et al (2004) Evaluation of the response to treatment of solid tumours—a consensus statement of the International Cancer Imaging Society. *Brit J Cancer* 90:2256–2260
5. Wormanns D, Kohl G, Klotz E et al (2004) Volumetric measurement of pulmonary nodules at multi-row detector CT: in vivo reproducibility. *Eur Radiol* 14:86–92
6. Marten K, Auer F, Schmidt S, Kohl G, Rummeny EJ, Engelke C (2006) Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria. *Eur Radiol* 16:781–790
7. Tran LN, Brown MS, Goldin JG et al (2004) Comparison of treatment response classifications between unidimensional, bidimensional, and volumetric measurements of metastatic lung lesions on chest computed tomography. *Acad Radiol* 11:1355–1360
8. Marten K, Engelke C. Computer-aided detection and automated CT volumetry of pulmonary nodules. *Eur Radiol* 2006; DOI 10.1007/s00330-006-0410-3 [Epub ahead of print]
9. Fan L, Qian J, Odry BL et al (2002) Automatic segmentation of pulmonary nodules by using dynamic cross-correlation for interactive CAD systems. In: Sonka M, Fitzpatrick JM (eds) *Medical imaging: Image processing. Proc SPIE* 4684:1362–1396
10. Fleiss JL, Cuzick J (1979) The reliability of dichotomous judgements: unequal numbers of judges per subject. *Appl Psychol Meas* 3:537–542
11. Fleiss JL, Nee CM, Landis JR (1979) Large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 86:974–977
12. Landis JR, Koch G (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
13. Bland M, Altman DG (1986) Statistical methods for assessing the difference between two methods of measurement. *Lancet* 1(8476):307–310
14. Maxwell AE (1970) Comparing the classification of subjects by two independent judges. *Br J Psychiatry* 116:651–655
15. Thiesse P, Ollivier L, Di Stefano-Louineau D et al (1997) Response rate accuracy in oncology trials: reasons for interobserver variability. *J Clin Oncol* 15:3507–3514
16. Gwyther SJ, Aapro MS, Hatty SR, Postmus PE, Smith IE (1999) Results of an independent oncology review board of pivotal clinical trials of gemcitabine in nonsmall cell lung cancer. *Anticancer Drugs* 10:693–698
17. Biganzoli L, Lohrisch C, Paridaens R, Therasse P, Piccart M (2002) Analysis of two EORTC trials in metastatic breast cancer (MBC) support the recommendation of external response review (ERR) in trials with response rate (RR) as primary endpoint. *Eur J Cancer* 36(Suppl. 5):S90
18. Watanabe H, Yamamoto S, Kunitoh H et al (2003) Tumor response to chemotherapy: the validity and reproducibility of RECIST guidelines in NSCLC patients. *Cancer Sci* 94:1015–1020

-
19. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, Munden RF (2003) Interobserver and intraobserver variability in measurement of nonsmall-cell carcinoma of the lung lesions: implications for assessment of tumour response. *J Clin Oncol* 21:2574–2582
 20. Grossi F, Belvedere O, Fasola G et al (1999) Tumour measurements on computer tomographic images of non-small cell lung cancer were similar among cancer professionals from different specialities. *J Clin Oncol* 17:25–30
 21. Yankelevitz DF, Gupta R, Zhao B, Henschke CI (2000) Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation. *Radiology* 217:251–256
 22. Kostis WJ, Yankelevitz DF, Reeves AP, Fluture SC, Henschke CI (2004) Small pulmonary nodules: reproducibility of three-dimensional volumetric measurement and estimation of time to follow-up CT. *Radiology* 231:446–452
 23. Chojniak R, Younes RN (2003) Pulmonary metastases tumor doubling time: assessment by computed tomography. *Am J Clin Oncol* 26:374–377
 24. Schwartz LH, Mazumdar M, Brown W, Smith A, Paniek DM (2003) Variability in response assessment in solid tumors: effect of number of lesions chosen for measurement. *Clin Cancer Res* 9: 4318–4323
 25. Schwartz M (1961) A biomathematical approach to clinical tumour growth. *Cancer* 14:1272–1294
 26. Zacharia TT, Saini S, Halpern EF, Sumner JE (2005) CT of colon cancer metastases to the liver using modified RECIST criteria: determining the ideal number of target lesions to measure. *AJR Am J Roentgenol* 186:1067–1070
 27. Ratain M (2004) Phase II studies of modern drugs directed against new targets: If you are fazed, too, then resist RECIST. *J Clin Oncol* 22:4442–4444
 28. Kernstine KH, Grannis FW Jr, Rotter AJ (2005) Is there a role for PET in the evaluation in subcentimeter pulmonary nodules? *Semin Thorac Cardiovasc Surg* 17:110–114