Acta Oncologica, 2010; 49: 509–514

informa
healthcare

ORIGINAL ARTICLE

Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria

CHIKAKO SUZUKI[1], MICHAEL R. TORKZAD[2], HANS JACOBSSON[1], GUNNAR ÅSTRÖM[3], ANDERS SUNDIN[1], THOMAS HATSCHEK[4], HIROFUMI FUJII[5] & LENNART BLOMQVIST[1]

[1]*Department of Diagnostic Radiology, Institution for Molecular Medicine and Surgery Karolinska University Hospital Solna and Karolinska Institutet, Stockholm, Sweden,* [2]*Department of Radiology, Uppsala University Hospital, Karolinska Institutet and Uppsala University, Uppsala, Sweden,* [3]*Departments of Oncology, Immunology and Radiology, Uppsala University Hospital, Uppsala, Sweden,* [4]*Department of Oncology, Karolinska University Hospital Solna, Stockholm, Sweden and* [5]*Functional Imaging Division, National Cancer Center East, Chiba, Japan*

Abstract

Background. Response Evaluation Criteria In Solid Tumors (RECIST) and WHO-criteria are used to evaluate treatment effects in clinical trials. The purpose of this study was to examine interobserver and intraobserver variations in radiological response assessment using these criteria. *Material and methods*. Thirty-nine patients were eligible. Each patient's series of CT images were reviewed. Each patient was classified into one of four categories according RECIST and WHO-criteria. To examine interobserver variation, response classifications were independently obtained by two radiologists. One radiologist repeated the procedure on two additional different occasions to examine intraobserver variation. Kappa statistics was applied to examine agreement. *Results*. Interobserver variation using RECIST and WHO-criteria were 0.53 (95% CI 0.33–0.72) and 0.60 (0.39–0.80), respectively. Response rates (RR) according to RECIST obtained by reader A and reader B were 33% and 21%, respectively. RR according to WHO-criteria obtained by reader A and reader B were 33% and 23% respectively. Intraobserver variation using RECIST and WHO-criteria ranged between 0.76–0.96 and 0.86–0.91, respectively. *Conclusion*. Radiological tumor response evaluation according to RECIST and WHO-criteria are subject to considerable inter- and intraobserver variability. Efforts are necessary to reduce inconsistencies from current response evaluation criteria.

Clinical trials are mandatory in the evaluation of new tumor treatments. A commonly studied indicator of the effect of an instituted therapy is the change of size of the malignant lesion(s). This is often assessed and quantified by various radiological techniques. A high accuracy and reproducibility are, for obvious reasons, necessary in order to achieve a meaningful evaluation of such studies.

In order to reduce the confusion caused by various methods used *ad hoc* for therapy evaluation of solid tumors with radiological methods, the World Health Organization criteria (WHO-criteria) were described in 1979 [1]. As an extension and modification of these definitions, the Response Evaluation Criteria In Solid Tumors (RECIST) were published in 2000 [2]. The RECIST has become dominating [3]; an updated version was published as RECIST 1.1 in 2009 [4]. Using these criteria, the therapy response is assessed by measuring tumor size before, during and after the treatment.

Many studies have compared two or three different measurements; RECIST (1-dimensional), WHO-criteria (2-dimensional), and volume (3-dimensional) measurements with various *ad hoc* techniques, and discussed the appropriate measurement for evaluation [2,5–10]. No previous study has evaluated inter- and intraobserver variability when a reader is confronted with a series of radiological images having to select and measure target lesions from various organs for evaluation.

Correspondence: Chikako Suzuki, Department of Diagnostic Radiology, Institution for Molecular Medicine and Surgery, Karolinska University Hospital Solna and Karolinska Institute, Stockholm SE-171 76, Sweden. Tel: +46 851770000. Fax: +46 851774583. E-mail Chikako.Tanaka@ki.se, Chikasakit@yahoo.co.jp

(Received 14 August 2009; accepted 15 February 2010)

ISSN 0284-186X print/ISSN 1651-226X online
DOI: 10.3109/02841861003705794

The purpose of this study was to examine the intra- and interobserver agreement at evaluations using both criteria. This was made by independent classifications of two different radiologists reading the same CT-studies, and by repeated readings of the same CT-studies by one radiologist, retrospectively.

Material and methods

This principally retrospective study was regarded as quality control by the local ethics committee why no patient information or consent in addition to the original trials was required.

The study was carried out using series of CT-examinations previously included in nine clinical trials of various anti-neoplastic treatments. Patients were eligible to the study if they fulfilled the following criteria:(1) included in a clinical trial, using conventional chemotherapy, without hormonal therapy or targeted agents and using RECIST as principle evaluation criteria between 2004 and 2005, (2) having all CT studies, from the baseline until drop-out or at January 31, 2006, i.e. the closing date of the study.

All CT-images were archived and reviewed on our institutional Picture Archiving and Communication Systems (PACS) (SECTRA, Linköping, Sweden). A total of 39 patients were studied. The study design is shown in Figure 1. One patient was included in two different trials with an interval of five months, and regarded as two different individuals. Twenty-four patients had breast cancer and 15 had colorectal cancer. The mean age was 56 years (range 40–80) and the median follow-up time was 139 days (range 29–408).

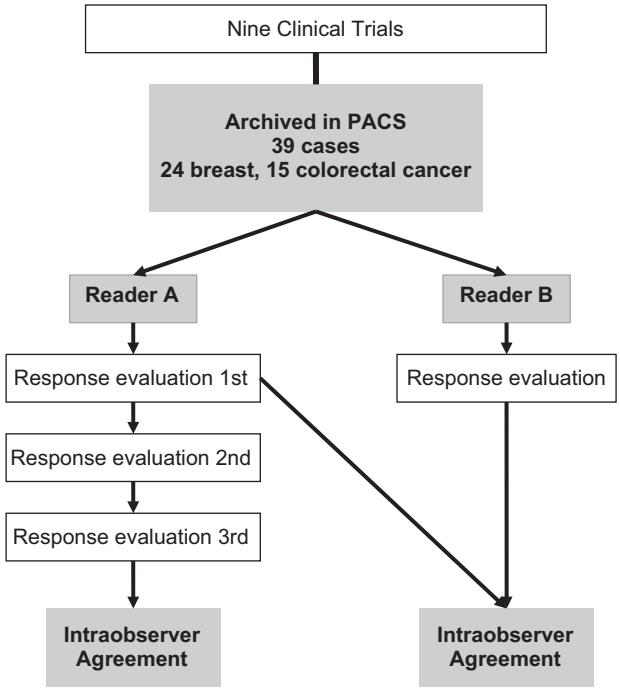


Figure 1. Scheme of the study.

All series of CT images for each patient were retrieved and reanalyzed. Tumor sizes were measured using electronic calipers. No values or markers were left in the PACS between the evaluations.

Two board certified radiologists with considerable experience of RECIST read the examinations (C.S. and M.T., denoted A and B, respectively). Both readers performed the same assessment procedure independently being blinded to each others results and clinical information of the individual patient. Each reader reviewed the series of CT-examinations from inclusion until the reader judged the patient showing progressive disease (PD) or until January 31, 2006. The reader selected target lesions and measured the longest diameter (LD) in the axial plane following RECIST. Simultaneously, the perpendicular longest diameter of the target lesion was obtained to allow response evaluation according to the WHO-criteria. Changes of tumor size as well as appearance of new lesion(s) and/or progression of non-target lesions were monitored. Finally each patient was classified into one of four categories: complete response (CR), partial response (PR), stable disease (SD) or PD according to the RECIST and WHO-criteria, respectively. A patient was classified as PD according to WHO-criteria if the sum of the products of LD and perpendicular LD increased 25% and more in one or more target lesion, which differs from RECIST's definition: 20% and more increase of the sum of LDs. When tumor response was classified as PD, the reason for this classification was noted as either: (1) increase in size of target lesions or (2) appearance of new lesion(s) and/or progression of non-target lesions. Objective response rate was defined as the percentage of the sum of CR and PR, those who were considered as "responder", in relation to all patients: (CR+PR)/all patients. Reader A repeated the above-mentioned procedure on two additional occasions with at least a six-week interval (Figure 1).

Statistical analyses

Wilcoxon matched pairs test was used to compare the number of selected target lesions by the readers. Friedman ANOVA by ranks was used to compare the number of selected target lesions among three repeated evaluations by reader A.

Kappa analysis was performed to evaluate interobserver and intraobserver agreements and to evaluate agreements between RECIST and WHO-criteria for each reader (Figure 1). Kappa values above 0.81, 0.61–0.8, 0.41–0.60, 0.21–0.40, and 0–0.2 indicated almost perfect, substantial, moderate, fair and slight agreements, respectively [11].

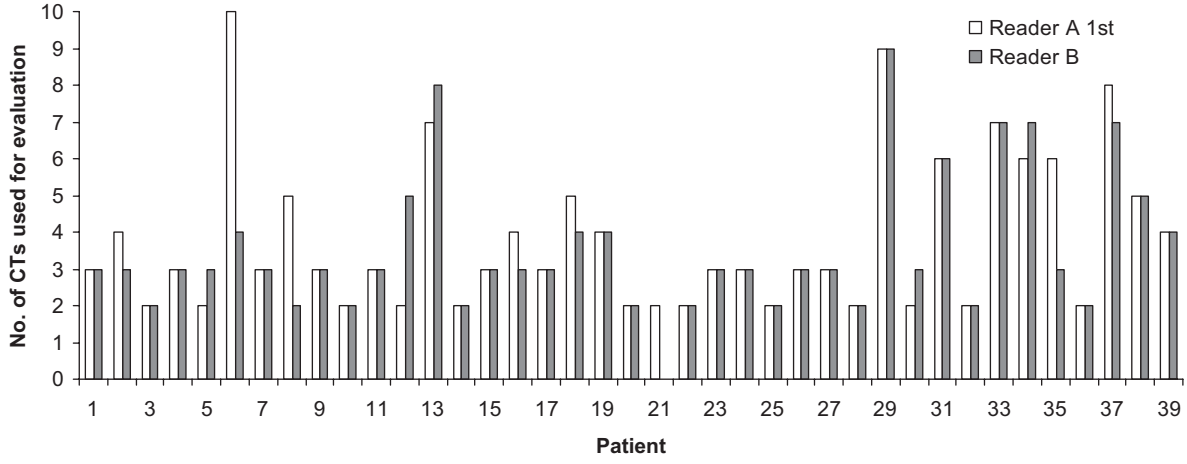


Figure 2. Number of CT examinations used for RECIST evaluation by reader A 1st and reader B. Each patient was followed from inclusion until when PD was demonstrated on CT or closing date of current study. Since reader B did not considered that patient No. 21 had any measurable lesion, radiological response evaluation was not performed in this case.

A value of $p<0.05$ was considered a statistically significant difference.

Statistical analysis was performed using StatXact 4 (CYTEL Software Corporation, Cambridge, MA, USA).

Results

Image interpretation

As shown in Figure 2, the number of CT-examinations chosen for response evaluation was different between the readers. This was mainly because of inconsistency in judging the same patient as PD at the same time. Reader B excluded one patient from the evaluation because of the absence of target lesions.

Figure 3 shows the numbers of selected target lesions. Statistically significant differences were demonstrated in the number of selected liver metastases, lymph node metastases, metastases in other organs and totally.

A total of 109 liver metastases were altogether selected by the readers. Seventy-one of the 109 liver metastases were different. Likewise, 10/35 lung metastases, 43/59 lymph node metastases, 13/16 other organ metastases and in total 137 of 219 metastases were differently selected by the readers. On the other hand, the readers selected identical lesions in five of 39 patients, these patients showed relatively few lesions; median 3. It was common that even though two readers selected the same lesion, the way of measurement was different (Figure 4).

Agreement in response evaluation

Figure 5 shows kappa coefficient values with 95% CI of interobserver agreement and intraobserver agreement. Interobserver agreement remained to be moderate and was generally lower than intraobserver agreement, which tended to be substantial to perfect.

Thirteen of 39 patients were differently classified by the readers according to RECIST (Table I, and 10 of 39 patients were differently classified according

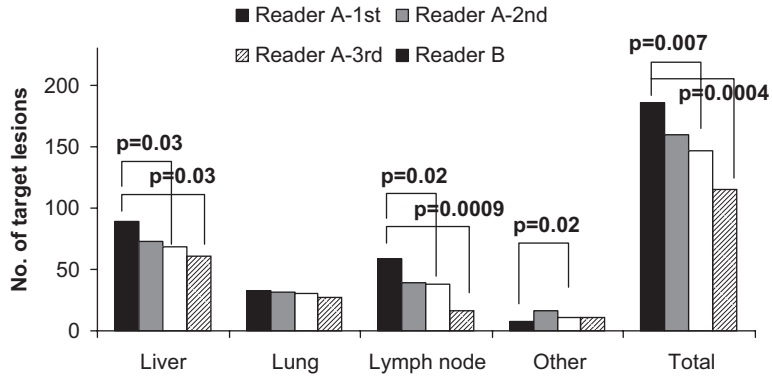


Figure 3. The number of target lesions selected by reader A and by reader B.

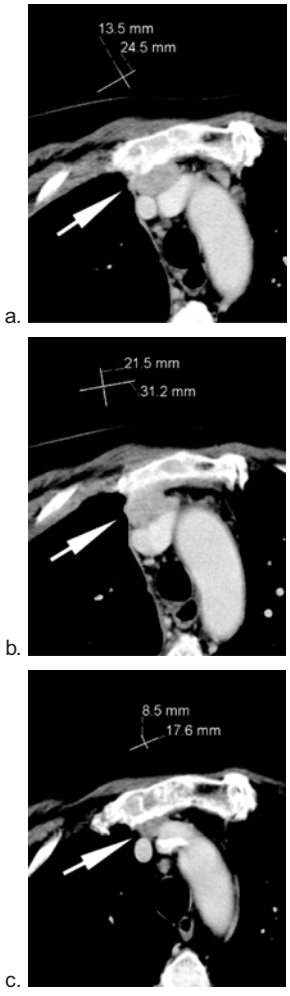


Figure 4. Possible cause for inconsistency; difference in measurement. A 62-year-old patient with retrosternal lymph node metastasis (arrow) before (a and b) and after therapy (c). Figure (a) and (b) represent two consecutive sections at the baseline study. If the tumor was measured as shown in (a), the patient will be classified as SD. On the other hand, if the same lesion was measured as shown in (b), the patient will be classified as PR.

Table I. Joint judgments of two readers regarding tumor responses according to RECIST.

		Reader A					
		CR	PR	SD	PD	Excluded	Total
Reader B	CR	2					2
	PR		5		1		6
	SD		5	6	5		16
	PD			1	13		14
	Excluded		1				1
	Total	2	11	7	19	0	39

Notes: CR: complete response, PR: partial response, SD: stable disease, PD: progressive disease.

to WHO-criteria (Table II). Response rates (RR) according to RECIST were 33% by reader A and 21% by reader B. RR according to WHO-criteria were 33% by reader A and 23% by reader B.

Interobserver agreement in the detection of new lesion(s) and progression of non-target lesions was moderate, $\kappa = 0.50$ (0.28–0.73) (Figure 5, Table III). Even among five patients whose target lesions were identical between the two readers, patients were judged differently mainly because inconsistency in interpretation of new lesions and progression of non-target lesions (Figure 6). Seven of 39 patients had received completely opposite evaluation regarding PD or not.

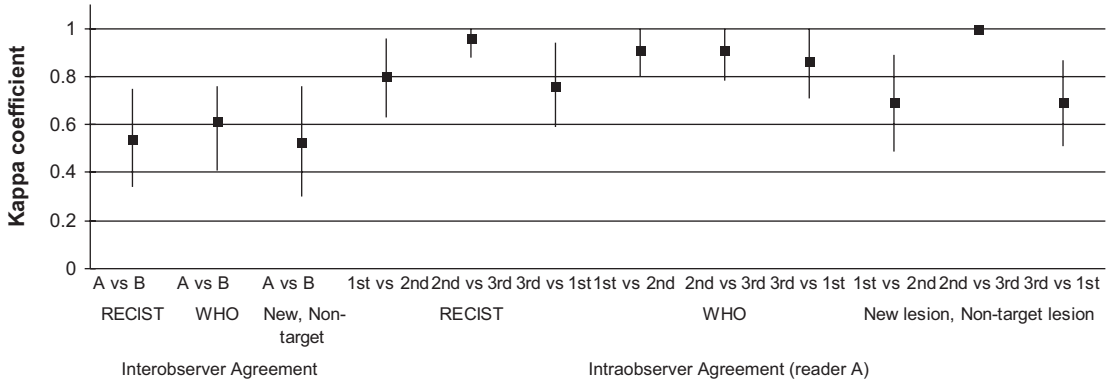
Intraobserver agreements in the detection of new lesion(s) or progression of non-target lesions ranged substantial to perfect (Table III).

Comparison to clinical judgment

Reader A evaluated falsely four patients and reader B evaluated falsely nine patients as PD, when patients were considered free from progression clinically.

Discussion

In order to allow adequate analyses in the evaluation of clinical trials, quantitative information is required.



95% CI upper	0.75	0.76	0.76	0.96	1	0.94	1	1	1	0.89	1	0.87
95% CI lower	0.34	0.41	0.3	0.63	0.88	0.59	0.8	0.78	0.71	0.49	1	0.51
Kappa coefficient	0.54	0.61	0.53	0.8	0.96	0.76	0.91	0.91	0.86	0.69	1	0.69

Figure 5. Non-weighted kappa coefficient value and corresponding 95% confidence interval (CI) for agreement.

Table II. Joint judgments of two readers regarding tumor responses according to WHO-criteria.

		Reader A					
		CR	PR	SD	PD	Excluded	Total
Reader B	CR	2					2
	PR		6		1		7
	SD		3	3	3		9
	PD		1	1	18		20
	Excluded		1				1
	Total	2	11	4	22	0	39

Notes: CR: complete response, PR: partial response, SD: stable disease, PD: progressive disease.

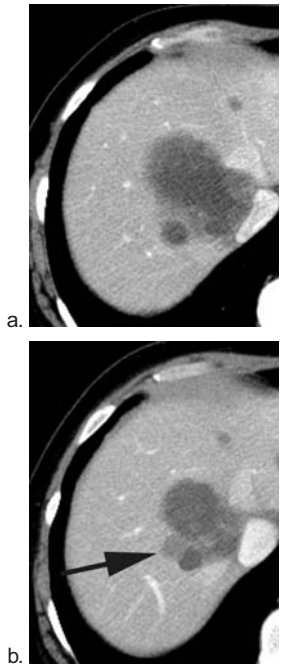


Figure 6. Possible cause for inconsistency; "new lesion". A 50-year-old patient with multiple liver metastases at the baseline study (a). After seven cycles of treatment, a low-attenuation lesion was depicted adjacent to the known metastases (arrow in b). Reader A interpreted this lesion as a "new lesion" indicating progressive disease (PD) regardless decrease size of other metastases, while reader B interpreted this lesion existed at baseline and classified as partial response (PR).

The current study deals with the conversion of the analogous information provided by the radiological studies into digital data which thereafter seldomly is questioned. Numerical information tends to appear very precise but cannot be more exact than how they are achieved. The WHO and RECIST-criteria were set up in order to make such conversions in a strict and standardized way. The crucial, but subjective steps, when a lesion is selected and thereafter measured have, however, surprisingly little been called in question.

The study showed significant inconsistencies between readers in the selection, in the measurement of the lesions as well as in tracking new lesions and in the attention of non-target lesions. This leads to considerable discrepancies in response evaluation using both WHO and RECIST-criteria despite standardized conditions on how to make the evaluations. The possible effects of these findings on the reliability of clinical trials are obvious.

The study is based on a restricted number of observations and on a heterogenous patient group why it must be regarded as a pilot study having to be repeated in larger materials and at other institutions. Nevertheless, and especially considering these limitations, the inconsistency of how lesions are selected, measured and interpreted between different readers is an important finding. Intraobserver agreement tended in general to be better than the interobserver agreement. This might be because the constancies in lesion selection, measurement and finding new lesion increased during repetition of evaluation, hence random errors decreased. This also indicates, on the other hand, a risk of underestimation of intraobserver inconsistency. Outside this, we could not find any systematic difference between the readers. Two radiologists with equivalent experience performed all evaluations using the same criteria and the same workstation.

There are emerging measurement software tools and values with intent to reduce both systematic and random error in the measurement of target of lesions [9,10,12–16]. Yet, it is questionable how much the measurement itself contributes to the consistency. This can be questioned considering almost perfect agreement in reader A's 2nd and 3rd evaluation in our study. On the other hand, little has been done to focus on the selection and the finding of new lesions

Table III. Joint judgments of two readers regarding detection of new lesions and/or progression of non-target lesions.

		Reader A				
		PD new/non	PD target	Not-PD	Excluded	Total
Reader B	PD new/non	8	1	1		10
	PD target	2	2			4
	Not-PD	6		18		24
	Excluded			1		1
	Total	16	3	20	0	39

Notes: PD new/non: progressive disease because of appearance of new lesions and/or progression of non-target lesions.
PD target: progressive disease because of increasing size of target lesion without new lesion or progression of non-target lesion.

and the progression of non-target lesions. This might be because of difficulties to deal with random error.

A challenge highlighted by this study is how to select lesions in a consistent way. Reducing the number of target-lesions as suggested by several retrospective studies may have negative impact on consistency [4,17–19]. This is because the contribution of a single lesion to the response evaluation increases when the number of lesions is reduced. It remains doubtful that only one or two arbitrary selected lesions can represent a patient's true response.

Possible improvements to increase consistency in the evaluations may be parallel readings or consensus readings by two or several radiologists or repeated readings by the same individual. This would, however, increase the cost and time for clinical trials already encumbered with such problems [3]. Another way of solving these problems may be that more detailed guidelines are added to the respective criteria. In a large perspective, however, these problems may be improved by emerging techniques.

In conclusion, the WHO and RECIST-criteria are indispensable to analyze important surrogate indicators such as response rate and progression-free survival in clinical trials. However this study casts doubt on consistencies in evaluations using the criteria. Efforts to minimize possible source of random errors and inconsistencies are required.

Acknowledgements

The authors thank Elisabeth Berg, BSc, The Medical Statistics Unit, Department of Learning, Informatics, Management and Ethics, Karolinska Institute, Stockholm, Sweden for professional statistical analysis. The authors greatly appreciate Maria Gustafsson Liljefors, MD, PhD, Katarina Bodén, MD, and Yvonne Eriksson-Alm, RT, for their advice during this study. Financial support was provided through the regional agreement on medical training and clinical research (ALF) between the Stockholm county council and the Karolinska Institute.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

[1] WHO handbook for reporting results of cancer treatment. Publication number 48. Geneva: World Health Organization; 1979. Available from: http://whqlibdoc.who.int/publications/9241700483.pdf.

[2] 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. 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;92:205–16.

[3] Nygren P, Blomqvist L, Bergh J, Astrom G. Radiological assessment of tumour response to anti-cancer drugs: Time to reappraise. Acta Oncol 2008;47:316–8.

[4] 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.

[5] James K, Eisenhauer E, Christian M, Terenziani M, Vena D, Muldal A, et al. Measuring response in solid tumors: Unidimensional versus bidimensional measurement. J Natl Cancer Inst 1999;91:523–8.

[6] Park JO, Lee SI, Song SY, Kim K, Kim WS, Jung CW, et al. Measuring response in solid tumors: Comparison of RECIST and WHO response criteria. Jpn J Clin Oncol 2003;33:533–7.

[7] Trillet-Lenoir V, Freyer G, Kaemmerlen P, Fond A, Pellet O, Lombard-Bohas C, et al. Assessment of tumour response to chemotherapy for metastatic colorectal cancer: Accuracy of the RECIST criteria. Br J Radiol 2002;75:903–8.

[8] Fabel M, von Tengg-Kobligk H, Giesel FL, Bornemann L, Dicken V, Kopp-Schneider A, et al. Semi-automated volumetric analysis of lymph node metastases in patients with malignant melanoma stage III/IV-A feasibility study. Eur Radiol 2008;18:1114–22.

[9] Sohaib SA, Turner B, Hanson JA, Farquharson M, Oliver RT, Reznek RH. CT assessment of tumour response to treatment: Comparison of linear, cross-sectional and volumetric measures of tumour size. Br J Radiol 2000;73:1178–84.

[10] Tran LN, Brown MS, Goldin JG, Yan X, Pais RC, McNitt-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.

[11] Kundel HL, Polansky M. Measurement of observer agreement. Radiology 2003;228:303–8.

[12] Marten K, Auer F, Schmidt S, Kohl G, Rummeny EJ, 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.

[13] Marten K, Auer F, Schmidt S, Rummeny EJ, 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.

[14] Honda O, Sumikawa H, Johkoh T, Tomiyama N, Mihara N, Inoue A, et al. Computer-assisted lung nodule volumetry from multi-detector row CT: Influence of image reconstruction parameters. Eur J Radiol 2007;62:106–13.

[15] Zhao B, Schwartz LH, Moskowitz CS, Wang L, Ginsberg MS, Cooper CA, et al. Pulmonary metastases: Effect of CT section thickness on measurement—initial experience. Radiology 2005;234:934–9.

[16] Jacene HA, Leboulleux S, Baba S, Chatzifotiadis D, Goudarzi B, Teytelbaum O, et al. Assessment of interobserver reproducibility in quantitative 18F-FDG PET and CT measurements of tumor response to therapy. J Nucl Med 2009; 50:1760–9.

[17] Schwartz LH, Mazumdar M, Brown W, Smith A, Panicek DM. Variability in response assessment in solid tumors: Effect of number of lesions chosen for measurement. Clin Cancer Res 2003;9:4318–23.

[18] Zacharia TT, Saini S, Halpern EF, Sumner JE. 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 2006;186:1067–70.

[19] Darkeh MH, Suzuki C, Torkzad MR. The minimum number of target lesions that need to be measured to be representative of the total number of target lesions (according to RECIST). Br J Radiol 2009;82:681–6.