Upsala Journal of Medical Sciences

ISSN: 0300-9734 (Print) 2000-1967 (Online) Journal homepage: https://www.tandfonline.com/loi/iups20

Tumor response evaluation criteria for HCC (hepatocellular carcinoma) treated using TACE (transcatheter arterial chemoembolization): RECIST (response evaluation criteria in solid tumors) version 1.1 and mRECIST (modified RECIST): JIVROSG-0602

Yozo Sato, Hirokazu Watanabe, Miyuki Sone, Hiroaki Onaya, Noriaki Sakamoto, Keigo Osuga, Masahide Takahashi, Yasuaki Arai & Japan Interventional Radiology In Oncology Study Group (JIVROSG)

To cite this article: Yozo Sato, Hirokazu Watanabe, Miyuki Sone, Hiroaki Onaya, Noriaki Sakamoto, Keigo Osuga, Masahide Takahashi, Yasuaki Arai & Japan Interventional Radiology In Oncology Study Group (JIVROSG) (2013) Tumor response evaluation criteria for HCC (hepatocellular carcinoma) treated using TACE (transcatheter arterial chemoembolization): RECIST (response evaluation criteria in solid tumors) version 1.1 and mRECIST (modified RECIST): JIVROSG-0602, Upsala Journal of Medical Sciences, 118:1, 16-22, DOI: 10.3109/03009734.2012.729104

To link to this article: https://doi.org/10.3109/03009734.2012.729104

Upsala Journal of Medical Sciences. 2013; 118: 16–22

informa
healthcare

ORIGINAL ARTICLE

Tumor response evaluation criteria for HCC (hepatocellular carcinoma) treated using TACE (transcatheter arterial chemoembolization): RECIST (response evaluation criteria in solid tumors) version 1.1 and mRECIST (modified RECIST): JIVROSG-0602

YOZO SATO[1], HIROKAZU WATANABE[2], MIYUKI SONE[2], HIROAKI ONAYA[2], NORIAKI SAKAMOTO[2], KEIGO OSUGA[3], MASAHIDE TAKAHASHI[2], YASUAKI ARAI[2] AND JAPAN INTERVENTIONAL RADIOLOGY IN ONCOLOGY STUDY GROUP (JIVROSG)

[1]*Department of Diagnostic and Interventional Radiology, Aichi Cancer Center Hospital, Nagoya, Japan,* [2]*Department of Diagnostic Radiology, National Cancer Center Hospital, Tokyo, Japan, and* [3]*Department of Diagnostic Radiology, Osaka University Graduate School of Medicine, Osaka, Japan*

Abstract

Background. Two standard sets of criteria are used to evaluate the tumor response of hepatocellular carcinoma (HCC): RECIST (Response Evaluation Criteria in Solid Tumors) and modified RECIST (mRECIST). The purpose was to compare two tumor response evaluation criteria, RECIST version 1.1 and mRECIST, for HCC treated using transcatheter arterial chemoembolization (TACE).

Methods. The radiological findings of patients who underwent TACE for HCCs in a multicenter clinical trial were examined. Sixty-five lesions in 21 patients treated with TACE without mixing iodized-oil were evaluated. The tumor size was evaluated by measuring the entire lesion, including the necrotic part, using RECIST version 1.1, whereas only the contrast-enhanced part observed during the arterial phase was measured using mRECIST. Five radiologists independently measured each lesion twice. To evaluate the inter-criteria reproducibility, the complete response (CR) rate, the response rate, the kappa statistics, and the proportion of agreement (PA) for response categories were calculated. The same analyses were conducted for inter- and intra-observer reproducibility.

Results. In the inter-criteria reproducibility study, the CR rate and the response rate obtained using mRECIST (56.9% and 79.7%) were higher than those obtained using RECIST version 1.1 (9.2% and 43.1%). In the inter- and intra-observer reproducibility study, mRECIST exhibited an 'almost perfect agreement', while RECIST version 1.1 exhibited a 'substantial agreement'.

Conclusions. Considerable differences in the CR rate and the response rate were observed. From the viewpoint of the high inter- and intra-observer reproducibility, mRECIST may be more suitable for tumor response criteria in clinical trials of TACE for HCC.

Key words: *Hepatocellular carcinoma, modified RECIST, RECIST version 1.1, reproducibility, tumor response*

Introduction

Two standard sets of criteria are used to evaluate the tumor response of hepatocellular carcinoma (HCC) treated using loco-regional therapy, such as transcatheter arterial embolization (TACE): RECIST (Response Evaluation Criteria in Solid Tumors) criteria (1) and modified RECIST (mRECIST) criteria (2).

RECIST criteria were published by the National Cancer Institute in 2000 with the objective of unifying

Correspondence: Hirokazu Watanabe, Department of Diagnostic Radiology, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan. Fax: +81-3-3542-3815. E-mail: hirwatan@ncc.go.jp

(Received 8 June 2012; accepted 6 September 2012)

ISSN 0300-9734 print/ISSN 2000-1967 online
DOI: 10.3109/03009734.2012.729104

the criteria used for response assessments. These criteria evaluate the unidimensional measurement of the longest diameter of the tumor lesions and have been used in most oncology trials. However, a number of questions and issues have arisen, leading to the development of revised RECIST (version 1.1) criteria (3). In the RECIST version 1.1 criteria, the major changes included the number of lesions to be assessed, the assessment of pathological lymph nodes, confirmation of a response, disease progression, and the necrotic tumor size (i.e. in cases where a lesion which was solid at baseline has become necrotic in the center, the longest diameter of the entire lesion should be followed).

In 2000, a panel of experts on HCC from the European Association for the Study of the Liver (EASL) agreed that estimating the reduction in viable tumor volume (as recognized using enhanced spiral computed tomography (CT)) should be considered the optimal method for assessing the local response to treatment in patients with HCC (4). Since then, most authors reporting the results of loco-regional therapy for HCC have evaluated tumor response according to this recommendation (5,6).

The aforementioned expert panel continued the concept of viable tumor endorsed by EASL and adapted the unidimensional measurement as a substitute for the bidimensional one in the determination of tumor response for target lesions in HCC (7). These amendments confirmed the American Association for the Study of Liver Disease (AASLD)–Journal of the National Cancer Institute (JNCI) guidelines and were defined as 'modified RECIST (mRECIST)' criteria (2). Therefore, mRECIST criteria were developed for loco-regional therapies to HCC. On the other hand, RECIST version 1.1 criteria were developed for systemic therapies; however, RECIST version 1.1 criteria are used in many oncology trials including loco-regional therapies for the treatment of HCC.

A study investigating the inter-criteria reproducibility between the older versions of criteria (RECIST version 1.0 and EASL) has been reported (8). Furthermore, a comparative study of tumor response by the updated criteria (RECIST version 1.1 and mRECIST) has been published (9). However, to the best of our knowledge, the inter- and intra-observer reproducibility between RECIST version 1.1 and mRECIST has not been investigated or reported.

Using these standardized criteria for evaluating tumor response in clinical trials, reproducible results should be obtained by all investigators. For a surrogate marker such as tumor response for therapy, both 'precision' (observer consistency study) and 'accuracy' (validation study comparing to gold standard) are evaluated. From the viewpoint of 'precision', we compared RECIST version 1.1 and mRECIST criteria by evaluating the inter- and intra-observer reproducibility.

The purpose of the present study was to clarify the differences in tumor response as evaluated using two updated sets of criteria (RECIST version 1.1 and mRECIST) by assessing the inter-criteria reproducibility. Moreover, another purpose of the present study was to investigate which set of criteria was superior for use as tumor response evaluation criteria in clinical trials of TACE for HCC by assessing the inter- and intra-observer reproducibility.

Materials and methods

We analyzed the radiological findings of patients who underwent pan-hepatic TACE for multiple HCCs in a multicenter clinical trial. In this trial, the eligibility criteria included patients with untreated, bilobar multiple HCCs, compensated Child–Pugh A or B cirrhosis, and the absence of vascular invasion or extrahepatic spread. TACE was performed using cisplatin (IA call, Nihon-Kayaku; 35–65 mg/m^2) and gelatin particles without mixing iodized-oil. The present study was conducted in accordance with the Helsinki Declaration, and the protocols were approved by the institutional review board. Informed written consent for the treatment protocols, including the secondary use of treatment-associated documents, was obtained from each patient. Twenty-one patients were entered from 19 July 2005 to 15 May 2007.

Image analysis

All patients underwent a dynamic study performed using a multi-slice CT scanner with non-ionic contrast medium. CT scans were obtained within two weeks before TACE and one month after TACE. Tumor assessments were made using a 5-mm interval, and axial images were obtained during the unenhanced phase, the arterial phase, and the portal venous or equilibrium phase.

Tumor response evaluation

Response was defined according to RECIST version 1.1 criteria measuring the entire lesion, including the necrotic part. On the other hand, mRECIST were used to evaluate the lesion taking tumor necrosis, recognized by the non-enhanced areas, into account. Both guidelines adopted the unidimensional measurement (Figure 1).

According to RECIST version 1.1 criteria, a complete response (CR) was defined as the disappearance

of all target lesions; a partial response (PR) was defined as at least a 30% decrease in the sum of the longest diameter of the target lesions; progressive disease (PD) was defined as at least a 20% increase in the sum of the longest diameter of the target lesions; and stable disease (SD) was defined as neither sufficient shrinkage to qualify for PR nor a sufficient increase to qualify for PD.

According to mRECIST criteria, CR was defined as the absence of enhanced tumor areas during the arterial phase, reflecting complete tissue necrosis; PR was defined as at least a 30% decrease, PD was defined as at least a 20% increase in the sum of the longest diameter in the enhanced tumor areas; and SD was defined using the same definition as that used in RECIST version 1.1 criteria.

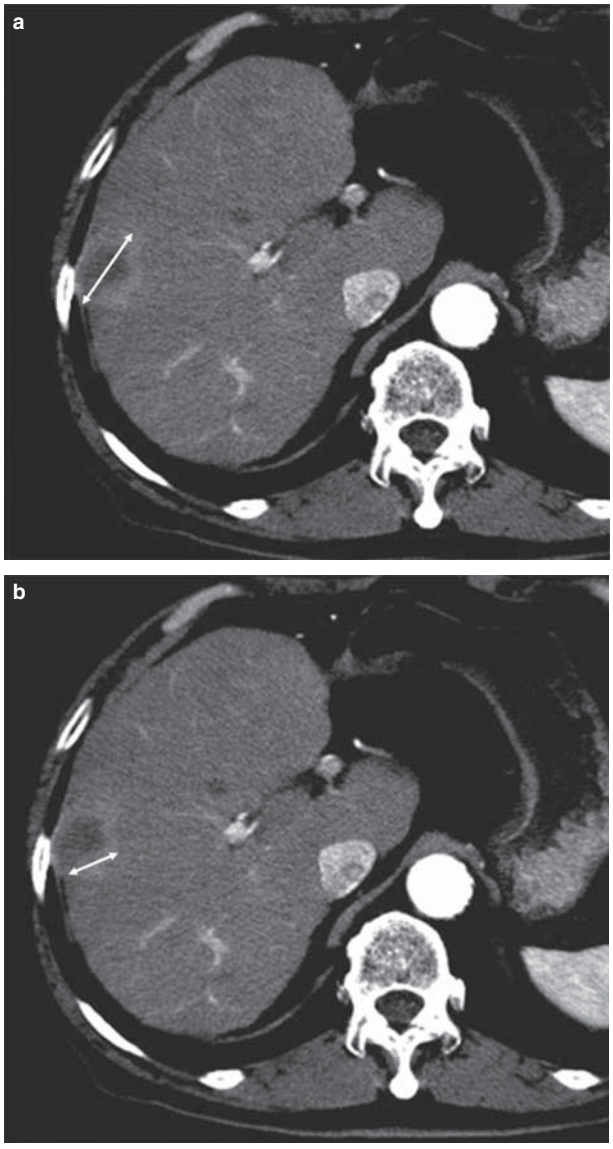

Figure 1. A: RECIST ver. 1.1: Response was defined according to a unidimensional measurement of the entire lesion, including the necrotic part. B: mRECIST: Response was defined according to a unidimensional measurement of the viable part, excluding the necrotic part.

Evaluation methods

Five observers measured 65 lesions in 21 patients independently. A total of 325 measurements were made for the first measurement. The second measurement was performed independently by the same five observers. The sum of the longest diameters for all the target lesions was calculated for baseline and post-treatment. The baseline sum was used as the reference from which the objective tumor response could be calculated. The percentage changes were calculated as the post-treatment value divided by the pre-treatment value. The percentage changes were then classified using RECIST version 1.1 and mRECIST tumor response classification systems. Tumor response was categorized as CR, PR, SD, or PD based on both sets of criteria. Furthermore, the CR rate and the response rate were also calculated.

All the images were collected from each institution and supplied to the Japan Interventional Radiology in Oncology Study Group (JIVROSG) Data Center using the WEB system.

Analysis of inter-criteria reproducibility

To examine the inter-criteria reproducibility between RECIST version 1.1 and mRECIST criteria, we estimated the kappa statistics and the proportion of agreement for the CR, PR, SD, and PD categories among the five observers. The data for the first measurements were analyzed to evaluate the inter-criteria reproducibility.

Analysis of inter-observer reproducibility

To examine the inter-observer reproducibility among the five observers, we estimated the kappa statistics and the proportion of agreement. Each pair yielded 10 pairs for comparison. The data for the first measurements were analyzed to evaluate the inter-observer reproducibility.

Analysis of intra-observer reproducibility

The data for the first and second measurements were compared to assess the intra-observer reproducibility for the same observer. The intra-observer reproducibility for the same observer yielded five pairs for comparison.

Statistics

Kappa statistics were performed to determine the concordance/agreement of the tumor response criteria. The potential kappa values ranged from –1.0 (complete disagreement) through 0 (chance agreement) to 1.0 (complete agreement). Interpretations of the strength of the agreement determined using the kappa values were given by adopting the criteria (9). The kappa values of the two agreements were compared for statistical significance using a paired *t* test. Comparisons between groups were done using the Fisher exact test. A conventional *P* value of 0.05 was considered statistically significant. All analyses were conducted using SPSS (version 17.0).

Results

Patient population

Sixty-five untreated lesions in 21 patients treated using pan-hepatic TACE were evaluated. The patients' characteristics were as follows (Table I), median age (range): 68 years (27–74 years); sex (male/female): 19/2; hepatitis C virus/hepatitis B virus/others: 12/3/6; Child–Pugh A/B: 20/1; total number of nodules (range): 65 nodules (1–5 nodules); mean tumor size (range): 20 mm (10–132 mm).

Inter-criteria reproducibility

The inter-criteria reproducibility using RECIST version 1.1 and mRECIST criteria is summarized in Tables II and III. Five observers measured 65 lesions independently, for a total of 325 measurements. According to RECIST version 1.1 criteria, the CR rate and the response rate were 9.2% and 43.1%, respectively; according to mRECIST criteria, the CR rate and the response rate were 56.9% and 79.7% (Table II).

Among the 185 CR lesions that were identified using mRECIST criteria, RECIST version 1.1 criteria classified the same responses as PR for 89 lesions, SD for 64 lesions, and PD for 2 lesions (Table III). The kappa value was 0.149 (95% CI 0.098–0.201), and the proportion of agreement was 35.5% (Table III).

Table I. Patients and characteristics.

No. of patients	21
Age, median (range)	68 (27–74)
Sex (male/female)	19/2
HCV/HBV/others	12/3/6
Child–Pugh A/B	20/1
No. of nodules, all (range)	65 (1–5)
Mean tumor size (range), mm	20 (10–132)

HCV = hepatitis C virus; HBV = hepatitis B virus.

Inter-observer reproducibility

The inter-observer reproducibility among the five observers was analyzed using the data for the first measurements, with each pair yielding 10 pairs for comparison. These 10 pairs for comparisons, or 650 measurements, are collectively shown in Table IV. For the inter-observer reproducibility for RECIST version 1.1, the kappa value was 0.628 (95% CI 0.571–0.684), and the proportion of agreement was 78.8%. For the inter-observer reproducibility for mRECIST, the kappa value was 0.829 (95% CI 0.792–0.866), and the proportion of agreement was 90.0%.

Intra-observer reproducibility

The intra-observer reproducibility was analyzed from the data for the first and second measurements, with each pair yielding five pairs for comparison. These five pairs for comparisons, or 325 measurements, are collectively shown in Table V. For the intra-observer reproducibility for RECIST version 1.1, the kappa value was 0.643 (95% CI 0.565–0.722), and the proportion of agreement was 79.4%. For the intra-observer reproducibility for mRECIST, the kappa value was 0.900 (95% CI 0.858–0.942), and the proportion of agreement was 94.2%.

Discussion

The inter-criteria reproducibility study between RECIST version 1.0 and EASL guidelines, and a comparative study of tumor response by RECIST and mRECIST have been reported (8,9). However, no information is available concerning the inter-observer reproducibility in those reports. In addition to performing an inter-criteria reproducibility study, we also estimated the inter- and intra-observer reproducibility to investigate which set of criteria (RECIST version 1.1 or mRECIST) is superior for performing tumor response evaluations in clinical trials of TACE for HCC.

Inter-criteria reproducibility

An evaluation of the tumor response according to RECIST version 1.0 and EASL guidelines after loco-regional therapies in patients with HCC has been reported. RECIST missed all the CRs obtained by

Table II. Inter-criteria reproducibility between RECIST version 1.1 and mRECIST criteria. Number of lesions (%).

Response category	Complete response	Partial response	Stable disease	Progressive disease	Overall response[a]
Response criteria					
RECIST	30 (9.2)	110 (33.8)	180 (55.4)	5 (1.5)	140 (43.1)
	$P < 0.001$				$P < 0.001$
mRECIST	185 (56.9)	74 (22.8)	65 (20)	1 (3)	259 (79.7)

[a]Complete response + partial response.
RECIST = Response Evaluation Criteria in Solid Tumors; mRECIST = modified RECIST.

Table III. Inter-criteria reproducibility between RECIST version 1.1 and mRECIST criteria: distribution chart.

		RECIST				
		Complete response	Partial response	Stable disease	Progressive disease	Total
mRECIST	Complete response	30	89	64	2	185
	Partial response	0	21	53	0	74
	Stable disease	0	0	63	2	65
	Progressive disease	0	0	0	1	1
Total		30	110	180	5	325

Proportion of agreement = 35.5%. Kappa = 0.149.

tumor necrosis and underestimated the extent of the partial tumor response because of tissue necrosis (8).

In our inter-criteria reproducibility study comparing RECIST version 1.1 and mRECIST criteria, similar results were obtained. The CR rate and the response rate obtained using mRECIST criteria were higher than those obtained using RECIST version 1.1 criteria (56.9% versus 9.2%, $P < 0.001$; 79.7% versus 43.1%, $P < 0.001$).

According to mRECIST criteria, if a tumor that was solid at baseline became entirely necrotic, all the tumors were evaluated as CR. On the other hand, using RECIST version 1.1 criteria, the necrotic tumor was evaluated as a non-CR based on the measurement of the entire lesion, leading to a different conclusion, such as PR, SD, or PD (Figure 2). Among 185 CR lesions that were identified using mRECIST criteria, 155 lesions (83.8%) were evaluated as non-CR using RECIST version 1.1 criteria. In particular, two lesions evaluated as CR using mRECIST criteria were categorized as PD using RECIST version 1.1 criteria; thus, two sets of criteria produced opposite conclusions (Table III). As the tumor size was very small and a 20% increase was thought to be within the range of measurement error, these two lesions were identified as PD using RECIST version 1.1 criteria. In some cases, this event might be caused by an increase in the necrotic tumor size secondary to chemoembolization. Therefore, the inter-criteria reproducibility between RECIST version 1.1 and mRECIST criteria for loco-regional therapy achieving complete tumor necrosis may have a low concordance.

The differences in the CR rate and the response rate between RECIST version 1.1 and mRECIST criteria indicate that the researchers should ascertain the presence or absence of 'm' (mRECIST? or RECIST?).

Table IV. Inter-observer reproducibility.

	Kappa	Proportion of agreement (%)
Inter-observer reproducibility		
RECIST	0.628 (95% CI 0.571–0.684)	78.8
mRECIST	0.829 (95% CI 0.792–0.866)	90.0

Inter- and intra-observer reproducibility

Standardized tumor response evaluation systems are considered to be reliable in clinical trials when they are reproducible among different observers. The importance of inter-observer reproducibility for any

Table V. Intra-observer reproducibility.

	Kappa	Proportion of agreement (%)
Intra-observer reproducibility		
RECIST	0.643 (95% CI 0.565–0.722)	79.4
mRECIST	0.900 (95% CI 0.858–0.942)	94.2

classification scheme has been discussed previously for other grading systems (10-14). Clinical investigators must take into account inter-observer reproducibility in tumor response evaluations, which can greatly affect the results of clinical trials.

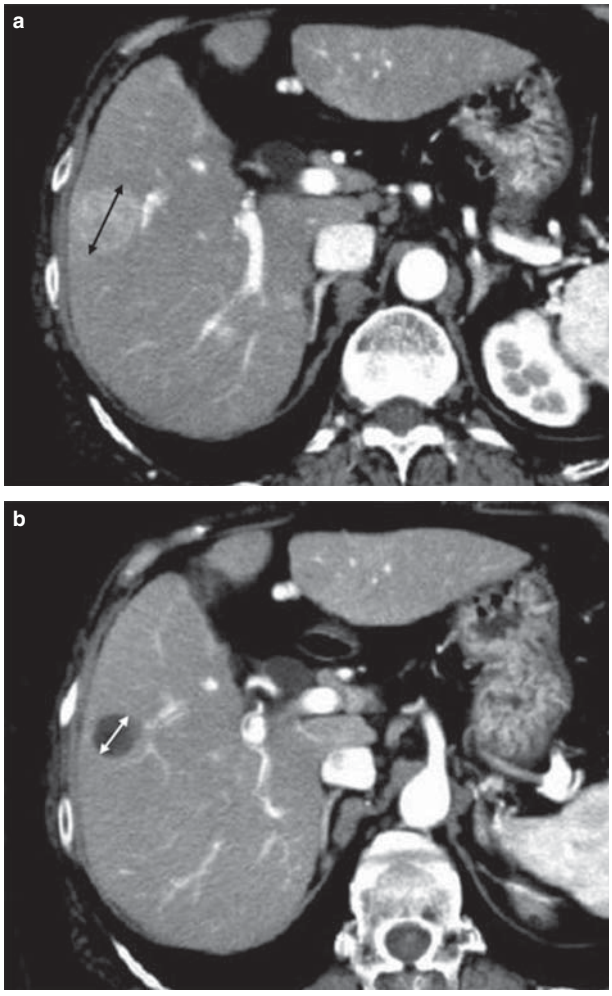

Figure 2. A: CT before TACE: Both criteria (RECIST version 1.1 and mRECIST) measured the longest diameter of the tumor. B: CT after TACE: The tumor had become entirely necrotic. The tumor response was evaluated as CR using mRECIST criteria (i.e. no measurement) and as non-CR using RECIST version 1.1 criteria (i.e. the measurement of the longest diameter of the entire tumor).

In our inter- and intra-observer reproducibility study, the kappa value and the proportion of agreement using mRECIST criteria ('almost perfect agreement') were higher than those for RECIST version 1.1 criteria ('substantial agreement'). In consideration of the high inter- and intra-observer reproducibility, mRECIST can be more recommended for use as tumor response criteria in clinical trials of TACE for HCC.

The present study had several limitations. The number of patients was relatively small, and the analyses were performed not on a per-patient basis, but on a per-lesion basis. To investigate which set of criteria was superior as tumor response criteria in clinical trials of TACE for HCC, the observer consistency study (inter- and intra-observer reproducibility between the two updated sets of criteria) were investigated in this study. A validation study comparing the updated criteria to the gold standard (i.e. overall survival) should be encouraged in future studies.

In conclusion, considering the differences in the CR rate and the response rate between RECIST version 1.1 and mRECIST criteria, close attention must be paid to the criteria used for a precise interpretation of the tumor response outcome. Furthermore, mRECIST criteria may be more suitable for tumor response criteria in clinical trials of TACE for HCC, compared with RECIST version 1.1 criteria, from the viewpoint of the high inter- and intra-observer reproducibility.

Acknowledgements

This study was undertaken as JIVROSG-0602. A part of this study was shown as a poster presentation at the meeting of the Cardiovascular and Interventional Radiological Society of Europe, Lisbon 2009.

Declaration of interest: This work was supported by the Grant-in-Aid for Cancer Research from the Japanese Ministry of Health, Labour and Welfare (20-15). The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1. 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.

2. Lencioni R, Llovet JM. Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. Semin Liver Dis. 2010;30:52–60.
3. 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.
4. Bruix J, Sherman M, Llovet JM, Beaugrand M, Lencioni R, Burroughs AK, et al. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. J Hepatol. 2001;35:421–30.
5. Varela M, Real MI, Burrel M, Forner A, Sala M, Brunet M, et al. Chemoembolization of hepatocellular carcinoma with drug eluting beads: efficacy and doxorubicin pharmacokinetics. J Hepatol. 2007;46:474–81.
6. Sala M, Llovet JM, Vilana R, Bianchi L, Solé M, Ayuso C, et al. Initial response to percutaneous ablation predicts survival in patients with hepatocellular carcinoma. Hepatology. 2004;40:1352–60.
7. Llovet JM, Bisceglie AD, Bruix J, Kramer BS, Lencioni R, Zhu AX, et al. Design and endpoints of clinical trials in hepatocellular carcinoma. J Natl Cancer Inst. 2008;100: 698–711.
8. Forner A, Ayuso C, Varela M, Rimola J, Hessheimer AJ, de Lope CR, et al. Evaluation of tumor response after locoregional therapies in hepatocellular carcinoma. Are response evaluation criteria in solid tumors reliable? Cancer. 2009;115:616–23.
9. Edeline J, Boucher E, Rolland Y, Vauléon E, Pracht M, Perrin C, et al. Comparison of tumor response by response evaluation criteria in solid tumors (RECIST) and modified RECIST in patients treated with sorafenib for hepatocellular carcinoma. Cancer. 2012;118:147–56.
10. Landis JR, Koch GG. The measurement of observer agreement for caterogical data. Biometrics. 1977;33:159–74.
11. Watanabe H, Kunitoh H, Yamamoto S, Kawasaki S, Inoue A, Hotta K, et al. Effect of the introduction of minimum lesion size on interobserver reproducibility using RECIST guidelines in non-small cell lung cancer patients. Cancer Sci. 2006;97: 214–18.
12. Al-Aynati M, Chen V, Salama S, Shuhaibar H, Treleaven D, Vincic L. Interobserver and intraobserver variability using the Fuhrman grading system for renal cell carcinoma. Arch Pathol Lab Med. 2003;127:593–6.
13. Hagen PJ, Hartmann IJ, Hoekstra OS, Stokkel MP, Postmus PE, Prins MH. Comparison of observer variability and accuracy of different criteria for lung scan interpretation. J Nucl Med. 2003;44:739–44.
14. Travis WD, Gal AA, Colby TV, Klimstra DS, Falk R, Koss MN. Reproducibility of neuroendocrine lung tumor classification. Hum Pathol. 1998;29:272–9.