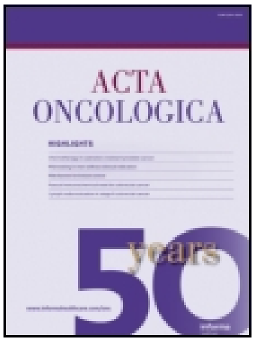

Acta Oncologica

Taylor & Francis
Taylor & Francis Group

ISSN: 0284-186X (Print) 1651-226X (Online) Journal homepage: informahealthcare.com/journals/ionc20

Observer variability in a phase II trial – assessing consistency in RECIST application

Kristin Skougaard, Mark James Dusgaard McCullagh, Dorte Nielsen, Helle Westergren Hendel, Benny Vittrup Jensen & Helle Hjorth Johannesen

To cite this article: Kristin Skougaard, Mark James Dusgaard McCullagh, Dorte Nielsen, Helle Westergren Hendel, Benny Vittrup Jensen & Helle Hjorth Johannesen (2012) Observer variability in a phase II trial – assessing consistency in RECIST application, Acta Oncologica, 51:6, 774-780, DOI: 10.3109/0284186X.2012.667149

To link to this article: https://doi.org/10.3109/0284186X.2012.667149

View supplementary material

Published online: 21 Mar 2012.

Submit your article to this journal

Article views: 1066

View related articles

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://informahealthcare.com/action/journalInformation?journalCode=ionc20

Acta Oncologica, 2012; 51: 774–780

informa
healthcare

ORIGINAL ARTICLE

Observer variability in a phase II trial – assessing consistency in RECIST application

KRISTIN SKOUGAARD[1], MARK JAMES DUSGAARD MCCULLAGH[2], DORTE NIELSEN[1], HELLE WESTERGREN HENDEL[3], BENNY VITTRUP JENSEN[1] & HELLE HJORTH JOHANNESEN[2]

[1]*Department of Oncology, Copenhagen University Hospital Herlev, Denmark,* [2]*Department of Radiology, Copenhagen University Hospital Herlev, Denmark, and* [3]*Department of Clinical Physiology and Nuclear Medicine, Copenhagen University Hospital Herlev, Denmark*

Abstract

Objective. To assess the consistency of Response Evaluation Criteria in Solid Tumours (RECIST) application in a phase II trial. *Material and methods*. Patients with metastatic non-resectable colorectal cancer treated with a combination of an antibody and a chemotherapeutic drug, were included. Computed tomography (CT) scans (thorax, abdomen and pelvis) were performed at baseline and after every fourth treatment cycle. RECIST was intended for response evaluation. The scans were consecutively read by a heterogeneous group of radiologists as a part of daily work and hereafter retrospectively reviewed by a dedicated experienced radiologist. Agreement on best overall response (BOR) between readers and reviewer was quantified using κ-coefficients and the discrepancy rate was correlated with the number of different readers per patient using a χ^2-test. *Results*. One hundred patients with 396 CT scans were included. Discrepancies between the readers and the reviewer were found in 47 patients. The majority of discrepancies concerned the application of RECIST. With the review, BOR changed in 17 patients, although, only in six patients the change was potentially treatment altering. Overall, the κ-coefficient of agreement between readers and reviewer was 0.71 (good). However, in the subgroup of responding patients the κ-coefficient was 0.21 (fair). The number of patients with discrepancies was significantly higher with three or more different readers per patient than with less ($p = 0.0003$). *Conclusion.* RECIST was not consistently applied and the majority of the reader discrepancies were RECIST related. Post review, 17 patients changed BOR; six patients in a potentially treatment altering manner. Additionally, we found that the part of patients with discrepancies increased significantly with more than three different readers per patient. The findings support a peer-review approach where a few dedicated radiologists perform double blinded readings of all the on-going cancer trial patients' CT scans.

Consistency in the production of trial results is necessary for meaningful comparison with other trials [1,2]. Radiological response evaluation is an essential component in the calculation of the trial results and therefore has to be carried out in a standardised and reproducible manner [2,3]. In 2000 a set of response criteria designed for clinical cancer treatment trials, the Response Evaluation Criteria in Solid Tumours (RECIST) [4], was published and in 2009 the revised version RECIST 1.1 was issued [5,6]. These are anatomical criteria based on tumour shrinkage or growth indicating response or non-response to treatment. Despite debate on various pitfalls in the construction, purpose and use of RECIST [7–12] consensus to use RECIST arose, and they are now predominant internationally [2,7,13,14]. A handful of studies have focussed on inter- and intra-observer variability in the use of RECIST [1–3,15], while only few of these investigate how consistently RECIST is used in practice [2,3]. Although response evaluation is often carried out as part of the daily work by numerous radiologists with widely varying levels of expertise, it is often implicitly assumed that the response evaluation is accurate and uniform [3]. A lot of faith is placed in the reported response rates from clinical phase II trials, but the conditions under which these

Correspondence: K. Skougaard, Department of Oncology, 54B1, Copenhagen University Hospital Herlev, Herlev Ringvej 75, 2730 Herlev, Denmark. Tel: +45 38689171. E-mail: kristinskougaard@dadlnet.dk

(Received 13 December 2011; accepted 2 February 2012)

ISSN 0284-186X print/ISSN 1651-226X online © 2012 Informa Healthcare
DOI: 10.3109/0284186X.2012.667149

response rates are obtained are rarely questioned. We believe, that whether the response evaluations are produced by a small group of dedicated radiologists, or as part of the daily routine by a large and heterogeneous group of radiologists, influences trial outcomes [3] as well as patients' treatment courses. In this paper we conducted a review of the computed tomography (CT) response evaluation of patients enrolled in a clinical phase II trial and thereby assessed the consistency of RECIST application. Furthermore, we estimated the inter-observer variability's influence on the trial outcome and the patients' treatment courses.

Material and methods

During 2006–2009 patients with metastatic non-resectable colorectal cancer referred to the Department of Oncology, Copenhagen University Hospital Herlev, Denmark for participation in a phase II trial were prospectively included. The patients were treated every second week with a monoclonal antibody and a chemotherapeutic drug as a third line therapy. Protocol details are outlined in Box 1.

The patients were scanned between 1–14 days before the first treatment and after every fourth treatment. In the present study, only patients that had completed four series of treatment and the first follow-up scan were included.

The scans were standard diagnostic CT examinations with both oral and intravenous contrast, covering the region of the thorax, abdomen and pelvis and were performed according to local standard guidelines. Iodinated contrast agent (Omnipaque 350 GE Healthcare, Oslo, Norway or Optiray 320 Covidien, Neustadt/Donau, Germany or Iomeron 350 Bracco, Milan, Italy) was given orally: 20 ml in 500 ml bottled water (4% solution) 30 min before CT, and intravenously: 100 ml with an injection flow of 5 ml/s. CT adaptation acquisition delay was 85 s. The majority of the patients was scanned on a helix dual-slice positron emission tomography (PET)/CT scanner (Philips GEMINI PET/CT, Philips Medical Systems, Cleveland, Ohio, USA) with 2 × 5 mm collimation (scan slice thickness 5 mm) and a minor part was scanned on a helix 16 slice CT scanner (Philips MX 8000 IDT, Philips Medical Systems, Cleveland, Ohio, USA) with 16 × 1.5 mm collimation, reconstructed to a scan slice thickness of 5 mm.

CT parameters varied between 140 kV, 150 mAs and 120 kV, 230 mAs due to tube limitations in the scanning system. The PET part of the scans was for experimental use only and the oncological decisions to continue or end treatment were based solely on the CT scan readings.

Each CT scan was single-read by an on-duty radiologist as part of the daily routine. Because the prospective reading of the trial-patients' CT scans was a part of the daily workload, it was performed by randomly different radiologists, e.g. in a patient with five scans (baseline and four follow-up scans) up to five different radiologists could have been involved: one reading the baseline scan, another reading the first follow-up scan, yet another reading the second follow-up scan, and so on. The scans were read on standard diagnostic workstations with high resolution displays and measurements were made with an electronic ruler and saved in the PACS system. The radiologists were all familiar with the trial protocol statement that RECIST [4] was to be prospectively applied for response evaluation. The PET and the CT scans were described separately in the nuclear medicine department and the radiology department, respectively. Thereafter, a joint conclusion, containing both convergent and divergent findings, was performed. The radiologists involved in the reading of the trial patients' scans are referred to as the readers. Among them was the reviewer; a dedicated radiologist specialised in onco-radiology with more than five years of experience. As the patients went off-study, she reviewed all CT scans, including her own, on equal terms: blinded to the initial descriptions, to the clinical decisions derived from them and to the PET results. Post-review, the two sets of CT descriptions (one from the initial reader and one from the reviewer) were compared and discrepancies registered. The procedure is depicted in Figure 1. Discrepancies were defined as differences between the initial prospective evaluation and the

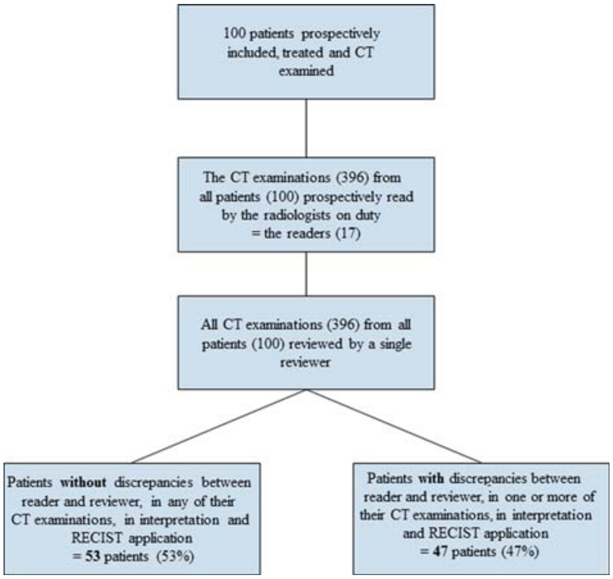
100 patients prospectively included, treated and CT examined
The CT examinations (396) from all patients (100) prospectively read by the radiologists on duty = the readers (17)
All CT examinations (396) from all patients (100) reviewed by a single reviewer
Patients **without** discrepancies between reader and reviewer, in any of their CT examinations, in interpretation and RECIST application = **53** patients (53%)
Patients **with** discrepancies between reader and reviewer, in one or more of their CT examinations, in interpretation and RECIST application = **47** patients (47%)

Figure 1. Study flow chart. Flow chart showing how the CT scans (396) of all the patients (100) were prospectively interpreted by the readers, while the study was on-going, and then reviewed by the reviewer after the study was closed.

reviewer's evaluation. The encountered discrepancies were described and divided into three categories (Table I):

1. Incorrect use of RECIST in selection of target lesions.
2. Discrepancies in measuring tumour (2.1 RECIST incorrectly or insufficiently applied and 2.2 Discrepancies in measuring tumour size).
3. Discrepancies in describing new or disappeared lesions.

Category 1 and 2.1 are RECIST related discrepancies whereas category 2.2 and 3 are discrepancies that could be present even though RECIST was applied correctly and defined as non-RECIST related.

The patients' treatment responses were calculated according to RECIST, dividing them into the four categories: complete response (CR), partial response (PR), stable disease (SD) or progressive disease (PD) [4]. This was done before and after the review. The best overall response (BOR) was registered for each patient before and after the review. The BOR distribution after the review was used for the final trial response rate calculation (Box 2).

Although recommended by RECIST [4], it was decided not to perform confirmation scans one month after the criteria for PR were first met. Patients with PR who still showed PR on the following CT scan four treatments later were considered as having confirmed PR (PRc). Patients with only one scan showing PR were considered as having unconfirmed PR (PRu).

Statistics

The degree of agreement between the readers and the reviewer was calculated using κ-coefficients. The part of the discrepancies between readers and reviewer concerning RECIST application was calculated in percentages. The correlation between the number of patients with CT discrepancies (Box 2) and the number of CT scans per patient was likewise calculated in percentages. The correlation between number of patients with CT discrepancies

Table I. Categorisation of discrepancies between readers and reviewer. BOR: best overall response.

	All patients		Subgroup of responding patients	
Categories of discrepancies	Patients with discrepancies (n = 47)	Patients with discrepancies and changed BOR (n = 17)	Patients with discrepancies (n = 17)	Patients with discrepancies and changed BOR (n = 12)
1. Incorrect use of RECIST in selection of target lesions				
Choosing truly non-measurable lesions as target lesions	14	6	5	5
Target lesions not representative for overall tumour burden	14	3	4	2
Not choosing any target lesions	7	3	3	3
2. Discrepancies in tumour measurements				
2.1 RECIST incorrectly or insufficiently applied				
Measuring shortest tumour diameter	4	3	1	1
Stops measuring target lesions or choosing new target lesions in the middle of the treatment course	4	0	1	0
Not describing the non-target lesions	3	1	0	0
2.2 Discrepancies in measuring tumour size				
Changing initial measurements in the middle of the treatment course	1	0	1	1
Measuring target lesions significantly too short or too long	4	1	1	1
Measuring one lesion as two separate and vice versa	3	0	1	1
Mistaking target lesions with each other	4	0	2	0
3. Discrepancies describing new or disappeared lesions				
Not seen new lesions seen by reviewer	6	2	1	1
Stating presence of new lesion not found by reviewer	4	1	1	1
Stating disappearance of lesion not agreed on by reviewer	5	1	4	2
Total numbers of significant discrepancies	73	21	25	18
RECIST related discrepancies (1. and 2.1)	46 (**63%**)	16 (**76%**)	14 (**56%**)	11 (**61%**)
Non-RECIST related discrepancies (2.2 and 3.)	27 (**37%**)	5 (**24%**)	11 (**44%**)	7 (**39%**)

and number of different readers per patient (two groups: three or less readers and more than three readers per patient) was calculated using a χ^2-test and an odds ratio (OR) with confidence intervals (CI). Level of significance was $p < 0.05$.

Results

One hundred patients (64 males, 36 females, median age: 63 years, 25–75 interquartile range: 58–70 years) were included and 396 CT scans were performed, read and reviewed. Eighty-one patients were PET/CT scanned (321 scans) and 19 patients were CT scanned (75 scans) throughout their treatment course. The median number of CT scans per patient was four (interquartile range: two to five). Disease involvement was seen in: liver, abdominal lymph nodes, lungs, peritoneum (carcinomatosis), rectum, bones and the spleen.

The CT readings were performed by 17 different radiologists (the readers): 13 specialists and four supervised residents. Four of the specialists had more than five years of onco-radiological experience.

The encountered discrepancies are listed in Table I. In 47 patients (47%) a total of 73 discrepancies were registered; 46 (63%) RECIST-related (Table I, category 1 and 2.1) and 27 (37%) non-RECIST related (Table I, category 2.2 and 3). In 27 patients one discrepancy was encountered, in 14 patients two discrepancies were encountered and in six patients three discrepancies were encountered. Of the discrepancies encountered among the 17 patients with changed BOR, 76% were RECIST related and 24% were non-RECIST related (Table I). In the subgroup of responding patients, the review gave rise to discrepancies in 17 patients, 56% were RECIST related and 44% were non-RECIST related (Table I). BOR changed in 12 of the responding patients and 61% of the encountered discrepancies in these patients were RECIST related whereas 39% were non-RECIST related (Table I). An example of a discrepancy is illustrated in Figure 2. Overall, there was agreement regarding BOR in 83 patients (83%) (bold writing Table II) and disagreement in the remaining 17 patients (17%). The corresponding κ-coefficient was 0.71 and strength of agreement therefore good. The BOR of the 17 patients changed almost equally in both directions: eight patients to a less favourable response and nine patients to a more favourable response (Table III) resulting in a response rate of 21% (21 patients) before and 19% (19 patients) after the review (Table II). In the subgroup of responding patients, there was agreement on BOR in 13 of 25 patients (52%) (bold writing Table IV) and disagreement in the remaining 12 patients (48%). The corresponding κ-coefficient was 0.21 and strength of agreement in this subgroup therefore fair. According to the review, five patients (5%) could have continued treatment as their last scan showed either SD or PR instead of PD as initially read (Table III, PRu to PD: 3 and SD to PD: 2). However, two of these five patients went off-study due to worsening of their clinical performance status and not due to their CT scan response. Furthermore, three patients (3%) could, according to the review, have continued treatment for longer (Table III, PD to SD: 3). No patients had CR.

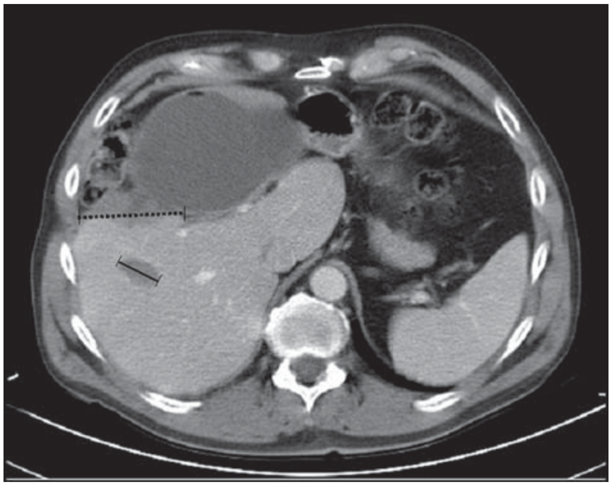

Figure 2. Example of a discrepancy between a reader and the reviewer. The reader chose a radio frequency (RF) ablation cavity as a target lesion (solid line) mistaking it for a measurable liver metastasis. The region in the rim of the large RF cavity is tumour tissue (dotted line). Differentiating tumour tissue from a RF cavity can be very difficult and requires training and sufficient information of previous treatment from the referring physician. The patient changed BOR from SD before the review to PR after the review.

The number of patients with discrepancies increased with increasing number of CT scans from an average of 39% with 2–5 CT scans to an average of 71% with 6–10 CT scans per patient (Table V). Likewise, the number of patients with discrepancies

Table II. Agreement (bold writing) between readers and reviewer on best overall response (BOR) of all patients.

	Response: Reviewer				
Response: Readers	PRu	PRc	SD	PD	Total: Readers
PRu	**2**	2	1	3	8
PRc	0	**11**	2	0	13
SD	1	3	**53**	2	59
PD	0	0	3	**17**	20
Total: Reviewer	3	16	59	22	**100**

κ-coefficient: 0.71

No patients had complete response (CR)

PD: progressive disease, PRc: confirmed partial response, PRu: unconfirmed partial response, SD: stable disease.

Table III. Changes in BOR after the review.

Negative BOR changes	Positive BOR changes
PRc to SD: 2	PRu to PRc: 2
PRu to PD: 3	SD to PRc: 3
PRu to SD: 1	SD to PRu: 1
SD to PD : 2	**PD to SD: 3**

Theoretically, five patients could have stopped treatment earlier and three patients could have continued treatment longer (bold writing).

increased with increasing number of different readers involved per patient from 0% with one reader (by chance, the same reader read all the patients scans) to 100% with six different readers (by chance, patients with six scans had a different radiologist describing each of their scans and patients with more than six scans had by chance, a few of the six different readers describing two or more of their scans) (Table VI). The χ^2-correlation between more than three readers and number of patients with discrepancies was significant with a p-value of 0.0003 (χ^2-value = 13.34) and an OR of 5.4 (95% CI 2.1–13.9). The cut-off at three readers was chosen from the distinct shift in the figures after more than three readers (Table VI).

Discussion

We assessed the consistency in application of RECIST by describing and quantifying the inter-observer variability and estimated the influence of the BOR disagreements on the trial response rate and on the individual patients' treatment courses. One or several discrepancies were found by the reviewer in 47% of the patients and the majority of the discrepancies were RECIST related. The overall agreement on BOR was good, while in the subgroup of responding patients discrepancies were however more pronounced and agreement only slightly better than chance. Even though BOR changed in 17% of the patients, the review only vaguely influenced the overall response rate as the responses changed almost equally in both directions. Still, six patients could theoretically have had a different treatment course. Of great interest, we additionally found that the part of patients with discrepancies increased significantly with more than three different readers per patient.

Table IV. Agreement (bold writing) between readers and reviewer on best overall response (BOR) of responding patients (PRc + PRu).

	Response: Reviewer				
Response: Readers	PRu	PRc	SD	PD	Total: Readers
PRu	**2**	2	1	3	8
PRc	0	**11**	2	0	13
SD	1	3	**0**	0	4
PD	0	0	0	**0**	0
Total: Reviewer	3	16	3	3	**25**

κ-coefficient: 0.21.

No patients had complete response (CR).

PD: progressive disease, PRc: confirmed partial response, PRu: unconfirmed partial response, SD: stable disease.

Table V. Number of CT scans per patient correlated to part of patients with discrepancies.

CT scans per patient	All patients no.	Patients with discrepancies no. (%)
2	7	3 (43)
3	22	8 (36)
4	31	13 (42)
5	14	5 (36)
6	9	5 (56)
7	6	5 (83)
8	6	4 (67)
9	3	3 (100)
10	2	1 (50)

The part of patients with discrepancies increased with increasing number of scans per patient.

The unchanged overall response rate in our study contrasts with other central reviews of phase II trials, where a reduction in response rate after review is usually found [3;14]. However, the level of disagreement and the types of discrepancies between observers is consistent with other studies [1–3;14;16;17].

It is a limitation of our study that the reviewer was also a reader, because this minor intra-observer element makes the inter-observer study irregular. Yet, inter-observer variability is usually reported larger than intra-observer variability [1;2;15] presumably resulting in limited influence on our study outcome. On the other hand, our study genuinely reflects daily work routines.

We believe the high rate of patients with changed BOR as well as the high discrepancy rate within the

Table VI. Number of different readers per patient correlated to part of patients with discrepancies.

Different readers per patient	Patients with discrepancies no. (total)	%
1	0 (4)	0
2	12 (29)	43
3	12 (36)	33
4	11 (15)	73
5	10 (14)	71
6	2 (2)	100

The part of patients with discrepancies increased with increasing number of different readers per patient and the number was significantly higher with more than three different readers per patient: p-value 0.0003 (χ^2-value = 13.34) and OR 5.4 (95% CI 2.1–13.9), than with less than three.

group of responding patients can be explained in part by the lack of consensus among the radiologists to consistently apply RECIST and in part by the management of these examinations as a part of daily workflow in the radiology department rather than for instance separating trial examinations from routine examinations. This is emphasised by the significantly increased number of discrepancies with increasing number of readers. A way to limit the number of readers and ensure a more consistent application of RECIST is the implementation of an internal, prospective, peer review system: blinded, routine double reading by a few (up to three) onco-radiologists of all radiological studies of trial patients while they are on-going. There are numerous ways to set up an internal peer review system [1;3;17], but the common purposes are: education, error reduction, competence ensuring and quality improvement [18–21]. The review process must be as fair and unbiased as possible, have a minimal effect on work flow and – very essentially – be non-punitive [19;22]. The peer-review system could be widened to include other imaging criteria where consistent evaluation and reporting is desirable. Nevertheless, creating a culture where all employees are comfortable with the displaying and registration of their errors and disagreements is a prolonged process and the consequential positive improvements must be evident [23]. Furthermore, for optimal radiological performance, regularly updated education on measurement and response evaluation criteria is important [2] as well as a clearly emphasising that the chosen trial-specific criteria should be applied consequently. In this case, it could lead to a possible elimination of the RECIST related discrepancies. Conversely, discrepancies that are not RECIST related will, to some degree, always be present although RECIST is followed stringently [24;25], as they are caused by different perceptions and doubt; normal human variation.

Even though the change in the fraction of responding patients before and after the review was small, this retrospective review was beneficial for the study as it improved the quality of the outcome data. Nevertheless, on an individual patient basis the changes in BOR were considerable and the patients did not benefit from the review as it was performed retrospectively. However, the number of patients that potentially could have had a different treatment course was relatively small. By reviewing the CT scans prospectively instead of retrospectively the risk of under- or over-treatment of the patients is presumably reduced. Not to deprive patients an effective treatment that might be their last option by incorrectly stating PD or to continue an ineffective treatment with considerable adverse effects by failing to state PD, is a primary interest in cancer care and it is equally important in order to spend the considerable funds used on anti-cancer drugs optimally. Recent protocols have implemented RECIST 1.1 [5] and thereby have the opportunity of using FDG-PET as an assisting tool in the identification of new lesions, presumably enhancing consensus on PD due to new lesions. Moreover, clinical relevant information from the referring clinicians is necessary for the radiologists to produce clinical relevant CT descriptions [11;17].

Conclusion

RECIST was not consistently applied and the main part of the observer variability was RECIST related. The review's influence on the trial outcome was notable as it changed the BOR of 17 patients; six patients in a potentially treatment altering manner. Additionally, we found that the part of patients with inter-observer discrepancies increased significantly with increasing number of different readers per patient supporting a change in management of trial examinations from part of daily workflow to a peer review form, where a few onco-radiologists perform double blinded readings of the on-going cancer trial patients' CT scans. We believe the peer-review form would promote response criteria application and in doing so, markedly reduce observer variability, increase reliability and reproducibility of trial outcomes and optimise the patients' treatment courses.

Acknowledgements

We would like to thank Henrik S. Thomsen, Professor of Radiology at Copenhagen University Hospital Herlev, for his critical review of the manuscript. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

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] Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: Implications for assessment of tumor response. J Clin Oncol 2003;21:2574–82.

[2] Suzuki C, Torkzad MR, Jacobsson H, Astrom G, Sundin A, Hatschek T, et al. Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. Acta Oncol 2010;49:509–14.

[3] Thiesse P, Ollivier L, Di Stefano-Louineau D, Negrier S, Savary J, Pignard K, et al. Response rate accuracy in oncology trials: Reasons for interobserver variability. Groupe

Francais d'Immunotherapie of the Federation Nationale des Centres de Lutte Contre le Cancer. J Clin Oncol 1997; 15:3507–14.

[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. 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.

[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] Bogaerts J, Ford R, Sargent D, Schwartz LH, Rubinstein L, Lacombe D, et al. Individual patient data analysis to assess modifications to the RECIST criteria. Eur J Cancer 2009; 45:248–60.

[7] Therasse P, Eisenhauer EA, Verweij J. RECIST revisited: A review of validation studies on tumour assessment. Eur J Cancer 2006;42:1031–9.

[8] Suzuki C, Jacobsson H, Hatschek T, Torkzad MR, Boden K, Eriksson-Alm Y, et al. Radiologic measurements of tumor response to treatment: Practical approaches and limitations. Radiographics 2008;28:329–44.

[9] Michaelis LC, Ratain MJ. Measuring response in a post-RECIST world: From black and white to shades of grey. Nat Rev Cancer 2006;6:409–14.

[10] Eisenhauer EA. Response evaluation: Beyond RECIST. Ann Oncol 2007;18(Suppl 9):ix29–32.

[11] 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.

[12] Jaffe CC. Measures of response: RECIST, WHO, and new alternatives. J Clin Oncol 2006;24:3245–51.

[13] 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.

[14] Therasse P. Measuring the clinical response. What does it mean? Eur J Cancer 2002;38:1817–23.

[15] Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, Tenhave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol 1996;167:851–4.

[16] Belton AL, Saini S, Liebermann K, Boland GW, Halpern EF. Tumour size measurement in an oncology clinical trial: Comparison between off-site and on-site measurements. Clin Radiol 2003;58:311–4.

[17] Ford R, Schwartz L, Dancey J, Dodd LE, Eisenhauer EA, Gwyther S, et al. Lessons learned from independent central review. Eur J Cancer 2009;45:268–74.

[18] Soffa DJ, Lewis RS, Sunshine JH, Bhargavan M. Disagreement in interpretation: A method for the development of benchmarks for quality assurance in imaging. J Am Coll Radiol 2004;1:212–7.

[19] Mahgerefteh S, Kruskal JB, Yam CS, Blachar A, Sosna J. Peer review in diagnostic radiology: Current state and a vision for the future. Radiographics 2009;29:1221–31.

[20] Halsted MJ. Radiology peer review as an opportunity to reduce errors and improve patient care. J Am Coll Radiol 2004;1:984–7.

[21] Lee JK. Quality – a radiology imperative: Interpretation accuracy and pertinence. J Am Coll Radiol 2007;4:162–5.

[22] FitzGerald R. Radiological error: Analysis, standard setting, targeted instruction and teamworking. Eur Radiol 2005; 15:1760–7.

[23] Swensen SJ, Johnson CD. Radiologic quality and safety: Mapping value into radiology. J Am Coll Radiol 2005;2: 992–1000.

[24] Abujudeh HH, Boland GW, Kaewlai R, Rabiner P, Halpern EF, Gazelle GS, et al. Abdominal and pelvic computed tomography (CT) interpretation: Discrepancy rates among experienced radiologists. Eur Radiol 2010; 20:1952–7.

[25] Goddard P, Leslie A, Jones A, Wakeley C, Kabala J. Error in radiology. Br J Radiol 2001;74:949–51.

Supplementary material available online

Supplementary Box 1 to be found online at http://informahealthcare.com/doi/abs/10.3109/0284186X.2012.667149