VOLUME 27 · NUMBER 19 · JULY 1 2009

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Evaluation of the Optimal Number of Lesions Needed for Tumor Evaluation Using the Response Evaluation Criteria in Solid Tumors: A North Central Cancer Treatment Group Investigation

Shauna L. Hillman, Ming-Wen An, Michael J. O'Connell, Richard M. Goldberg, Paul Schaefer, Jan C. Buckner, and Daniel J. Sargent

From the Mayo Clinic and Mayo Foundation, Rochester, MN; Allegheny Cancer Center, Pittsburgh, PA; The University of North Carolina at Chapel Hill, Chapel Hill, NC; and the Toledo Community Hospital Oncology Program Community Clinical Oncology Program, Toledo, OH.

Submitted May 27, 2008; accepted January 23, 2009; published online ahead of print at www.jco.org on May 4, 2009.

This study was conducted as a collaborative trial of the North Central Cancer Treatment Group and Mayo Clinic and was supported in part by Public Health Service Grant No. CA-25224.

Presented in part in oral format at the 39th Annual Meeting of the American Society of Clinical Oncology, Chicago, IL, May 31 to June 3, 2003.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Shauna Hillman, MS, Mayo Clinic, 200 First St SW, Rochester, MN 55905; e-mail: hillman.shauna@mayo.edu.

0732-183X/09/2719-3205/$20.00

DOI: 10.1200/JCO.2008.18.3269

ABSTRACT

Purpose
In February 2000, the criteria for measuring tumor shrinkage as an indicator of antitumor activity were redefined by the Response Evaluation Criteria in Solid Tumors (RECIST). This resulted in simplifying bidimensional to unidimensional measurement of lesions. Under RECIST, all lesions, up to 10, must be measured. Scanning and measuring multiple lesions is costly, time-consuming, and a disincentive to participation in clinical trials. We investigated whether fewer than 10 lesions can be measured without compromising the accuracy of assessing a regimen's activity.

Patients and Methods
Thirty-two North Central Cancer Treatment Group trials including 2,374 patients were analyzed. Twelve studies were conducted before RECIST; 20 were conducted post-RECIST. Agreement between objective status by cycle, confirmed response, overall response rate, and time to progression (TTP) was evaluated based on all 10 versus the largest one through five lesions.

Results
The median number of lesions reported on RECIST trials did not differ from pre-RECIST trials (median = 2.0). One lesion at baseline was reported in 49% of patients, two lesions in 28% of patients, three lesions in 12% of patients, four lesions in 6% of patients, and five lesions in 5% of patients in post-RECIST trials. Utilizing the largest two lesions produced excellent concordance with that using all lesions for all end points. In no trial did the overall response rate differ by more than 3% when two versus all lesions were considered. Evaluating more than two lesions did not significantly improve agreement.

Conclusion
Based on these trials, the assessment of more than two lesions did not alter the conclusions regarding a treatment's efficacy as judged by response rate or TTP.

J Clin Oncol 27:3205-3210. © 2009 by American Society of Clinical Oncology

INTRODUCTION

Anticancer cytotoxic agents are often evaluated for antitumor activity by measuring tumor shrinkage. In the late 1970s, the International Union Against Cancer and the WHO introduced specific criteria for the codification of tumor response evaluation. Over time, various groups developed diverging criteria for the assessment of tumor response. In 1994, the European Organisation for Research and Treatment of Cancer (EORTC), the National Cancer Institute (NCI) of the United States, and the National Cancer Institute of Canada Clinical Trials Group (NCIC CTG) established a task force—the Response Evaluation Criteria in Solid Tumors (RECIST) —to review existing criteria for evaluating response in solid tumors resulting in a consensus document.[1] RECIST is intended for trials where tumor response is the primary end point, helps to further standardize definitions and methodology, and simplifies data collection by eliminating the need for bidimensional measurements. RECIST is also an essential element of clinical trials assessing time to tumor progression (TTP), as in those trials, a progression event is typically defined using RECIST.

One element of RECIST was a standardization of the number of lesions per patient required to be evaluated. Under RECIST, all lesions, up to a maximum of 10 lesions with a maximum of five per organ, must be evaluated and reported. We are

unaware of data to support the choice of 10. This requirement raises multiple issues in the practical conduct of clinical trials. Specifically, there is a concern that payers may not be willing to reimburse for the scans necessary to follow up to 10 lesions, particularly if all scans would not otherwise be required for routine clinical care. In addition, the cost of collecting, processing, and auditing the additional data is substantial. Difficulty with compliance is also a concern, as well as the potential for a negative effect on participation rates for trials requiring the RECIST. The NCIC CTG Clinical Research Associate Committee cited tumor measurements as a major factor contributing to increased workload.[2]

Tumor shrinkage and TTP continue to be critical end points in most clinical trials. The United States Food and Drug Administration has consistently recognized tumor shrinkage as a measure of clinical activity, and has allowed this evidence to be the basis for accelerated approval of cancer drugs in some situations.[3] In addition to tumor response, TTP, or the closely related end point of progression-free survival, are being increasingly used as a clinical trial end point. These two considerations—the need to reduce the burden of clinical trials and the continued importance of tumor assessments—imply that determining the minimum number of tumor measurements that can be assessed without compromising an accurate reflection of a regimen's activity is critical.

Based on these considerations, we performed a retrospective pooled analysis of NCCTG phase II/III clinical trials. The primary aim of this analysis is to answer the question—can we assess fewer than 10 lesions without compromising an accurate reflection of a regimen's activity?

PATIENTS AND METHODS

Individual patient data on 2,374 patients were pooled from 32 trials conducted by the NCCTG open between August of 1998 and September of 2002. All trials that opened between these dates that collected radiographic measurement data were included. A single trial, trial 96-32-55, did not have adequate measurement data collected to allow inclusion.[4] All trials and all patient data collected as of September 20, 2002, were included for this analysis regardless of data maturity. Twelve trials were conducted before implementation of the RECIST (pre-RECIST), 20 were post-RECIST implementation. Trials included 12 GI trials, 10 lung trials, seven breast trials, two melanoma trials, and one mesothelioma trial. Data were largely collected from patients enrolled through NCCTG, but also included data from 925 patients enrolled to NCCTG studies through other oncology cooperative groups including, Cancer and Leukemia Group B (CALGB; 323 patients), Southwest Oncology Group (SWOG; 266 patients), Eastern Cooperative Oncology Group (ECOG; 212 patients), and the NCIC CTG (124 patients). Eighty-eight percent of the patients (n = 2,096) included in this analysis were evaluated using computer tomography (CT), 4% with chest x-ray (n = 88), 2% with physical examination (n = 56), 0.9% with MRI (n = 21), 0.2% with ultrasound (n = 4), 3% using a mix of imaging techniques (n = 75), and 1% with unknown evaluation techniques (n = 34).

For patients entered onto pre-RECIST trials, measurable disease was defined as lesions on physical examination or x-ray with clearly measurable perpendicular diameters and/or liver edge at least 5 cm below the costal margin lesions. Only measurable lesions were used for this analysis. For studies that collected bidimensional measurements, the longer of the two measurements for each lesion was used. The standard RECIST definitions were applied to determine objective and progression status.[1]

Four end points were evaluated: per patient concordance between reported objective status for each cycle; confirmed response per patient; TTP; and the overall study response rate. To evaluate agreement, each end point was compared based on all measured lesions for each patient versus the longest one to five lesions for that same patient. For example, for the first end point, objective status by cycle, the objective status (possible values of complete response [CR], partial response [PR], stable disease, and disease progression) based on all lesions reported was compared to the objective status computed using only the longest lesion. This comparison was repeated for each evaluation cycle. We performed similar comparisons of the objective status based on all lesions with objective status based on the longest two, three, four, and five lesions. Agreement was achieved when the objective status was consistent. An objective status originally reported as a PR based on all lesions but calculated as a CR when a subset of lesions was used was considered nonagreement. Any evaluations where all measurements were not available were excluded for this end point. Disease evaluations were conducted per protocol and included a variety of evaluation schedules. Common schedules included every cycle, and every other cycle starting with either cycle 1 or 2. Some protocols also require a confirmation of initial response thus changing the schedule of evaluation once an initial response is observed. The variability of these schedules results in a different number of patients evaluated for objective status for each cycle. Due to the most common schedule being every other cycle starting with cycle 2, this time point has the largest number of patients with evaluations and is therefore included as a representative example in the results section. Analyses were conducted for cycles 1 to 6 in order to identify trends over the entire course of treatment.

A response was considered confirmed if an objective status of CR or PR was reported on two consecutive evaluations. Agreement was achieved when a confirmed response was reported for a given patient based on all lesions and was also computed when only the longest one to five lesions were considered, or when a confirmed response was not reported for a given patient on all reported lesions and based on the longest one to five lesions. Since the trials analyzed were at different stages of data maturity, some patients had not yet had two evaluations, such patients were excluded form analysis (n = 267).

The cycle of progression was determined in all patients where a progression was reported in either the complete data (all lesions) or where a progression was calculated when considering only the longest one to five lesions. The cycle in which a progression was first recorded using all reported lesions was compared to the cycle in which a progression was first noted if only the longest one to five lesions were considered. Agreement was achieved when the cycle of progression was consistent. Patients were not included for this end point if they were alive at the time of the analysis without a progression reported. Any patients who died without progression or reported a progression based on the presence of new lesions or clinical deterioration were considered an agreement since the time of progression would not change based on the number of lesions reported.

To evaluate the overall study response rate, the study confirmed response rate was calculated using all reported lesions for each patient and also calculated using only the longest one to five lesions. Patients were included in this analysis if they had at least two evaluations.

A sensitivity analysis was performed in patients that reported five or more lesions. This analysis was conducted to evaluate the end points in those patients who have the greatest opportunity for differential response status based on the number of lesions considered.

RESULTS

The median number of lesions measured for pre-RECIST studies did not differ from post-RECIST studies (median = 2). In the pre-RECIST and the post-RECIST trials, 5% (n = 108) of patients had five or more lesions reported. The percentage of patients with five or more lesions did not differ by cooperative group. Figure 1 presents the distribution of number of lesions measured at baseline.

The agreement between the objective status determined using all lesions versus the objective status at each cycle determined by using only the longest one, two, or three lesions was high. For cycle 2, there

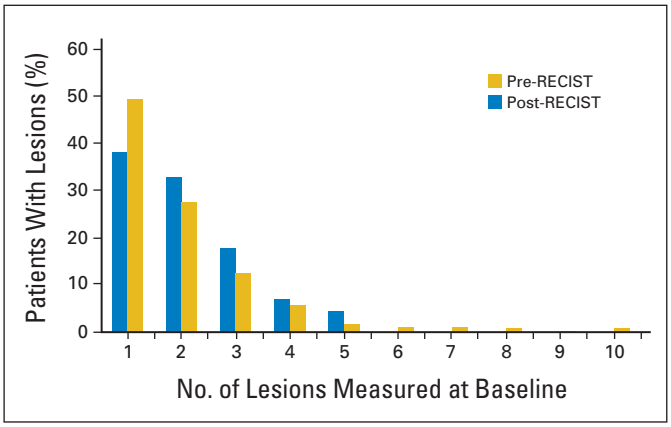

Fig 1. Number of lesions measured at baseline. RECIST, Response Evaluation Criteria in Solid Tumors.

was 93% agreement (1,164 of 1,254) using the longest lesion as compared to using all reported lesions, 98% agreement (1,241 of 1,263) using the longest two lesions, and 99.8% agreement (1,265 of 1,268) using the longest three lesions. These results were consistent over the course of treatment. (Fig 2). In order to further examine the disagreements, Table 1 provides objective status when including all lesions as compared to the longest two for treatment cycles 1 and 2 respectively. For cycle 1 data, 1% of the patients had an improved objective status when considering only the longest two lesions and 0.7% had a decline. This was similar for cycle 2 with 1.3% improved and 0.4% with a decline. We also examined the agreement rates by number of baseline lesions reported, assessment method, and disease site (Table 2). Results were similar across all factors considered.

The per patient agreement between confirmed response status when considering only the longest, longest two and longest three lesions compared to all lesions reported was also high. Considering only the longest lesion resulted in 96% agreement (2,012 of 2,103), longest two resulted in 99% agreement (2,079 of 2,104), and longest three resulted in 99.8% agreement (2,100 of 2,105), all as compared to the confirmed response status when considering all reported lesions.

Tumor progression was reported in 1,719 patients. Of these, 35% (n = 602) reported tumor progression based on new lesions or clinical deterioration, 39% (n = 670) went off therapy without progression due to treatment completion, adverse events, refusal or other reasons, but reported progression during the long-term follow-up portion of

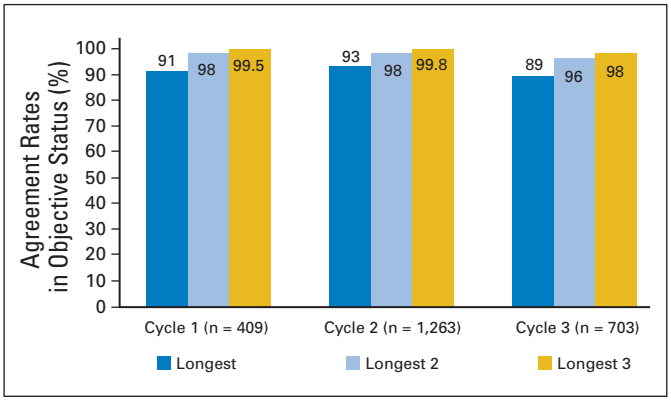

Fig 2. Agreement rates: objective status.

Table 1. Objective Status Summary

All Lesions by Cycle	Longest 2 Lesions			
	CR	PR	Stable	PROG
1 (n = 409)				
CR	3	0	0	0
PR	0	66	1	0
Stable	0	2	228	3
PROG	0	0	2	104
2 (n = 1,263)				
CR	34	0	0	0
PR	1	287	4	0
Stable	0	9	662	1
PROG	0	1	6	258

Abbreviations: CR, complete response; PR, partial response; PROG, progression.

the trial where measurements are not routinely collected and 26% (n = 447) reported progression based on tumor measurements.

The agreement rate on the cycle of disease progression based on consideration of the longest lesion, longest two lesions and longest three lesions compared to all lesions reported was also high, particularly when the longest two or three lesions were considered. Compared to all reported lesions, 91% agreement (1,552 of 1,701) was obtained based on the single longest lesion, with the longest two lesions, 97% agreement (1,665 of 1,716), and 99% agreement (1,705 of 1,718) when the longest three lesions were considered.

A large percentage of patients were reported to have a small number of lesions, thus the estimate of a rate of agreement obtained from a comparison of end points based on all lesions versus the longest one, two, or three lesions could be overly optimistic. We therefore evaluated the same end points in the subset of patients who reported five or more lesions, as this subset of 108 patients (5% of the total) has the greatest opportunity for differential response status. The metric of objective status by cycle was difficult to evaluate as small patient numbers were available for any given single cycle (Table 2). For the confirmed response and cycle of progression end points, the agreement rates remained high. The confirmed response agreement rates were 93% (83 of 89) when considering the longest single lesion, 93% (83 of 89) when considering the longest two, and 98% (87 of 89) when considering the longest three compared to all reported lesions. Regarding cycle of progression, again the agreement rates were high when considering the longest two or three lesions. The agreement rates are 82% (65 of 79) when considering the longest single lesion, 86% (68 of 79) when considering the longest two, and 89% (70 of 79) when considering the longest three lesions.

Lastly, we compared the overall study response rate that would have been estimated if for each study, the patient level response had been determined using just each patient's longest two lesions. In the 12 pre-RECIST trials, the largest absolute difference between the study level response rate using only two lesions to using all reported lesions was 1%, and the largest absolute difference in the overall study response rate for the 20 post-RECIST trials was 3%, with all but three trials showing no difference in overall response rate (Tables 3 and 4).

DISCUSSION

In this pooled analysis, we evaluated the value added as a result of considering multiple lesions per patient in the assessment of cancer

Table 2. Agreement Rates: Objective Status by Factors

Parameter	Lesion					
	Longest		Longest 2		Longest 3	
	No.	%	No.	%	No.	%
No. of baseline lesions*						
1	538/538	100.0	538/538	100.0	538/538	100.0
2	345/385	89.6	391/391	100.0	391/391	100.0
3	175/203	86.2	192/207	92.3	206/206	100.0
4	71/83	85.5	80/87	92.0	86/88	97.7
5+	35/45	77.8	42/45	93.3	44/45	97.8
Assessment method*						
CT	1,021/1,097	93.1	1,086/1,106	98.2	1,104/1,106	99.8
CXR	45/45	100.0	47/47	100.0	47/47	100.0
PE	32/35	91.4	34/35	97.1	35/35	100.0
Mix	37/44	84.1	45/48	93.8	47/48	97.9
Other†	29/33	87.9	31/32	96.9	16/16	100.0
Disease						
Breast	159/168	94.6	166/170	97.6	170/170	100.0
GI	728/785	92.7	780/793	98.4	791/794	99.6
Lung	243/264	92.0	260/268	97.0	267/267	100.0
Other	34/37	91.9	37/37	100.0	37/37	100.0

Abbreviations: CT, computed tomography; CXR, chest x-ray; PE, physical examination; Mix, mixture of imaging techniques.
*Using cycle 2 agreement data.
†Other: magnetic resonance imaging (n = 12), ultrasound (n = 4), and unknown (n = 17).

therapies in more than 2,300 patients from 32 phase II/III clinical trials. Based on our data, excellent concordance per patient is observed in key trial end points using the measurements from two lesions per patient compared to all lesions. Similarly, at the study level, assessing more than two lesions does not alter the determination of response rate. This data strongly support the consideration of a revision for RECIST. Such simplification would enhance the utility of RECIST.

Mazumdar et al[5] and Schwartz et al[6] have recently presented a theoretical approach to assessing the minimum number of tumors required for assessing treatment response. Based on their model, they proposed six lesions be measured, based on a consideration that a $\geq$ 20% increment in variability is unacceptable. We note that the selection of six lesions exceeds our recommendation of two, and this may be due to the use of strictly tumor measurements to define progression. We found a relatively low percentage of patients actually manifested progression based on tumor measurements (26%). Tumor progression was more frequently reported based on new lesions and clinical deterioration (35%). This may imply that treatment fails because of emergence of resistant clones or that it is easier to report new lesions rather than measure existing ones, unfortunately our data collection methods did not allow us to distinguish between these possibilities. It is also difficult to determine if symptomatic deterioration is intertwined with treatment toxicity as again this data was not collected. The remaining 39% of patients went off therapy without progression due to treatment completion, adverse events, or refusal. These factors contribute to the high rate of agreement specifically in TTP and overall study response rate when only two lesions are assessed compared to all lesions measured.

Table 3. Response Rates in Pre–Response Evaluation Criteria in Solid Tumors Studies

Study	Disease Group	No. of Patients	Response Rate		Absolute Difference	Disagreement	
			All Lesions	Longest 2 Lesions		No.	%
N9741	GI	1,022	31	32	1	21	2.1
983252	Breast	56	71	71	0	0	0.0
974651	GI	34	56	56	0	0	0.0
982052	Lung	29	45	45	0	0	0.0
N9841	GI	246	20	20	0	0	0.0
982452	Lung	102	19	19	0	2	2.0
984351	GI	38	13	13	0	0	0.0
983253	Breast	21	10	10	0	0	0.0
982453	Lung	26	8	8	0	0	0.0
983251	Breast	21	5	5	0	0	0.0
972451	Lung	45	0	0	0	0	0.0

NOTE. Studies with N $\geq$ 20.

Table 4. Response Rates for Post–Response Evaluation Criteria in Solid Tumors Studies

Study	Disease Group	No. of Patients	Response Rate		Absolute Difference	Disagreement	
			All Lesions	Longest 2		No.	%
N0043	GI	38	11	8	3	1	2.6
N9921	Lung	50	12	14	2	1	2.0
N9932	Breast	48	60	60	0	0	0.0
N9941	GI	46	26	26	0	0	0.0
N0026	Lung	37	14	14	0	0	0.0
N0021	Mesothelioma	21	10	10	0	0	0.0
N0022	Lung	54	6	6	0	0	0.0
N0032	Breast	20	5	5	0	0	0.0
N0031	Breast	23	4	4	0	0	0.0
N0042	GI	26	4	4	0	0	0.0
N9946	GI	35	3	3	0	0	0.0
N9975	Melanoma	27	0	0	0	0	0.0

NOTE. Studies with N ≥ 20.

Zacharia et al[7] also undertook a similar evaluation of the optimal number of lesions to be measured to assess response in patients undergoing chemotherapy for colon cancer metastases to the liver. Although their analysis was restricted to patients with colon cancer metastases to the liver and was quite small (n = 30), they also suggest that it may be possible to reduce the number of lesions measured in clinical trials. Patients utilized for the Zachareia et al analysis were recruited prospectively, while our analysis included data from actual phase II/III clinical trials. We purposely included all modes of assessment, and trials across several disease sites and cooperative groups as we sought to ensure that our analysis reflected the real data encountered when evaluating tumor response.

In a recent article, Eisenhauer summarizes the accomplishments of RECIST and outlines the areas that need attention.[8] She concluded that the minimum number of lesions to assess antitumor activity was likely fewer than the current 10 lesions, but indicated that little data was available to base a revised recommendation.[8] Our pooled analysis was undertaken in a large sample of more than 2,300 patients across 32 trials and provides the data needed to support such revisions. We found that for the majority of patients being enrolled onto these clinical trials, a relatively small number of lesions were reported (median = 2). It is unclear if this is representative of the larger population of oncology patients, or if patients who tend to have more lesions are less likely to enroll on clinical trials due to the impact of tumor burden on their performance status, are less willing to participate in clinical trials, have physicians less likely to place them on clinical trials due to the increase in workload, or have additional lesions that were not reported on the case report forms. We found that the agreement rate for the typical trial end points assessed in oncology was very high when comparing the longest two lesions to all lesions reported. These rates remain high when considering only those patients with five or more reported lesions.

There are several limitations to the analyses. The data were largely from one cooperative group. A relatively small number of patients had a large number of lesions measured. The findings are also limited to GI, breast, and lung cancer, as the majority of our data came from those tumor types. CT scanning was the modality used to assess tumor measurements most often. While these factors limit the generalizability of our conclusions, these three disease sites are among the most common sites for the testing of new agents, and are most often followed by CT scans. Another limitation is that the number of measurable lesions may potentially be under-reported in cooperative group studies. The presence and potential impact on such under-reporting cannot be assessed in this data. Thus, while our analysis cannot demonstrate the concordance between measuring a small number of lesions versus some unknown truth we feel that this data does allow the conclusion that within the context of current clinical trial practice, assessing a smaller number of lesions will not impact trial outcomes.

There have been many recent proposals for improvement in evaluation of tumor activity, including use of continuous tumor measurements;[9] however, at this time, RECIST remains the standard and therefore these results remain highly relevant. If our results can be confirmed in other data sets, RECIST guidelines should be re-evaluated to allow the observation of two lesions to provide an adequate measure of antitumor activity.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: Shauna L. Hillman, Michael J. O'Connell, Richard M. Goldberg, Daniel J. Sargent
Financial support: Jan C. Buckner, Daniel J. Sargent
Administrative support: Jan C. Buckner, Daniel J. Sargent
Provision of study materials or patients: Richard M. Goldberg, Paul Schaefer, Jan C. Buckner, Daniel J. Sargent
Collection and assembly of data: Shauna L. Hillman, Ming-Wen An, Daniel J. Sargent
Data analysis and interpretation: Shauna L. Hillman, Ming-Wen An, Jan C. Buckner, Daniel J. Sargent
Manuscript writing: Shauna L. Hillman, Ming-Wen An, Michael J. O'Connell, Richard M. Goldberg, Jan C. Buckner, Daniel J. Sargent
Final approval of manuscript: Shauna L. Hillman, Ming-Wen An, Michael J. O'Connell, Richard M. Goldberg, Paul Schaefer, Jan C. Buckner, Daniel J. Sargent

REFERENCES

1. Therasse P, Arbuck SG, Eisenhauer EA, et al: New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 92:205-216, 2000

2. Roche K, Paul N, Smuck B, et al: Factors affecting workload of cancer clinical trials: Results of a multicenter study of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol 20:545-556, 2002

3. Johnson J, Williams G, Pazadur R: End points and United States Food and Drug Administration approval of oncology drugs. J Clin Oncol 21:1404-1411, 2003

4. Perez EA, Hillman DW, Mailliard JA, et al: Randomized phase II Study of two irinotecan scheduling for patients with metastatic breast cancer refractory to an anthracycline, a taxane, or both. J Clin Oncol 22:2849-2855, 2004

5. Mazumdar M, Smith A, Debroy P, et al: A theoretical approach to choosing the minimum number of multiple tumors required for assessing treatment response. J Clin Epidemiol 58:150-153, 2005

6. Schwartz L, Mazumdar M, Brown W, et al: Variability in response assessment in solid tumors. Clinical Cancer Res 9:4318-4323, 2003

7. Zacharia TT, Saini S, Halpern EF, et al: 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, 2006

8. Eisenhauer EA: Response evaluation: Beyond RECIST. Ann Oncol 18:ix29-ix32, 2007

9. Karrison TG, Maitland ML, Stadler WM, et al: Design of phase II cancer trials using a continuous endpoint of change in tumor size: Application to a study of sorafinib and erlotinib in non-small-cell lung cancer. J Natl Cancer Inst 99:1455-1461, 2007

