Featured Article

Variability in Response Assessment in Solid Tumors: Effect of Number of Lesions Chosen for Measurement

Lawrence H. Schwartz,[1] Madhu Mazumdar, Wendy Brown, Alex Smith, and David M. Panicek

Departments of Radiology [L. H. S., W. B., D. M. P.] and Epidemiology and Biostatistics [M. M., A. S.], Memorial Sloan-Kettering Cancer Center, New York, New York 10021, and Weill Medical College of Cornell University, New York, New York [L. H. S., M. M., D. M. P.]

Abstract

Purpose: **This study was performed to systematically evaluate the variability in tumor response assessments that occurs depending on how many tumor deposits are selected for measurement at imaging.**

Experimental Design: **The two largest perpendicular diameters of all tumor deposits in 36 patients were measured on computed tomography scans obtained at baseline and first posttherapy follow-up. A computerized modeling analysis of those data was performed to determine each patient's therapeutic response category assignment for every possible number of lesions in a grouping. The variance in the sum of measurements of these lesion groupings was calculated, and the frequency of response assessment categories was plotted against the number of lesions.**

Results: **The computerized analysis of the resultant 1,833,821 possible combinations of tumor deposits showed that when six lesions were measured bidimensionally and four lesions were measured undimensionally, the average variance decreased by 90%. The number of different response assessment categories into which a patient was assigned decreased with increasing lesion grouping size. When six or more lesions were measured bidimensionally, 9% of all possible lesion groupings still fell into a second response category, reflecting the effect of which particular lesions are chosen for measurement.**

Conclusions: **Measuring larger numbers of lesions will decrease the variance. In this population, the variance decreased by at least 90% when six or more lesions were measured bidimensionally. Further confirmatory studies with larger series of patients are warranted before adopting this number as a criterion in clinical trials for assessing the activity of antineoplastic therapies.**

Introduction

Assessment of therapeutic response to chemotherapy and radiation therapy is critical in judging the success or failure of a therapy. In clinical practice, radiologists and oncologists measure lesions before and after therapy to assess response. The WHO has recommended a standardized approach to assess response, wherein the largest diameter and the largest perpendicular diameter of each selected tumor deposit are measured on CT[2] or magnetic resonance images; the diameters are multiplied for each deposit, and the resulting "cross-products" of all deposits measured are then summed (1). The sum of the tumor measurements can be compared with the sum of the same tumor deposits assessed on follow-up scans; the percentage change in the sums reflects the degree of tumor response or progression. Tumor response is further categorized based on this percentage difference (and other factors, such as appearance of new tumor deposits) as PD, SD, PR, or CR.

Recently, the RECIST Group has issued new guidelines for evaluating tumor response, in which only the largest diameter of a tumor (*i.e.*, a unidimensional measurement) is used (2). The use of unidimensional measurements has been validated by the RECIST Group (2) and others (3, 4). In addition, the percentage changes in total tumor burden that define the categories of PR, CR, and PD have been modified to reflect differences inherent in the volumes derived from unidimensional and bidimensional measurements (2).

The RECIST criteria also suggest that 5 lesions/organ and 10 lesions in total, representative of all lesions in the patient, should be measured at baseline examination (2). These 10 lesions are generally considered to be "target" lesions; other "nontarget" lesions should also be evaluated, possibly in a more subjective fashion. These target lesions are assumed to be representative of a patient's entire tumor burden and, thus, the overall disease status of the patient. Many patients have more than a total of 10 tumor deposits or 5 tumor deposits per organ site, yet fewer than 10 lesions per patient are actually recorded in many trials.

Studies have shown significant variability in therapeutic response assessment. For instance, in a French multicenter trial, major disagreements occurred in 40% of cases, and minor disagreements occurred in 10.5% of cases; the reasons for the interobserver variability included errors in tumor measurements, errors in selection of measurable targets, as well as radiological

Received 2/10/03; revised 4/8/03; accepted 6/26/03.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
This work was supported in part by the Mr. William H. Goodwin and Mrs. Alice Goodwin and the Commonwealth Cancer Foundation for Research, The Experimental Therapeutics Center of Memorial Sloan-Kettering Cancer Center, and the Byrne Foundation.
[1] To whom requests for reprints should be addressed, at Department of Radiology, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021. Phone: (212) 639-5511; Fax: (212) 794-4010; E-mail: schwartl@mskcc.org.

[2] The abbreviations used are: CT, computed tomography; CR, complete response; PD, progressive disease; PR, partial response; RECIST, Response Evaluation Criteria in Solid Tumors; SD, stable disease.

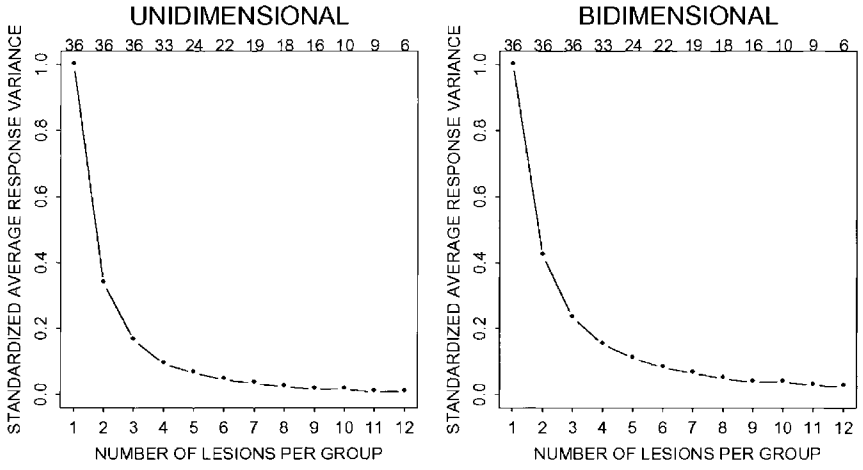

Fig. 1 Average response variance plot for unidimensional and bidimensional tumor measurements. As the number of lesions measured increases, the variance decreases for both unidimensional and bidimensional measurements.

technical issues and other diseases mistaken for metastases (5). One group of researchers has suggested that a radiologist "should measure and report a significant number of each patient's tumor sites" to assess response, without defining what constitutes that "significant" number (6).

Variability in response assessments based on the number of tumors measured has not been systematically studied; in fact, the available recommendations about the number of tumors to be measured appear to be arbitrary. The purpose of our study was to evaluate one of the several sources of variability in response assessment: how the number of tumors selected for measurement at CT affects tumor response assessment. We performed this analysis in the context of both WHO and RECIST criteria, but we did not evaluate those criteria themselves.

Patients and Methods

Patient Population

Forty patients enrolled in clinical trials and whose CT images were present in our Picture Archiving and Communication System between 1998 and 2002 were randomly selected from our patient database. The patients' CT scans were retrospectively reviewed, and four patients with fewer than 4 lesions were then excluded because all lesions would normally be measured if three or fewer were present. The study population therefore consisted of 36 solid tumor cancer patients [29 males and 7 females; average age, 59 years (range, 35–87 years)]. The primary tumors were cancers of the kidney ($n = 18$), colon ($n = 12$), prostate ($n = 2$), breast ($n = 3$), and bladder ($n = 1$). In 21 patients, the baseline and follow-up CT scans were of the chest, abdomen, and pelvis; in the remaining 15 patients, the scans were of the abdomen and pelvis. In all cases, the same body part (chest, abdomen, or pelvis) was available at baseline and follow-up. All CT scans were obtained with i.v. and oral contrast materials.

Tumor Measurements on CT Scans

All tumor sites in each patient were remeasured at baseline and at initial follow-up CT. Tumor measurements were obtained on a Picture Archiving and Communication System workstation (GE Medical Systems, Chicago, IL) with the electronic caliper tool that allows the user to draw a thin electronic line on the computer monitor. Images could be magnified and window/level settings could be adjusted at the radiologist's discretion to best display each tumor deposit. The largest perpendicular diameters were obtained for each tumor deposit and recorded by one radiologist. The radiologist measured both baseline and initial follow-up scans in the same evaluation session. Lesions were measured at baseline if the minimal perpendicular diameter was >1 cm at baseline.

Response Categories

Unidimensional (maximum diameter) percentage changes in tumor size were categorized according to the RECIST classification criteria: CR, −100%; PR, −100% to −30%; SD, −30% to 20%; and PD, ≥20%. Bidimensional (cross-product) percentage changes were categorized according to the WHO classification criteria: CR, −100%; PR, −100% to −50%; SD, −50% to 25%; and PD, ≥25%.

Data Analysis

Lesion Grouping. Response assessments were performed on a computer for all possible lesion combinations and for all possible numbers of lesions in each grouping that could be selected for measurement in each patient. For example, if a patient had four measurable lesions at baseline, then one, two, three, or all four lesions could be grouped, and their measurements summed. The total number of combinations of lesions that could be chosen, however, is greater; with four measurable lesions, there are a total of 15 combinations of lesions that could be chosen for measurement: each lesion individually (representing 4 combinations), 6 combinations of two lesions, 4 combinations of three lesions, and 1 combination of four lesions.

Ranked Response Categories. Patients were classified according to the standard tumor response categories of CR, PR, SD, and PD using the WHO criteria (Ref. 1; for bidimensional measurements) and the RECIST criteria (Ref. 2; for unidimensional measurements). To evaluate response assessments independent of those categories, a RANK category system was created. RANK1 represents the response category containing the

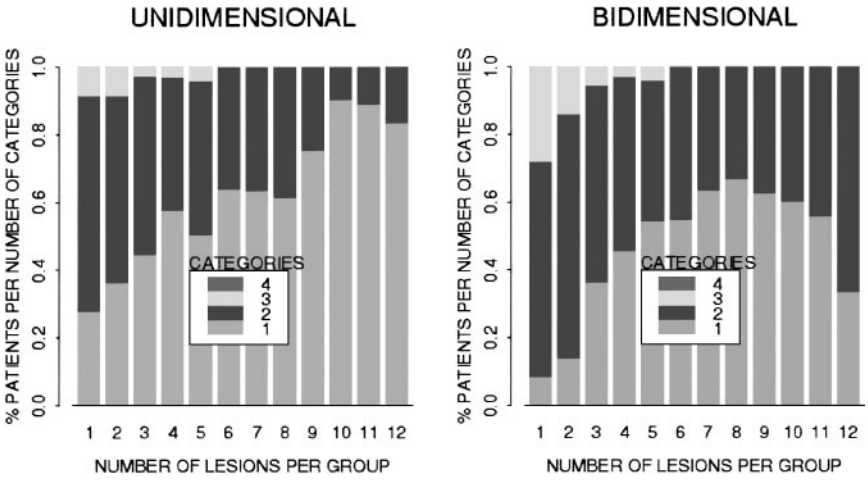

Fig. 2 Patient classification plot for unidimensional and bidimensional tumor measurements. When more than five lesions were measured (either unidimensionally or bidimensionally), no patients were categorized in more than two response categories.

greatest proportion of measurement groupings; RANK2 and RANK3 represent the categories with the second highest and the lowest proportions, respectively. None of the patients was categorized in all four categories (see "Appendix B: Ranked Response Categories").

Average Response Variance Calculation. Response categorization is based on the underlying continuous variable of percentage changes in the sum of the tumor measurements. A measure of average variance is calculated for this variable. For each patient and each choice of grouping size (k) from 1 to 12, a simple variance is calculated using the squared deviation of their mean and dividing by the number of possible combinations. An average response variance was then estimated by taking the mean of these variances across all patients (see "Appendix C: Average Variance Calculation").

Descriptive Plots

Three plots were graphed to illustrate the changes in response assessment based on the number of lesions evaluated to determine the effect of the number of lesions chosen for measurement on the robustness of response evaluation.

Variance Plot. Variance (the mean variability in response) was plotted with the X axis representing the number of lesions included in the grouping and the Y axis representing the average variance for that grouping.

Patient Classification Plot. The Y axis of this plot is the proportion of the time that the response of the 36 patients was classified in one, two, or three different response categories; the X axis is the number of lesions measured in that grouping.

Average RANK Classification Plot. The Y axis of this plot is the proportion of RANK classification, independent of the actual response category; the X axis represents the number of lesions measured in each grouping. Each patient contributes variably to this plot, depending on the number of lesions and the number of groupings that these lesions represent. This plot reflects the dominance of the most common response category.

In interpreting each of these plots, it should be noted that as the number of lesions analyzed (k) increases, fewer of the 36 patients in the study population actually contributed to those analyses. For example, when k is increased to 12, only six patients had enough measurable tumor deposits to produce one or more groupings of that size.

Results

Lesions. At baseline and again at initial follow-up, 324 lesions (mean, 9 lesions/patient; range, 4–20 lesions/patient) were measured. These tumor deposits were located in lymph nodes (n = 125), liver (n = 104), lung (n = 70), and other locations (n = 25, including in adrenal, kidney, and psoas muscle).

Lesion Groupings. Based on these 324 lesions and considering grouping sizes of up to 15 lesions, a total of 1,833,821 groupings of lesions (range, 15–1,042,379) in these 36 patients were analyzed.

As lesion number increases, the number of possible groupings of tumors increases rapidly. As greater numbers of lesions are measured and grouped together in the response assessment, variance decreases for both unidimensional and bidimensional measurements (Fig. 1).

No patients were categorized in more than two response categories when more than five lesions were measured (either unidimensionally or bidimensionally; Fig. 2). The number of patients that fell into two categories decreased as the number of lesions included in the measurement grouping increased to 10 lesions. However, the percentage of groupings approaches 100% for RANK1 at six lesions measured bidimensionally and at nine lesions measured unidimensionally (Fig. 3). This weighting reflects the overall percentage of combinations of lesions and ranks according to their preponderance. For example, a patient may have tumor responses in two categories but with nearly all response assessments in RANK1 (Table 1).

Summary Statistics for One Patient. The calculated statistics for one patient with 10 lesions evaluated unidimensionally are shown in Table 2, demonstrating that the variance decreases with an increase in the number of lesions in a grouping. Also notice that this patient would be classified as PR if all of the 10 lesions were considered. At least three lesions would need to be measured for PR to become the dominant response category. The percentage of PR classifications increases from 52% to 78% as the number of lesions considered increases from

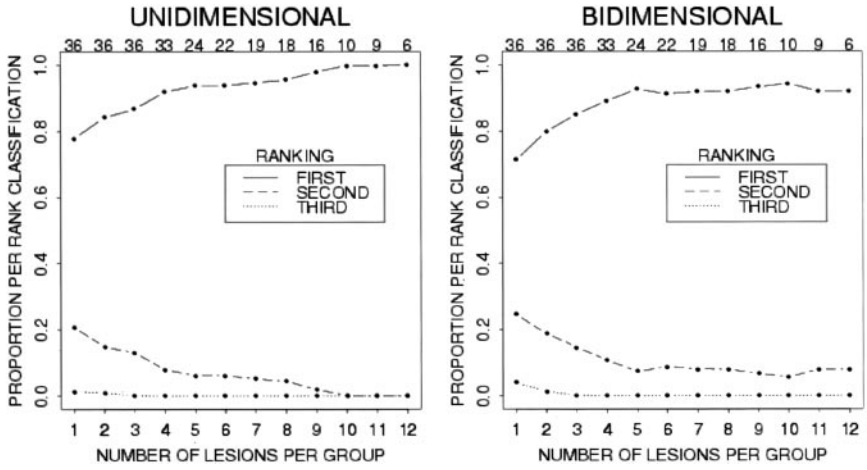

Fig. 3 Rank agreement response plot for unidimensional and bidimensional tumor measurements. The percentage of groupings for RANK1 approaches 100% when six lesions are measured bidimensionally and nine lesions are measured unidimensionally. *Numbers* along the *top* of each plot represent the number of patients contributing data to each lesion grouping size (listed on *X* axis).

Table 1 Total number of lesions, response assessments, response ranks, and response categories in six sample patients, analyzing groupings of five lesions with RECIST criteria

Patient 4 has response assessments falling into three different categories, but the proportion falling in RANK1 is very high (92%); in contrast, patient 3 has response assessments falling into fewer (two) categories, but RANK1 is only 60%. A large proportion of lesion combinations (40%) will fall into a second response category. Therefore, patient 3 is nearly evenly split between two response categories, whereas patient 4's response almost always (92%) falls into one category (but with the potential for two other response categories 8% of the time).

Patient no.	Total no. of lesions	No. of groupings	Response assessment				Response rank			No. of response categories
			CR	PR	SD	PD	1	2	3	
1	7	21	0	0	0	21	1	0	0	1
2	16	4368	0	0	3697	671	0.85	0.15	0	2
3	10	252	0	100	152	0	0.6	0.4	0	2
4	10	252	1	232	19	0	0.92	0.08	0.004	3
5	12	792	0	0	31	761	0.96	0.04	0	2
6	15	3003	0	0	3003	0	1	0	0	1

three to eight. The number of different response categories into which the patient could have been classified remained three until five lesions were considered and converged to one response category only when eight lesions were considered. In other words, the response assessment becomes more uniform as a larger number of lesions is considered.

General Trends. For bidimensional criteria, when six lesions are considered, the average variance decreased by 90%. For unidimensional criteria when more than four lesions are considered, the average variance decreased by 90% (Fig. 1). Also for both methods, beyond six lesions, lesion combinations fell into one response category with a ranked proportion of 91% (Fig. 3). However, even if six lesions were measured, 36% of patients would have a non-zero probability of falling into two response categories (Fig. 2).

Discussion

Accurate and reproducible measurements on radiological images are needed for evaluating tumor response to therapy in clinical practice. The importance of such measurements has become even more critical because the United States Food and Drug Administration recently recognized that tumor shrinkage is often an appropriate surrogate indicator for the effectiveness of a treatment and accepts such shrinkage as evidence to allow accelerated approval of cancer drugs in some situations (7). A mixed response to therapy can confound results obtained with traditional assessment criteria when cross-products of a varying mixture of enlarging and shrinking lesions are summed. Both intraobserver and interobserver variations in tumor measurements exist for a variety of reasons, such as size of initial lesion, irregular shape, and poorly defined margins of a lesion; the phase of i.v. contrast administration and the exact levels through which a lesion is scanned; and the measurement technique used [*e.g.,* hand-held calipers, electronic calipers, and automated techniques (5, 6, 8, 9)].

Our study was undertaken in an attempt to study one of the sources of variability in response assessment (namely, the number of lesions measured) and to determine the "optimal" number of lesions to measure that would minimize variance, such that if an equal number (but different set) of lesions were measured, then a similar response assessment would result. It is clear that the number of lesions selected for measurement in a given patient, as well as which particular lesions are selected, can influence the overall response assessment for that patient. We also wanted to investigate at what number of lesions the dominant response category converges to the one that would be obtained if all lesions were measured. However, selecting an "optimal" number requires making value judgments and a de-

Table 2 Calculated tumor response assessments, response ranks, and response categories for one patient, analyzing 10 lesions with RECIST criteria

No. of lesions in grouping	No. of groupings	Variance	Response assessment				Response rank			No. of response categories
			CR	PR	SD	PD	1	2	3	
1	10	0.158	5	1	4	0	0.50	0.40	0.10	3
2	45	0.067	10	17	18	0	0.40	0.38	0.22	3
3	120	0.038	10	63	47	0	0.52	0.39	0.08	3
4	210	0.025	5	133	72	0	0.63	0.34	0.02	3
5	252	0.016	1	170	81	0	0.67	0.32	0.004	3
6	210	0.011	0	148	62	0	0.70	0.30	0	2
7	120	0.007	0	90	30	0	0.75	0.25	0	2
8	45	0.004	0	35	10	0	0.78	0.22	0	2
9	10	0.002	0	10	0	0	1	0	0	1
10	1	—	0	1	0	0	1	0	0	1

cision analysis about the relative merits of the effort expended *versus* the expected benefits. For instance, in a single patient, it would require relatively little extra effort to measure all disease, with possible benefit to this individual patient. In a large Phase III trial, the effort expended is considerably greater and must be weighed against generating results with greater variance. In part, this decision depends on the design of the clinical trial and how this data will be compared with past trials.

Some patients fall into two response assessment categories even though many lesions were measured in each grouping. This likely reflects the heterogeneity and biology of tumor response, current limitations in tumor measurement techniques, and the fact that these patients may have responses or progressions close to the cutoff. In the latter scenario, lesion selection will be especially critical because it is more likely that some lesion groupings will result in a different response categorization. Whereas measuring six or more lesions resulted in no substantial decrease in variance, 9% of all lesion groupings fell into a second response category, and 38% of patients had at least one such lesion grouping that fell into a second response category.

Measurements of tumors serve as surrogate indicators of therapeutic response; other surrogate markers exist as well. Ultimately, it will be necessary to compare various surrogate markers with each other and with patient survival. The purpose of our study was to address one of the most commonly used surrogate markers, *i.e.*, tumor measurements, and to assess variability in this surrogate based on the sample size of tumor burden measured. The extent of acceptable variance will depend on the type of clinical study being performed, the use of other available surrogate markers, the relative importance of imaging as an end point, and whether comparison with other data sets will be performed.

Ideally, all tumor deposits present in a patient would be measured, but currently this is not practical for several reasons. Some lesions have imaging characteristics that make them too variable to consistently measure and would be considered "inevaluable" by current methodologies. Also, some patients' tumor burden is so great that measuring all lesions, even in one dimension, would be prohibitively time-consuming. Additionally, it may not be possible to determine whether every lesion in a patient with multiple lesions represents a cancer deposit, or whether some represent concurrent benign lesions. These problems are further compounded in the patient with more than one primary cancer.

We used standard methodology for both unidimensional and bidimensional measurements in this study. Other approaches exist, as well, such as manual boundary definition, autocontour techniques (9), and volumetric assessments (10–13); use of these approaches might further reduce variability.

Despite the endorsement of unidimensional measurements by the RECIST Group (2) and others (3, 4), their acceptance is not universal (14). Note that our study was not designed to assess the relative merits of unidimensional, bidimensional, and three-dimensional measurements or to measure intraobserver or interobserver variability.

There are several limitations in our study. Our patient population was a heterogeneous group, comprising several tumor types, target organs, and several types of systemic therapies. It would be helpful to apply our methodology to a larger patient population with more uniform tumor types and therapies and to compare those results with those obtained with other (nonimaging) surrogates for therapy response. Moreover, we did not control for differences in tumor response rate or rate of tumor progression. The magnitude of the average response variance is heavily affected by the large number of possible groupings of lesions, which is an artifact of the number of lesions present in a given patient, and needs to be carefully interpreted. Finally, radiologists' selection of lesions is frequently not random or by chance. The radiologist may choose not to include lesions that are of odd shape or that are located in a region that may be difficult to reproducibly assess at follow-up, such as a lung base or the liver dome. We did not assess the effect of nonrandom lesion selection in this study.

Our results underscore that the selection of lesions to be measured is a fundamental issue that can profoundly affect the assessment of a patient's response to therapy. Additionally, as a potential direct result of this issue, substantial societal resources can be wasted on new therapies that are inappropriately deemed to be effective, and effective therapies can be erroneously declared ineffective. Measuring as large a number of lesions as practicable will decrease the likelihood of such errors, but only up to a point; in our study, measuring more than six lesions did not yield substantial improvement. Even with measuring six lesions, the dominant response category was achieved in only

90% of lesion groupings. The degree of acceptable variance should be assessed based on the individual clinical or study requirements and the specific historical comparisons that will be made in the study.

Acknowledgments

We thank Dr. Zvi Fuks for helpful suggestions and careful review of the manuscript.

Appendices

Appendix A: Lesion Grouping. Mathematically, the number of possible lesion groupings was calculated as follows: for a patient with n lesions and lesion grouping size k (the number of lesions chosen to measure, where $k < n$), then a total of $m = {}^{n}C_{k}$ (pronounced "n choose k") possible groupings could be observed. For the sum of all possible groupings, all ms are summed.

In the above example, with $n = 4$, if we consider $k = 2$ (*i.e.*, two lesions are selected to be in the grouping), there are $m = 4!/2! \times (4 - 2)! = 4 \times 3 \times 2 \times 1/(2 \times 1)\ (2 \times 1) = 6$ possible combinations of lesions that could be selected. If all four lesions are included in the measurement (*i.e.*, $k = 4$), then only one combination is possible ($m = 1$). The total number of possible combinations of lesions would be 15 in this example.

Appendix B: Ranked Response Categories. For example, consider two sample patients with 7 and 16 lesions, respectively (rows 1 and 2 in Table 1). When five lesions are to be selected for response assessment, these patient can be assessed in 21 and 4368 possible groupings, respectively. The first patient falls in PD all 21 times, thereby making RANK1, RANK2, and RANK3 100%, 0%, and 0%, respectively. The second patient falls in SD 85% (3697 of 4368) of the time and in PD 15% (671 of 4368) of the time, thereby making RANK1, RANK2, and RANK3 85%, 15%, and 0%, respectively. Average rankings when five lesions are considered are then calculated by averaging RANK1 over 36 patients.

Appendix C: Average Variance Calculation. For example, the patient specified in Table 2 contributed a total of 10 lesions for evaluation. When considered one lesion at a time ($k = 1$), the percentage changes in lesion diameter were -0.50, -1.00, -1.00, -1.00, -1.00, -0.21, -0.17, -0.29, -0.16, and -1.00. The simple variance of these values is 0.158 (*i.e.*, the first number under variance in Table 2). Similar variance calculations were done for all 36 patients for a grouping size of one lesion ($k = 1$), and the average of these variances was estimated (0.077; Fig. 1).

References

1. Miller, A. B., Hoogstraten, B., Staquet, M., and Winkler, A. Reporting results of cancer treatment. Cancer (Phila.), *47:* 207–214, 1981.

2. Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A. T., Christian, M. C., and Gwyther, S. G. New guidelines to evaluate the response to treatment in solid tumors. J. Natl. Cancer Inst. (Bethesda), *92:* 205–216, 2000.

3. James, K., Eisenhauer, E., Christian, M., Terenziani, M., Vena, D., Muldal, A., and Therasse, P. Measuring response in solid tumors: unidimensional *versus* bidimensional measurement. J. Natl. Cancer Inst. (Bethesda), *91:* 523–528, 1999.

4. Dachman, A. H., MacEneaney, P. M., Adedipe, A., Carlin, M., and Schumm, L. P. Tumor size on computed tomography scans: is one measurement enough? Cancer (Phila.), *91:* 555–560, 2001.

5. Thiesse, P., Ollivier, L., Di Stefano-Louineau, D., Nãgrier, S., Savary, J., Pignard, K., Lasset, C., and Escudier, B. Response rate accuracy in oncology trials: reasons for interobserver variability. J. Clin. Oncol., *15:* 3507–3514, 1997.

6. Hopper, K. D., Kasales, C. J., Van Slyke, M. A., Schwartz, T. A., TenHave, T. R., and Jozefiak, J. A. Analysis of interobserver and intraobserver variability in CT tumor measurements. Am. J. Roentgenol., *167:* 851–854, 1996.

7. United States Food and Drug Administration Backgrounder. Cancer Therapies: Accelerating Approval and Expanding Access (March 29, 1996). Available at http://www.fda.gov/opacom/backgrounders/cancerbg.html.

8. Van Hoe, L., Van Cutsem, E., Vergote, I., Baert, A. L., Bellon, E., Dupont, P., and Marchal G. Size quantification of liver metastases in patients undergoing cancer treatment: reproducibility of one-, two-, and three-dimensional measurements determined with spiral CT. Radiology, *202:* 671–675, 1997.

9. Schwartz, L. H., Ginsberg, M. S., DeCorato, D., Rothenberg, L. N., Einstein, S., Kijewski, P., and Panicek, D. M. Evaluation of tumor measurements in oncology: use of film-based and electronic techniques. J. Clin. Oncol., *18:* 2179–2184, 2000.

10. Breiman, R. S., Beck, J. W., Korobkin, M., Glenny, R., Akwari, O. E., Heaston, D. K., Moore, A. V., and Ram, P. C. Volume determinations using computed tomography. Am. J. Roentgenol., *138:* 329–333, 1982.

11. Eggli, K. D., Close, P., Dillon, P. W., Umlauf, M., and Hopper, K. D. Three-dimensional quantitation of pediatric tumor bulk. Pediatr. Radiol., *25:* 1–6, 1995.

12. Hopper, K. D., Kasales, C. J., Eggli, K. D., TenHave, T. R., Belman, N. M., Potok, P. S., Van Slyke, M. A., Olt, G. J., Close, P., Lipton, A., Harvey, H. A., and Hartzel, J. S. The impact of 2D *versus* 3D quantitation of tumor bulk determination on current methods of assessing response to treatment. J. Comput. Assist. Tomogr., *20:* 930–937, 1996.

13. Prasad, S. R., Jhaveri, K. S., Saini, S., Hahn, P. F., Halpern, E. F., and Sumner, J. E. CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques–initial observations. Radiology, *225:* 416–419, 2002.

14. Saini, S. Radiologic measurement of tumor size in clinical trials: past, present, and future. Am. J. Roentgenol., *176:* 333–334, 2001.