Variability in Mesothelioma Tumor Response Classification

Samuel G. Armato, III[1]
Joseph L. Ogarek[1]
Adam Starkey[1]
Nicholas J. Vogelzang[2]
Hedy L. Kindler[3]
Masha Kocherginsky[4]
Heber MacMahon[1]

OBJECTIVE. The objective of our study was to evaluate observer variability in the measurement of temporal change in mesothelioma tumor thickness and in the resulting tumor response classification from CT scans. In addition, the performance of a semiautomated measurement method was evaluated.

MATERIALS AND METHODS. Four observers individually used an interface that displayed two serial CT scans from the same patient to measure mesothelioma tumor thickness on the follow-up CT scans of 22 patients based on baseline scan measurements. During one session, observers acquired measurements on the follow-up scans based on written reports of baseline scan measurements; in another session, baseline scan measurements were superimposed on the baseline scan for direct visual comparison. Follow-up scan measurements also were obtained from a semiautomated method. Measurement variability and tumor response classification concordance were evaluated for manual measurements acquired in both modes and for semiautomated measurements.

RESULTS. Although only a small increase in tumor response classification concordance rate was obtained with the visual approach (84.8%) relative to the more standard written-report approach (82.6%), the actual measurements acquired by observers were statistically significantly different between the two approaches ($p = 0.03$). Both the semiautomated measurements and the resulting tumor response classifications were consistent with manual measurements.

CONCLUSION. The presentation of baseline scan tumor measurements affects measurements acquired on follow-up scans and could influence tumor response classification. The potential utility of semiautomated tumor thickness measurements was shown in the context of measuring tumor response.

Keywords: computer-aided diagnosis, CT, lung cancer, mesothelioma, oncologic imaging, tumor response

DOI:10.2214/AJR.05.0078

Received January 14, 2005; accepted after revision February 22, 2005.

S. G. Armato, III, and H. MacMahon are minor shareholders in R2 Technology, Inc. (Sunnyvale, CA).

Presented in part at the 2004 annual meeting of the American Society of Clinical Oncology, New Orleans, LA.

Supported in part by the Mesothelioma Applied Research Foundation and funding from The University of Chicago Cancer Research Center through a gift from Fay and Cal Sawyier.

[1]Department of Radiology, The University of Chicago, 5841 S Maryland Ave., Chicago, IL 60637. Address correspondence to S. G. Armato, III (s-armato@uchicago.edu).

[2]Nevada Cancer Institute, Las Vegas, NV 89135.

[3]Department of Medicine, The University of Chicago, Chicago, IL 60637.

[4]Department of Health Studies, The University of Chicago, Chicago, IL 60637.

AJR 2006; 186:1000–1006

0361–803X/06/1864–1000

Malignant mesothelioma is a devastating asbestos-induced malignancy affecting an estimated 2,500 Americans annually. Approximately 80% of mesotheliomas are cancers of the pleura. Systemic chemotherapy is the only potential treatment option for most of these patients due to advanced disease at presentation [1]. The long-standing therapeutic nihilism regarding chemotherapy for this disease is no longer warranted because of the recent introduction of several active new agents such as pemetrexed, gemcitabine, and vinorelbine [2–4]. Chemotherapy has been shown to improve symptoms and prolong survival in patients with mesothelioma, and many promising new agents for the treatment of this disease are currently under investigation [5]. The development of these new active treatments emphasizes the need for accurate tumor measurement techniques with which to assess tumor response to therapy.

CT has been established as an effective imaging technique for the evaluation of mesothelioma [6–10]. Nevertheless, the progression of mesothelioma and the response of this disease to therapy have been difficult for clinicians to evaluate consistently and reproducibly due, in part, to the circumferential morphology and axial extent of this tumor. Such difficulties impair the evaluation of patient prognosis and hinder an accurate assessment of clinical trials.

The notion of tumor response is fundamental in oncology. Assessment of disease progression or response to therapy is necessary for the clinical management of patients and is critical for the evaluation of clinical trials. The radiologic assessment of response of patients enrolled in clinical trials has gained acceptance as a surrogate for patient survival outcomes during the regulatory approval process [11]. This

radiologic assessment, however, necessitates quantitative tumor measurements and the standardization of tumor response criteria based on such measurements.

In 1981, the World Health Organization (WHO) [12] recommended the radiologic quantification of solid tumors through bidimensional measurements on imaging studies. Tumor response then was determined from a comparison of a lesion's bidimensional measurements across temporally sequential imaging studies [12]. Nearly two decades later, the Response Evaluation Criteria in Solid Tumors (RECIST) guidelines advocated the replacement of bidimensional tumor measurements with one-dimensional measurements of the longest diameter of the lesion [13, 14].

The measurement guidelines offered by WHO and RECIST were designed for compact tumors and thus are generally not as appropriate for the circumferential growth pattern and often scalloped morphology of mesothelioma. Accordingly, alternative CT measurement protocols have been proposed specifically for mesothelioma. For one such protocol, between one and three one-dimensional measurements of pleural thickness are obtained on each of three CT sections [2, 3] according to a modified RECIST approach [15]. The sum of these one-dimensional measurements is used to represent tumor burden. The RECIST guidelines for tumor response classification then are applied to the summed measurements obtained from temporally sequential CT scans.

The actual manner in which tumor measurement protocols are implemented raises issues of consistency and reproducibility. In studies unrelated to mesothelioma, inter- and intraobserver variability in the selection and measurement of lesions on CT scans have been reported [16–18]; however, the circumferential morphology and axial extent of mesothelioma further complicate the measurement of this specific type of tumor. Such difficulties may impair the accurate evaluation of patient prognosis and may hinder accurate evaluation of clinical trials.

We previously articulated a three-step process for the manual measurement of mesothelioma on CT scans that involves, first, selection of a limited number of CT sections in which the disease is most prominent; second, identification of specific sites within these sections; and, third, the actual measurement of tumor thickness at these sites [19]. In a previous study [19], we assessed observer measurement variability in the third step on the basis of a single CT scan from each of 22 patients with mesothelioma; the 95% limits of agreement for relative interobserver difference of tumor thickness measurements spanned a range of 30%. We noted the expectation of increased variability if observers had been allowed to implement all three steps of the measurement process and if temporally sequential scans of the same patient had been evaluated as they are in clinical practice.

The purpose of the present study, therefore, was to evaluate the variability of mesothelioma tumor measurements and the resulting tumor response classifications among observers based on temporally sequential CT scans. Variability in this measurement task will directly impact the treatment of patients with mesothelioma and the outcomes of clinical trials that seek to validate the efficacy of novel therapeutic agents. Specifically, this study compares the variability of manual measurements obtained through a computer interface that stores the precise locations of measurements acquired from the baseline CT scan as a visual assist to the acquisition of measurements from the follow-up scan with the variability of manual measurements acquired from the follow-up scan based on standard written radiology reports of baseline scan measurements.

Materials and Methods

The database consisted of 44 diagnostic thoracic helical CT scans: two scans acquired from each of 22 patients with biopsy-proven malignant pleural mesothelioma. These patients (six women, 16 men; age range at baseline, 38–80 years; mean age, 68 years) had been enrolled in chemotherapy clinical trials at our institution, and no scan was acquired specifically for this study. Informed consent was obtained for the research use of each patient's CT scans.

The CT examinations were performed on a HiSpeed Advantage or LightSpeed CT scanner (GE Healthcare). Each CT section was reconstructed as a 512 × 512 pixel image matrix. The scans were acquired with a collimation of 7 (n = 37), 7.5 (n = 1), 5 (n = 3), or 10 (n = 3) mm. The reconstruction interval was 7 or 10 mm for scans with 10-mm collimation and 5 mm for most sections in all other scans. The time interval between the sequential scans for the 22 patients ranged from 7 to 30 weeks (mean, 15 weeks). The baseline scans, along with the corresponding tumor thickness measurements, were obtained from a previous study [19]. Variability of change in tumor thickness measurements was evaluated without regard to the eventual clinically determined tumor response classification.

Visualization Interface

A computer interface was developed to simultaneously display two CT scans from the same patient and to allow the manual measurement of structures within a scan. The measurement of a structure (e.g., mesothelioma tumor thickness) is obtained from the length of a line segment the user constructs within a specific CT section. To construct a line segment, an initial end point is placed when the user clicks the mouse with the cursor at the desired image location. The user drags the mouse to create a visible line segment extending from the initial end point. The terminal end point is established when the user releases the mouse button to fix the line segment within the image. The corresponding lengths of the line segments are displayed, and attributes of each line segment (length, section, and spatial coordinates of the end points) are stored to allow comparisons of measurements among observers and across serial scans of the same patient. The storage of image annotation and measurement data through an integrated database and user interface has been shown as a viable model for radiologic image interpretation [20].

The interface was used by observers to measure mesothelioma tumor thickness on follow-up scans in two modes: one in which the observer acquired measurements in the follow-up scan based on written radiology reports of baseline scan measurements and another in which the observer acquired measurements in the follow-up scan with the baseline scan measurements superimposed on the baseline scan for direct visual comparison. Measurement variability and tumor response classification concordance were evaluated for measurements acquired in both modes.

Baseline Tumor Thickness Measurements

In conjunction with an earlier study [19], an experienced thoracic radiologist used the interface to review the 22 baseline CT scans. During this initial review, the radiologist implemented our clinical protocol for the measurement of mesothelioma on CT scans [2] by selecting for each scan three sections on which measurements would be acquired and then identifying up to three measurement sites on each chosen section. The radiologist acquired measurements through a single-scan version of the interface. In total, 134 measurements were made on 66 sections of the 22 baseline scans. The selected sections and the line segments constructed to capture tumor thickness at each measurement site were stored in the computer.

The same radiologist reviewed the tumor thickness measurements and the superimposed line segments and translated these measurements and their spatial locations into a dictated report. Measurements were verbally expressed in a manner consistent with clinical practice at our institution at that time—for ex-

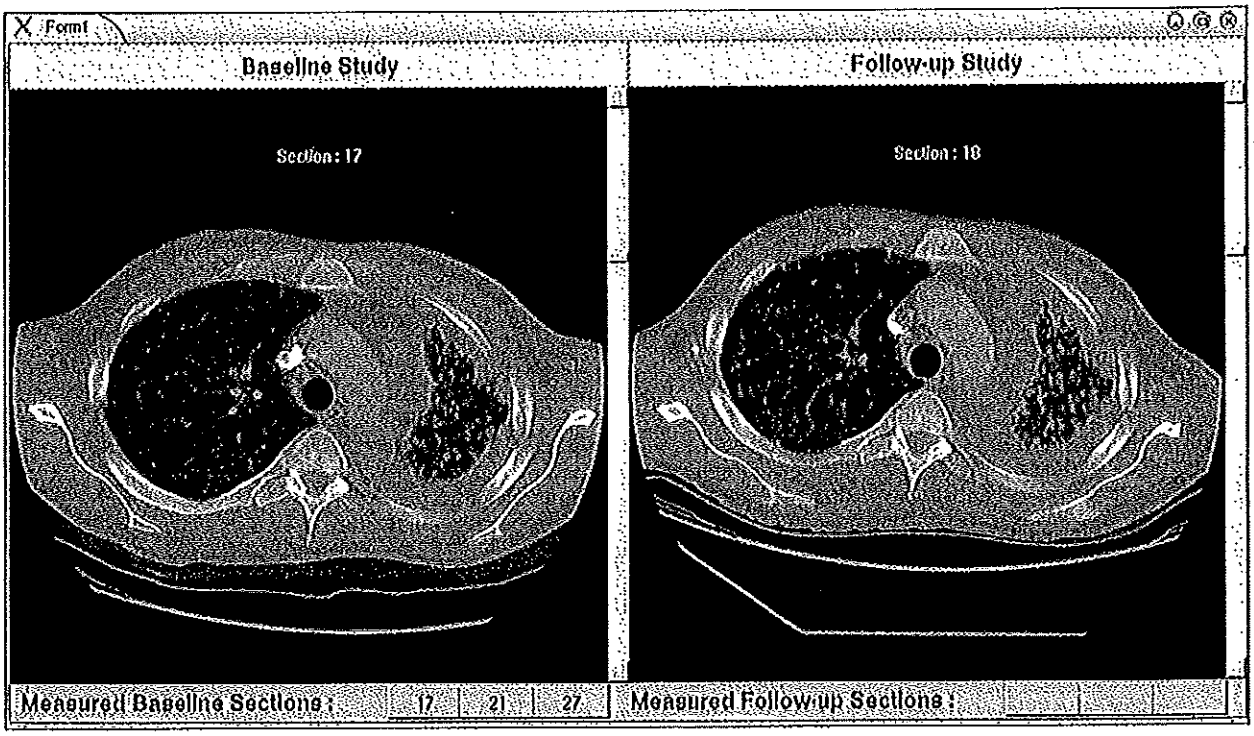


Fig. 1—Screen capture shows interface used for side-by-side comparison of temporally sequential CT scans from 38-year-old man. Baseline scan is viewed on left side, and follow-up scan (for which measurements were acquired for this study) is viewed on right side. For visual approach of measurement, line segments representing baseline scan measurements were superimposed on baseline scan images for direct visual comparison.

ample, "section 9, at approximately the 9-o'clock position, in the left paratracheal area: 35 mm." The transcribed report simulated a written radiology report for the 22 baseline scans. To avoid bias, this radiologist provided only baseline scan measurements and did not participate in the remainder of this study.

Follow-Up Tumor Thickness Measurements

Four observers (two attending thoracic radiologists and two attending oncologists) independently reviewed the 22 CT scan pairs in random order through the interface. The scan pairs were displayed through the interface with the baseline scan always displayed on the left side of the interface and the corresponding follow-up scan displayed on the right (Fig. 1). Each scan was displayed one section at a time, and the interface provided slider bars for the independent control of the displayed section for both scans. The task of the observers was to measure tumor thickness on the follow-up scans by constructing line segments that spanned the mesothelioma tumor at sites that corresponded to the sites at which measurements had been made in the baseline scans. These follow-up measurements were captured and stored through the interface.

Baseline scan measurements acquired by the initial radiologist were presented to the observers through two approaches: as a written report and visually as tumor-spanning line segments superimposed on the three sections of the baseline scan (Fig. 2). The observers independently measured tumor thickness on the follow-up scan of each of the 22 patients during two sessions separated by at least 3 weeks. During each session, baseline scan measurements were presented through one of the two approaches. For the visual approach, the initial radiologist's tumor-spanning line segments and corresponding tumor thickness measurements were superimposed on these baseline scan sections; for the written-report approach, these visual cues were not provided, and the baseline scan sections were displayed without annotation. Two observers performed the visual approach during the first of their sessions, while the other two observers performed the written-report approach first. For the baseline scan measurements shown in Figure 3A, Figure 3B depicts the tumor thickness measurements acquired by one observer in the corresponding follow-up scan section (selected by that observer) based on both approaches; the potential for substantially different tumor thickness measurements with and without the visual aid is evident.

During an observer's first session, the observer manually selected the three follow-up scan sections that anatomically matched the three baseline scan sections on which measurements had been made by the initial radiologist. Interobserver variability in the manual selection of anatomically matched sections between temporally sequential CT scans was evaluated and has been reported previously [21]. The observer then acquired an equal number of measurements in the selected follow-up scan sections based on either the visual information or the written report regarding the baseline scan measurements. During an observer's second session, the observer was constrained to acquire measurements from the same sections in the follow-up scans that were selected by that observer during the first session to eliminate intraobserver section selection differences as a possible source of measurement variability. At the completion of the study, a one-to-one correspondence existed between the initial radiologist's 134 baseline scan measurements and each of the two sets of follow-up scan measurements made by each observer.

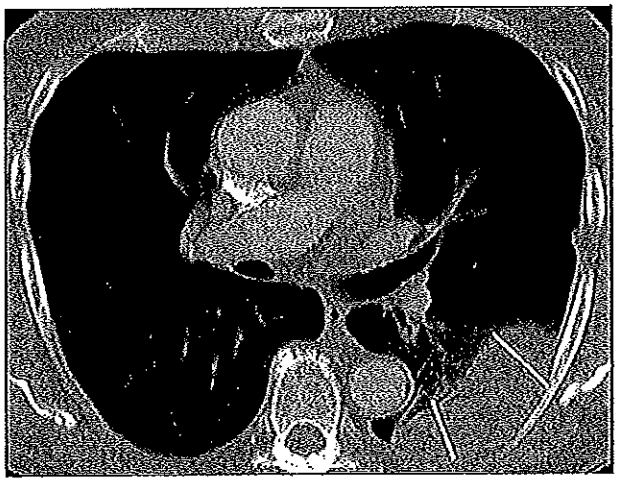
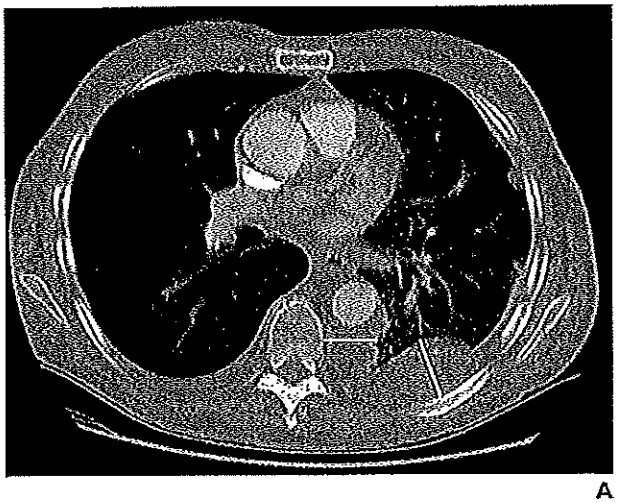
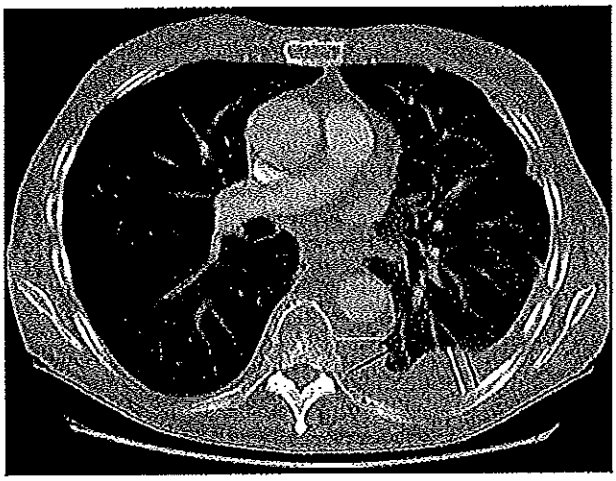


Fig. 2—Representative baseline CT scan section of 77-year-old man with tumor thickness measurements acquired by initial radiologist shown as tumor-spanning line segments (*white lines*) superimposed for direct visual comparison.



A



B

Fig. 3—66-year-old man.
A, Baseline scan measurements are shown superimposed on one section of baseline CT scans.
B, Tumor thickness measurements acquired by observer on corresponding follow-up scan section (as selected by that observer) based on written report of baseline scan measurements (black line segments) and with benefit of visual aid (white line segments). Baseline measurements were not superimposed during implementation of written-report approach.

Tumor Response Classification

In accordance with the RECIST guidelines, each case was classified as showing partial response if the sum of the one-dimensional measurements in the follow-up CT scan was less than 30% lower than the sum from the baseline scan, progressive disease if the sum of the one-dimensional measurements in the follow-up CT scan was more than 20% greater than the sum from the baseline scan, or stable disease if the extent of summed measurement reduction was not great enough to qualify as partial response or the extent of measurement increase was not great enough to qualify as progressive disease [14]. No case in the database exhibited the fourth RECIST classification, complete response. For a given case, concordance was achieved if two different methods (e.g., two different observers or two different measurement approaches) yielded the same tumor response classification.

Semiautomated Tumor Thickness Measurements

This study presented an opportunity to further evaluate the efficacy of the semiautomated method we developed to quantify mesothelioma tumor thickness [19, 22]. The semiautomated method requires as input a user-specified point along the outer tumor margin. Based on this point, the computer automatically identifies a corresponding point along the inner tumor margin, and the distance between these two points is returned as the semiautomated tumor thickness measurement. In the present study, the outer tumor margin points identified by each observer during the visual-based measurement session were input to the semiautomated method at the completion of the session. Tumor response classifications were derived from the baseline scan measurements of the initial radiologist and the semiautomated measurements of the corresponding follow-up scan based on input from each observer.

Statistical Analysis

Interobserver agreement between the RECIST classifications of all pairs of observers was assessed by calculating the concordance rates for the written-report and visual approaches. Intraobserver concordance rates between the two measurement approaches were computed to assess the agreement of the two methods. Concordance rates between the visual approach and the corresponding classifications derived from the semiautomated method also were computed. Kappa coefficients [23] were used to assess the overall interobserver agreement for each of the two measurement methods and for the semiautomated method. The kappa coefficient assumes that the two observers agree when they obtain the same RECIST tumor response classification and that they disagree otherwise.

Agreement between and within observers also was assessed using the actual continuous summed measurements that the observers obtained with the visual and written-report measurement approaches. A mixed-effects analysis of variance model was used to determine whether the summed measurements acquired from each follow-up scan differed between the visual and written-report approaches:

$$y_{ijk} = \mu + \alpha_k + s_i + r_j + e_{ijk},$$

where α_k is the fixed effect representing the measurement approach ($k = 1$ for the visual approach, $k = 2$ for the written-report approach); s_i and r_j are the random effects of the scan and the observer, respectively; and e_{ijk} is the error term. The model was fitted using the PROC MIXED function of SAS software, version 8 (SAS Institute). Differences in percent measurement change between corresponding baseline and follow-up scans were assessed using a similar mixed-effects model. In all of these models, the responses were log-transformed when appropriate to satisfy the normality assumption.

Similar models were used for the comparison of the visual-based measurements and the corresponding semiautomated measurements and percent changes relative to baseline. Due to the natural pairing of the measurements between the two methods (the semiautomated measurements used the initial end point selected by the observer during acquisition of the visual-based measurements), the response variable used was $\log(y_{ij1} / y_{ij2})$, the log-transformed ratio of the measurements, for the summed measurements and $\text{sgn}(y_{ij1} / y_{ij2}) \times \log(|y_{ij1} / y_{ij2}|)$ for percent change rather than the actual measurements as in the models described. This transformation eliminates the fixed effect in the previous model; therefore, only the intercept term is relevant in this random effects model. Accordingly, the null hypothesis corresponds to a test of whether the ratio of the measurements based on the two methods is equal to 1.

Results

The first task performed by the observers before measurements on a follow-up CT scan could be acquired was selection of the sections in the follow-up scan that corresponded to the sections of the baseline scan on which measurements had been made. Accordingly, each observer was presented with 66 section-matching tasks (three sections in each of the 22 follow-up scans), which introduced interobserver variability as previously reported [21]. All observers selected for measurement the exact same section of the follow-up scan for 14 (21.2%) of the 66 baseline scan sections. For 31 (47.0%) of the 66 baseline scan sections, the sections of the follow-up scan selected by observers spanned two contiguous sections; 12.1% of the matching tasks yielded selected follow-up scan sections that spanned four or more contiguous sections.

The tumor response classifications derived from the fixed set of baseline scan measurements and the corresponding follow-up scan measurements acquired by four observers for both measurement approaches are presented in Table 1. Tumor response classifications are based on RECIST criteria: partial response (PR), stable disease (SD), and progressive disease (PD). Table 2 presents the intraobserver tumor response classification concordance rates based on the two measurement approaches, which ranged from 77.3% to 90.9%, with a mean of 83.0% across observers.

Table 2 also presents the concordance rates between the tumor response classifications obtained by each observer from the visual-based measurement approach and the corresponding tumor response classifications derived from the semiautomated method. Except observer C, the concordance rates between each observer's tumor response classification based on the visual approach and that based on the semiautomated measurement method was greater than the concordance rates between each observer's classifications based on the written-report and visual approaches. The concordance rates between each observer's visual-approach classifications and their corresponding semiautomated tumor response classifications ranged from 86.4% to 90.9%, with a mean of 88.6% across observers.

Table 3 presents the tumor response classification concordance rates for each pair of observers. The classification concordance rates based on the written-report measurements and the visual measurements are given above and below the diagonal, respectively. The concordance rates with the visual approach are higher than the corresponding concordance rates with the written-report approach for four of the six pairwise comparisons. Overall, of the 132 pairwise comparisons of tumor response classification for each of the measurement methods (six combinations of observers for each of 22

TABLE 1: Tumor Response Classifications for the Eight Cases That Showed Discordant Tumor Response Classifications by the Four Observers with the Two Measurement Approaches

Case No.	Written-Report Approach				Visual Approach			
	Observer A	Observer B	Observer C	Observer D	Observer A	Observer B	Observer C	Observer D
4	SD	SD	SD	PR	SD	SD	SD	SD
5	SD	PR	SD	PR	PR	SD	SD	PR
6	SD	PR	SD	SD	SD	SD	SD	SD
7	PR	SD	PR	PR	SD	PR	PR	SD
9	SD	PD	SD	SD	PD	PD	SD	PD
14	PD	PD	SD	SD	PD	SD	PD	PD
19	SD	SD	SD	SD	SD	SD	SD	PR
22	SD	PD	SD	SD	PD	PD	PD	SD

Note—Tumor response classifications are in accordance with Response Evaluation Criteria in Solid Tumors (RECIST) guidelines: PR = partial response, SD = stable disease, and PD = progressive disease.

TABLE 2: Tumor Response Classification Concordance Rates Between the Written-Report Approach and the Visual Approach Measurements of Each Observer and Between the Visual Approach Measurements of Each Observer and the Measurements Derived from the Semiautomated Method

Observer	Concordance Rate (%) and No. of Cases of Agreement / Total No. of Cases	
	Written-Report Approach / Visual Approach	Visual Approach / Semiautomated Approach
A	81.8 (18/22)	90.9 (20/22)
B	81.8 (18/22)	90.9 (20/22)
C	90.9 (20/22)	86.4 (19/22)
D	77.3 (17/22)	86.4 (19/22)

TABLE 3: Tumor Response Classification Concordance Rates Between the Written-Report Approach Follow-Up Scan Measurements of Different Observers (Cells Above the Diagonal) and Tumor Response Classification Concordance Rates Between the Visual Approach Follow-Up Scan Measurements of Different Observers (Cells Below the Diagonal)

Observer	Observer			
	A	B	C	D
A	—	77.3	95.5	86.4
B	86.4	—	72.7	72.7
C	86.4	90.9	—	90.9
D	90.9	77.3	77.3	—

Note—Data are percentages; dash (—) indicates not applicable.

cases), observers achieved an 84.8% concordance rate ($n = 112$) with the visual approach and an 82.6% concordance rate ($n = 109$) for the written-report approach. Of the 132 pairwise comparisons of tumor response classification for the semiautomated method applied to the input of each observer, the semiautomated method achieved an 84.1% concordance rate ($n = 111$) across observers.

The kappa statistic for interobserver agreement, when computed for the tumor response classifications based on the written-report measurements, was 0.66 with a 95% confidence interval (CI) of 0.53–0.79, and a kappa statistic value of 0.72 with a 95% CI of 0.60–0.85 was obtained for the tumor response classifications from the visual-based measurements across observers. Although agreement of tumor response classification across observers was greater for the visual approach, the two kappa statistic values are close. A similar level of agreement was obtained among the tumor response classifications from the semiautomated measurements across observers, with a kappa statistic value of 0.70 (95% CI, 0.58–0.83).

To analyze the measurement data directly (instead of the measurements as filtered through the RECIST classifications), we used the mixed-effects model to assess differences in the summed lengths (i.e., numeric values) for the follow-up scans of all 22 cases across all four observers and in the percent change values between the follow-up scans and baseline scans of all 22 cases across all four observers. Statistically significant differences were observed between the two methods for both the summed lengths and percent change analyses ($p = 0.03$ and $p < 0.01$, respectively). The estimated mean of the summed lengths of the follow-up scans was statistically significantly lower for the written-report approach: The average decrease in tumor thickness between the baseline scans and the corresponding follow-up scans was estimated to be 7.3% using the written-report approach and 1.9% for the visual approach.

The semiautomated measurements and the resulting percent change values also were analyzed using a mixed-effects model, which was applied to the difference in the log-transformed, summed visual-approach-based measurements and the corresponding semiautomated measurements of each observer. The model was applied similarly to the percent change in tumor size (relative to the baseline scan measurements) that resulted from the two methods. The semiautomated measurements were not statistically significantly different from the visual approach measurements for both the summed lengths and percent change analyses ($p = 0.35$ and $p = 0.56$, respectively).

Discussion

This study assessed interobserver variability in the evaluation of tumor response in patients with mesothelioma. In a context that involved solid tumors other than mesothelioma, other researchers have recommended the reporting of tumor response from measurements acquired by a single observer who measures both baseline and follow-up scans simultaneously [17]. Because measurements of the baseline scan would have been acquired at the time of that scan, the simultaneous measurement of baseline and follow-up scans after each follow-up scan introduces a redundancy that would increase the workload of the radiologist or clinician. Depending on the number of scans acquired from a particular patient, this workload increase could become substantial. The visual approach, in which follow-up scans are measured in conjunction with the display of measurements on the baseline CT scan images, was intended to reduce the effect of a major component of interobserver variability in the evaluation of tumor response: differences in the interpretation of baseline scan measurements as recorded in a written report. Through the visual approach, observers were able to see exactly where measurements had been made on the baseline scan; the observers were aware that in this study, as in clinical practice, consistency was important. This approach should, in principle, obviate simultaneous measurements by the same individual to achieve consistent measurements across serial CT scans.

Despite the expected impact of the visual approach, the resulting concordance rate among observers who all viewed the same measurements on the baseline CT scans was only slightly higher than the concordance rate derived from measurements acquired with the written-report approach. One reason for only a small increase in concordance rate may have been the variability introduced by the range of follow-up scan sections selected by observers to match each baseline scan section. Although presenting observers with the same fixed set of 66 follow-up scan sections (three sections for each of the 22 follow-up scans) may have controlled for this effect, such a design would have strayed from the clinical implementation we intended to simulate. The small difference in concordance rates also may indicate that the sample size was too small.

Perhaps a more important reason for the small difference between concordance rates may lie in the very nature of response classifications. Grouping quantitatively continuous measurements (i.e., the measurement values that were actually acquired by the observers) into three distinct categories to obtain the tumor response classification of RECIST inherently smoothes variations that may be present in the underlying, higher resolution measurement data. This phenomenon clearly is present in this study: When the summed measurements of each observer for each follow-up scan were evaluated, statistically significant differences were obtained between the summed measurements acquired with the written-report approach and the visual approach. Significant differences also were observed between the percent change values obtained by observers with the two approaches. Observers clearly acquired measurements in a different manner between the two measurement approaches, even though this study was unable to show a significant difference in the translation of these measurements into the

more relevant clinical concept of tumor response classification.

No truth exists with regard to the measurement of mesothelioma, so this study makes no attempt to determine which of the two measurement techniques is better than the other in an absolute sense; however, relative differences clearly exist. The study was performed with actual CT scans by observers familiar with the clinical protocol for the measurement of mesothelioma. The acquisition of a limited number of linear measurements from a limited number of CT sections is consistent with current clinical practice. Volumetric tumor analysis was not performed in this study because it is not the current standard for evaluation of this disease.

The CT scans used in this study had been acquired with a collimation and reconstruction interval that were thicker (5–10 mm) than those currently available with state-of-the-art scanners. In general, curved tumor boundaries are less clearly demarcated on thick-section CT scans relative to thin-section scans; however, the growth pattern of mesothelioma and the fact that measurements were obtained from selected CT sections in which the chest wall, and hence the tumor, was more axially oriented likely minimized this effect. Nevertheless, additional studies are planned with higher resolution CT scans.

This study validated the semiautomated methods of measuring mesothelioma tumor thickness that we have developed. On a set of follow-up CT scans that had not been used in the development of the computerized techniques, the concordance rate between the response classifications generated by the semiautomated method and by the visual approach of the corresponding observer was comparable to the concordance rate of observers among themselves with the visual approach. Furthermore, the summed measurements and the percent change values generated by the semiautomated method were not statistically significantly different from the corresponding manual measurements obtained with the visual approach. These encouraging results show the potential of semiautomated tumor thickness measurement methods in a clinical setting.

In conclusion, the manner in which baseline scan tumor thickness measurements are presented to observers will affect the measurements acquired by the observers on the follow-up scans and could influence tumor response classification. Because consistency of tumor measurements across serial CT scans is important for the proper assessment of disease progression or response to therapy, clinicians must give careful consideration to the measurement acquisition process. In addition, this study validated our semiautomated methods of tumor thickness measurement and showed the potential of these methods in the context of measuring tumor response.

Acknowledgments

We thank John Pennessy and Alexandra Funaki for their participation in the tumor measurement sessions and Geoffrey Oxnard for enlightening discussions.

References

1. Kindler HL, Vogelzang NJ. Mesothelioma. In: Vokes EE, Golomb HM, eds. *Oncologic therapies*. Berlin, Germany: Springer-Verlag, 2003:415–423
2. Vogelzang NJ, Rusthoven JJ, Symanowski J, et al. Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma. *J Clin Oncol* 2003; 21:2636–2644
3. Byrne MJ, Davidson JA, Musk AW, et al. Cisplatin and gemcitabine treatment for malignant mesothelioma: a phase II study. *J Clin Oncol* 1999; 17:25–30
4. Steele JP, Shamash J, Evans MT, Gower NH, Tischkowitz MD, Rudd RM. Phase II study of vinorelbine in patients with malignant pleural mesothelioma. *J Clin Oncol* 2000; 18:3912–3917
5. Kindler HL. Moving beyond chemotherapy: novel cytostatic agents for malignant mesothelioma. *Lung Cancer* 2004; 45[suppl]:S125–S127
6. Kawashima A, Libshitz HI. Malignant pleural mesothelioma: CT manifestations in 50 cases. *AJR* 1990; 155:965–969
7. Heelan RT, Rusch VW, Begg CB, Panicek DM, Caravelli JF, Eisen C. Staging of malignant pleural mesothelioma: comparison of CT and MR imaging. *AJR* 1999; 172:1039–1047
8. Marom EM, Erasmus JJ, Pass HI, Patz EF Jr. The role of imaging in malignant pleural mesothelioma. *Semin Oncol* 2002; 29:26–35
9. Ng CS, Munden RF, Libshitz HI. Malignant pleural mesothelioma: the spectrum of manifestations on CT in 70 cases. *Clin Radiol* 1999; 54:415–421
10. Yilmaz UM, Utkaner G, Yalniz E, Kumcuoglu Z. Computed tomographic findings of environmental asbestos-related malignant pleural mesothelioma. *Respirology* 1998; 3:33–38
11. Saini S. Radiologic measurement of tumor size in clinical trials: past, present, and future. *AJR* 2001; 176:333–334
12. Miller AB, Hogestraeten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981; 47:207–214
13. James K, Eisenhauer E, Christian M, et al. Measuring response in solid tumors: unidimensional versus bidimensional measurement. *J Natl Cancer Inst* 1999; 91:523–528
14. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 2000; 92:205–216
15. Byrne MJ, Nowak AK. Modified RECIST criteria for assessment of response in malignant pleural mesothelioma. *Ann Oncol* 2004; 15:257–260
16. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR* 1996; 167:851–854
17. Erasmus JJ, Gladish GW, Broemeling L, 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; 21:2574–2582
18. Thiesse P, Ollivier L, Di Stefano-Louineau D, et al. Response rate accuracy in oncology trials: reasons for interobserver variability. *J Clin Oncol* 1997; 15:3507–3514
19. Armato SG III, Oxnard GR, MacMahon H, et al. Measurement of mesothelioma on thoracic CT scans: a comparison of manual and computer-assisted techniques. *Med Phys* 2004; 31:1105–1115
20. Aberle DR, Dionisio JDN, McNitt-Gray MF, et al. Integrated multimedia timeline of medical images and data for thoracic oncology patients. *RadioGraphics* 1996; 16:669–681
21. Sensakovic WF, Armato SG III, Starkey A, Ogarek JL. Automated matching of temporally sequential CT sections. *Med Phys* 2004; 31:3417–3424
22. Armato SG III, Oxnard GR, Kocherginsky M, Vogelzang NJ, Kindler HL, MacMahon H. Evaluation of semi-automated measurements of mesothelioma tumor thickness on CT scans. 2005; *Acad Radiol* 12:1301–1309
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33:159–174