Translational Oncology
www.transonc.com

Volume 5 Number 1 February 2012 pp. 19–25 19

Measurement of Tumor Volumes Improves RECIST-Based Response Assessments in Advanced Lung Cancer[1]

P. David Mozley*, Claus Bendtsen†, Binsheng Zhao‡, Lawrence H. Schwartz‡, Matthias Thorn§, Yuanxin Rong¶, Luduan Zhang#, Andrea Perrone, René Korn†† and Andrew J. Buckler‡‡**

*Merck Research Laboratories, West Point, PA, USA; †AstraZeneca, Nether Alderley, Cheshire, UK; ‡Columbia University, New York, NY, USA; §Siemens AG, Forchheim, Germany; ¶Perceptive Informatics, Inc, Billerica, MA, USA; #Intio, Inc, Broomfield, CO, USA; **BioClinica, Inc, Newtown, PA, USA; ††Definiens, AG, Munich, Germany; ‡‡BBMSC, LLC, Weham, MA, USA

Abstract

OBJECTIVE: This study was designed to characterize the reproducibility of measurement for tumor volumes and their longest tumor diameters (LDs) and estimate the potential impact of using changes in tumor volumes instead of LDs as the basis for response assessments. *METHODS:* We studied patients with advanced lung cancer who have been observed longitudinally with x-ray computed tomography in a multinational trial. A total of 71 time points from 10 patients with 13 morphologically complex target lesions were analyzed. A total of 6461 volume measurements and their corresponding LDs were made by seven independent teams using their own work flows and image analysis tools. Interteam agreement and overall interrater concurrence were characterized. *RESULTS:* Interteam agreement between volume measurements was better than between LD measurements ($\iota = 0.945$ *vs* 0.734, $P = .005$). The variability in determining the nadir was lower for volumes than for LDs ($P = .005$). Use of standard thresholds for the RECIST-based method and use of experimentally determined cutoffs for categorizing responses showed that volume measurements had a significantly greater sensitivity for detecting partial responses and disease progression. Earlier detection of progression would have led to earlier changes in patient management in most cases. *CONCLUSIONS:* Our findings indicate that measurement of changes in tumor volumes is adequately reproducible. Using tumor volumes as the basis for response assessments could have a positive impact on both patient management and clinical trials. More authoritative work to qualify or discard changes in volume as the basis for response assessments should proceed.

Translational Oncology (2012) 5, 19–25

Introduction

X-ray computed tomography (CT) is often an effective imaging technique for assessing responses to treatment in patients with solid tumors. Qualitative impressions based on nothing more than visual inspections of the images are frequently sufficient for making major clinical management decisions. However, quantification becomes more crucial when treatment effects are not robust, for example, when tumor masses change only slowly over the course of illness or when the differences between two arms of a clinical trial are not large. The need to distinguish between measurement noise and small but biologically true changes in health status becomes particularly important when options exist for patients who are not responding adequately to their current therapeutic regimens. Ethically, these patients deserve access to alternatives as soon as confidence emerges that their current regimens are futile. Scientifically, objective radiologic evidence might be the best way to evaluate the effectiveness of investigational treatments whenever patients will switch to new therapeutic regimes, which will then confound

Address all correspondence to: P. David Mozley, MD, Merck Research Laboratories, 770 Sumneytown Pike, WP42-305, West Point, PA 19486-0004. E-mail: mozley@merck.com

[1]The original clinical trial was sponsored by Merck & Co, Inc.

Received 13 August 2011; Revised 25 October 2011; Accepted 25 October 2011

DOI 10.1593/tlo.11232

the use of survival time as an end point by exerting new, off-study influences on the course of their illness. In short, the field needs more sensitive measures of response [1], so do all of the other stakeholders in the treatment of individual patients with cancer [2]. Expensive therapeutic regimens would be more cost effective if they were stopped just as soon as evidence of futility emerged.

Most assessments rely on the Response Evaluation Criteria in Solid Tumors (RECIST) [3]. The current standard of care uses electronic calipers to measure a single, in-plane line length, the longest diameter (LD), as a proxy for the mass of a tumor. Simple measures of LD can be adequate [4,5]. Using LDs has advantages, including simplicity and the widespread access of health care workers to measurement tools that require very little technical training to use. However, concerns about the precision, accuracy, and sensitivity of using LDs as a quantitative measurement of tumor mass have been raised [6,7].

Some of these concerns could be addressed by semiautomated image analysis algorithms. For example, the RECIST 1.1 Work Group alluded to a future state in which the variability in tumor measurements could be decreased by "software tools that calculate the maximal diameter for a perimeter of a tumor" [8]. In theory, demarcating the boundary of a mass on every slice that it is visible, and then interrogating every slice to find the greatest distance between any two in-plane pixels, could improve both repeatability and reproducibility. It could eliminate some of the subjectivity in selecting the sole slice for measurement, decrease the judgment associated with how to draw the line, and reduce some of the factors that regulatory authorities have noted can adversely influence the placement of the calipers tips, such as display contrast, ambient room light, viewing angle, and others [9]. Moreover, although automation might not eliminate the variability associated with selecting the edge between neoplastic and normal tissue, it could stabilize the bias over time to facilitate the assessment of change. However, questions would still remain about how well any single line reflects the true tumor burden, particularly when the geometries of tumor masses are complex.

Many investigators have suggested that measuring the volume of the whole tumor could solve some of these problems and have clinically significant effects on patient management [10,11]. Indeed, a few studies have shown that volumetric image analysis (VIA) can add value [12,13]. However, technical problems have delayed the adoption of VIA [14]. Historically, substantial amounts of effort in time and manpower have been required for VIA. And some reports about the precision [15–18] and accuracy [19] of volume measurements have led to concerns that variability in volume measurements can be mistaken for medically meaningful changes, leading to errors in management.

Recent assessments of VIA technology are more optimistic [20]. One reported suggested that intrarater and interrater variability can be as little as 1% when analyzing well-demarcated tumors with simple geometric shapes in a single image set [21]. For more complex tumors, new algorithms can produce intrarater and interrater measurements of change with reliabilities of about ±5% on serially acquired image sets [22]. For patients with advanced lung cancer, within-subjects pairs of image sets acquired after very short time intervals in recent "coffee break" studies showed that short-term reproducibility was increased to 95% confidence intervals of about ±15%, provided that the masses were not too small [23]. If such a high level of reproducibility can be regularly achieved, then for many tumor morphologies, VIA would be substantially more sensitive than the current practice of using relatively large changes in LDs.

As encouraging as these results are, they do not directly address the question of whether VIA is a better method than LDs or automated measurements for managing individual patients with lung cancer or for making decisions in clinical trial settings. Also, most studies comparing the precision of LDs to VIA were done within single centers, were limited to three or fewer image analysts, and used manually placed electronic calipers to measure LD [11–14]. Accordingly, our investigation was designed as a proof-of-concept study. We conducted a head-to-head comparison between semiautomated diameter measurements, called auto-LDs, and VIA. We worked with a small subset of the data from a multinational clinical trial of an investigational new drug in which image quality tends to be variable and often less than ideal [24]. In conformance with current standards for evaluating novel diagnostic technologies [25], value was defined as the theoretical potential for a method to have a unique and meaningful effect on clinical trials or on individual patient management. Measurements were made to support or disprove several key hypotheses, which included these negative claims: (1) interrater reliability is higher for auto-LDs than whole tumor volumes, (2) VIA fails more often than auto-LDs because not all tumors have adequately demarcated boundaries on every slice, and (3) VIA increases costs, effort, and the amount of time required to analyze the images but has no added impact on patient management when compared to auto-LDs.

Materials and Methods

Context

The Quantitative Imaging Biomarker Alliance (QIBA) [26,27] constructed a "process map" [28] for either qualifying or discarding VIA as a useful method for improving the standard of care for patients with advanced stage lung cancer. This study was only one of many groundwork projects described by the qualification process map.

Patients

Seventy-one time points from 10 cases were retrospectively selected from a sample of 253 patients with stage IIIb or IV, non–small cell lung cancer. These subjects had volunteered to participate in a randomized, double-blind, clinical trial conducted at about 70 sites worldwide. They were all treated with a doublet chemotherapy regimen that represented the then current standard of care for years 2007 to 2009. They were then randomized to receive either an investigational new drug or placebo along with the doublet. The cases were selected from an imaging-only archive in the chronological order of their enrollment if, and only if, they had five or more analyzable CT scans after their baseline scan. No other information was known at the time of selection or ever made available to the image analysts.

CT Scans

Because the study was conducted at about 70 different sites worldwide, the image acquisition technique used the local standard of care as starting point for performing the scans. Then, the imaging manuals specified that all scans had to be acquired and processed with parameters that led to single breath hold images of the chest with reconstruction intervals of 5 mm or less without gaps. Sites were strongly encouraged, but not absolutely required, to use the same machine and identical image acquisition parameters each time they scanned their patients.

Target Lesions

Tumors were preselected as target lesions for this pilot study of technical feasibility if, and only if, they were predominantly in the

lung. The project manager assembled and distributed screenshots of the target lesions, which were marked up on only one slice of the baseline scan in PowerPoint slide format.

Image Analysts

The QIBA workforce for this particular project came from a diverse group with expertise in image processing and analysis, including two imaging CROs, an imaging device manufacturer, two image analysis software development companies, an academic cancer center, and a biopharmaceutical company.

The workforce was given no study-specific training or instructions other than to quantify the volumes of the marked tumors at each time point and semiautomatically generate the corresponding LDs. The "auto-LD" was predefined as the greatest in-plane distance between two pixels on the edges making up the boundaries around the tumor in the whole stack of tomographic images on which the target lesion was visible. Once the project began, there was no communication between any of the image analysis teams or any feedback from QIBA or the project manager.

Image Analysis

The sponsor and one software development company used the same image analysis tool. All of the other analyses were conducted with different tools that had been developed by the parent organizations of the analysts. All of the software tools deployed semiautomatic edge detection algorithms of one type or another. An operator could constrain or extend the boundaries as judgment indicated. Practicing radiologists supervised the analyses for the two CROs and the academic site. The remainder relied exclusively on analysts with no formal medical practice credentials.

All results were finalized by the individual image analysis teams, and then forwarded to the project manager, who summed the measures for patients with more than one target lesion, deidentified the image analysts, and then distributed the end points for statistical analysis. The two end points were the sum of LDs (auto-SLDs) for each marked target lesion and the corresponding sum of tumor volumes at each time point.

Statistical Analyses

For the purpose of simulation, an assumption was made that the date of the baseline scan corresponded to the day treatment started, and the last scan corresponded to the day the subject came off trial. All statistical analyses were conducted in with a commercially available software packaged called "R." Overall interrater agreement was assessed using the ι measure of agreement of Janson & Olsson. Individual interrater agreement was assessed using the concordance correlation coefficient. Interrater variability was assessed through (1) the coefficient of variation (CV) of the volumes per time point and (2) the SD of the differences between each postbaseline time point measurement and the baseline measurement divided by the baseline value.

The null hypothesis of equal variance with respect to determining the nadir of measurements with VIA or auto-SLD was assessed using Levene test. For the test, the difference between time to nadir for each patient and rater and the average time to nadir for each patient using the same type of measurement were used.

In an attempt to perform a fair comparison between VIA and auto-SLD, an analogous strategy of constructing categorical response variables was used for determining volumetric changes and disease progression. A negative change from baseline was used to define a partial response (PR), and a positive change from the nadir was used to declare progressive disease (PD). Volumetric thresholds for categorical responses were based on RECIST 1.1 thresholds and also were determined experimentally based on results for interrater reliability. Experimentally based thresholds were constructed by determining the SD of measurement among all image analysts. Cutoffs for PD and PR were then set symmetrically as ±2SD, respectively. For the auto-SLDs, RECIST 1.1 values were used. A 30% or greater decrease in the auto-SLD defined PR, whereas an increase of 20% or more defined PD.

Both time to PR as well as the theoretical progression-free survival (PFS) interval were measured. Results were computed for each image analysis team independently and then recalculated for pooled data from all teams.

Differences between the average auto-SLD values and average volumetric measurements were assessed using the log-rank test.

To identify and quantify systematic biases among image analysis teams and understand how they relate to each other, a hierarchical clustering based on average linkage was performed. A quantitative dendrogram, or tree diagram, was constructed [29]. The dendrogram was designed to show the arrangement of the clusters. It was constructed using 1 minus the concordance correlation coefficient as a metric [30].

Results

Subjects

All 10 subjects included in the analysis met diagnostic criteria for stage IV lung cancer. The mean time between the baseline scan and the last scan was 289 ± 64 days (median = 274 days, range = 213-391 days). Patients were scanned an average of 7.3 ± 1.4 times (median = 7 times, range = 6-10 times).

Characteristics of the Target Lesions

There was general agreement among teams that the target lesions were quite difficult to assess. Subjectively, there was incomplete confidence in the placement of some indistinct boundaries, particularly where masses wrapped around blood vessels, invaded the mediastinum, or were colocated with other pulmonary pathology, such as collapsed lung segments and effusions. Qualitatively, many of the target lesions appeared spiculated and asymmetrically lobulated, whereas 70% were attached to the pleura. Figure 1 shows typical examples.

A total of 6461 measurements of tumor volume and 6461 measurements of their corresponding LDs were made by the seven independent image analysis teams.

Using data from all teams showed that at *baseline*, the 13 total target lesions in this sample had an average sum of LDs (auto-SLD) of 6.6 ± 3.1 cm (median = 7.4 cm, range = 2.2-11.3 cm). During the entire study, the average auto-SLD from all 71 time points was 5.3 ± 2.2 cm (median = 5.3 cm, range = 2.1-11.3 cm).

Using data from all teams showed that at *baseline*, the target lesions in this sample had an average per patient volume of 106 ± 125 mL (median = 63 mL, range = 1.7-395 mL). During the entire study, the average per patient volume from all 71 time points was 42 ± 58 mL (median = 23 mL, range = 1.8-394 mL).

Technical Performance: Precision of Measurement

For auto-SLD measurements, the ι value describing overall interrater agreement was 0.734. Individual interrater agreement described

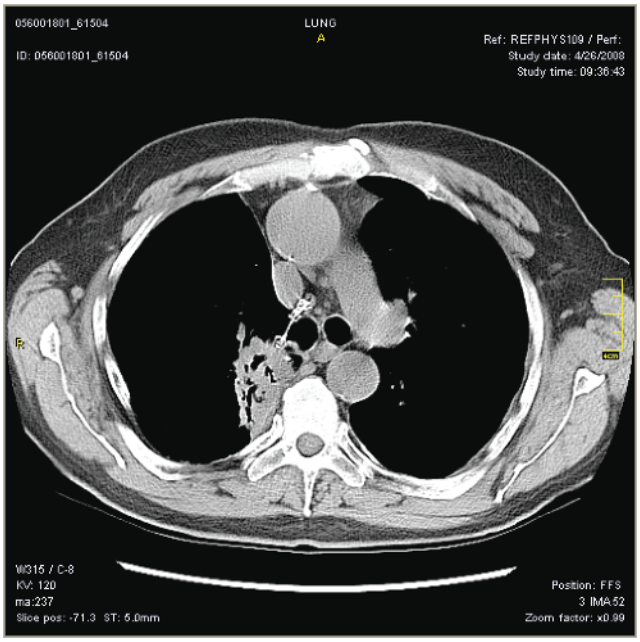

Figure 1. Typical mass. All patients in this sample had stage IV non–small cell lung cancer. Many of the target lesions were morphologically complex and associated with comorbid pulmonary disease.

by the concordance correlation coefficients varied between 0.58 and 0.96.

For the sum of volumes, overall interrater agreement was higher, as measured by an ı of 0.945. Individual interrater agreement described by the concordance correlation coefficients varied between 0.86 and 0.993.

Analyzing the agreement between absolute volumes at face value showed a mean CV of 25% and an upper 95% confidence interval on the CV of 59%. However, once normalized to baseline, the findings showed that differences between image analysis teams had an SD of only 12.4% (95% CI = −21.3% to 31.5%). When considering only measurements at which the time point volume was less than the baseline volume, the SD between image analysis teams was only 8.6% (95% CI = −16.7% to 18.7%). The overall dynamic range showed an SD from baseline of 39.9% (95% CI = −94.2% to 55.3%).

For auto-SLD data, the mean CV was 16.1% (upper 95% CI = 44.9%). After baseline normalization, the difference from average had an SD of 7.2% (95% CI = −14.1% to 13.5%) and the dynamic range an SD of 18.3% (95% CI = −63.1% to 12.9%).

Concordance Defining the Nadir

Examining the auto-SLD measurements showed that there were no cases in which all image analysis teams agreed on the exact time point that corresponded to the lowest value. There were two cases where the discrepancy was within only one time point, so that only two time points described the nadir, and four others where the discrepancy was within ±1 time point, so that three time points described the nadir. In one case, one team measured the nadir at baseline. Another team measured the nadir at the very last time point, which corresponded to the fifth postbaseline scan. Every time point from baseline to the end of treatment was selected as the nadir by at least one team. Taking the extreme values for all the cases showed that the maximum discrepancy per subject averaged 136 ± 70 days (median = 121 days, range = 56-291 days).

For volume measurements, there were two cases in which all seven of the image analyst teams agreed on the exact time point that corresponded to the volumetric nadir. The variance was one time point in two cases, and ±1 time points in the other six cases. These variances corresponded to relatively small changes in tumor volume in the middle of treatment, that is, at the bottoms of the wells of "U"-shaped response curves. Taking the extreme values showed that the maximum discrepancy per subject averaged 61 ± 39 days (median = 84 days, range = 0-97 days).

Variability in determining the time to the nadir with volume measurements was less than with auto-SLD measurements (P = .005; two-sided P value, Levene test).

Effect on Best Overall Response

The average auto-SLD values showed that, while on the trial, only 3 of the 10 patients would have met RECIST criteria for a PR.

For volumes, using 2 SDs as a criterion for change produced a threshold for PR of just less than a 25% decrease in volume. At a threshold that required more than 25% reduction in volume for the assessment of PR, VIA showed that 9 of the 10 patients met criteria for PR. If the threshold had required a change of more than 33%, then 8 of the 10 patients would have met criteria for PR. In one patient, tumor size continuously increased at every time point. In the nine others, the volume of the nadir was less than the volume of the baseline by a mean of 59% (median = 55%, range = 33%-95%).

Figure 2 shows the Kaplan-Meier curves for time to PR. Using the results of the log-rank test, we rejected the null hypothesis of no difference in time to PR by the average of the auto-SLD *versus* the average of the volumetric analysis (P = .004; two-sided P value, exact log-rank test).

Influence on PFS

Auto-SLD values for the selected target lesions showed that only 4 of the 10 subjects would have ever met RECIST criteria for PD while on trial.

Using the previously established threshold for change, that is, a greater than 25% increase in volume to declare PD, VIA showed that 8 of the 10 would have met criteria for PD, with an average increase from the nadir of 93% (median = 75%, range = 38%-242%). In one case, both SLDs and volumes suggested that the selected target lesions were continuously decreasing in size at every time point. In the other case, where all analysts agreed that the tumor burden increased after an initial response, three teams found increases of more than 25%, whereas four teams reported increases ranging from only 7% to 12%. The seven-team average was 14.7%.

In one case, the average auto-SLD values met RECIST criteria for PD at the same time as VIA. In all other cases, VIA showed PD before auto-SLD.

In eight of the nine cases where the target lesions increased in size during the trial, using volumes would have decreased RECIST defined PFS by an average of 62 ± 47 days (median = 42 days, range = 0-144 days).

Figure 3 shows the Kaplan-Meier curves for PFS. Based on the log-rank test for the whole sample, we rejected the null hypothesis of no difference between average volumes and average auto-SLD (P = .02; two-sided P value, exact log-rank test).

Systematic Biases

Hierarchical clustering of all quantitative measurements revealed that biases among the image analysis teams were consistent for measurements

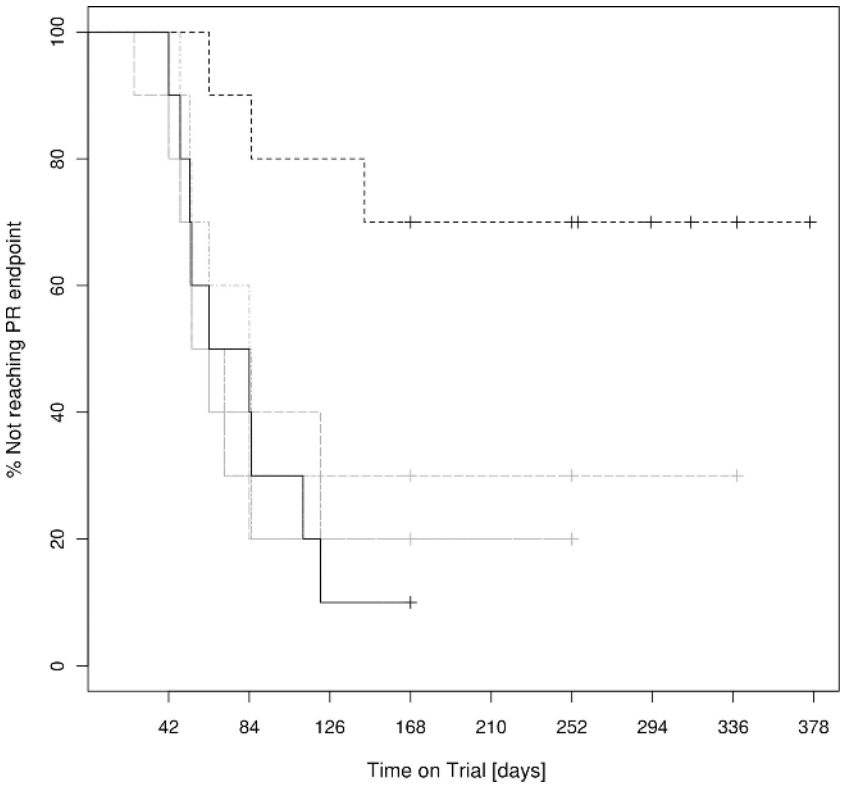

Figure 2. Kaplan-Meier curves of PR. Kaplan-Meier curves show that using average changes in volume (black solid line) instead of average changes in auto-SLD (black dashed line) as the basis for categorical response classifies significantly more patients as partial responders (PR). The Kaplan-Meier curves in gray represent volumetric responses produced by each of the seven independent image analysis teams. There appear to be fewer than seven gray curves because there was so much overlap between teams. Censoring events are shown as (+).

of both volumes and auto-SLDs. For example, one image analysis team produced estimates of absolute volumes that were lower than the median 97% of the time, whereas their corresponding auto-SLD measurements were lower than the median 85% of the time. In contrast, another team produced estimates that were higher than or equal to the median 85% of the time for volumes and 82% of the time for auto-SLD. The image analysis teams that were supervised by expert radiologists consistently clustered in either the highest- or the lowest-volume groupings, whereas those teams without expert radiologist supervision consistently clustered together in the middle.

Although an auditable trail of the time required was not kept, all of the teams estimated that it took between 5 and 20 minutes to import the images and export the results, depending on lesion complexity.

Discussion

The results show that there is at least one context in which volumes outperform auto-SLDs as the input measurement for RECIST-based assessments. In this population, which had geometrically complex tumors and comorbid pulmonary disease, changes in volumes were more consistently measured at all time points than were their corresponding auto-SLDs. As a consequence, volumes were more sensitive for detecting favorable tumor responses as well as for diagnosing disease progression. A majority of the patients who had to be right censored at the time they came off trial because their auto-SLDs never extended beyond the stable disease range would have been reclassified if volumes had been used as the basis for their response assessments.

In our attempt to determine whether measuring tumor volumes can add more value than estimating the LDs, we felt compelled to use the standard range for auto-SLDs established by RECIST. However, for the new measure of volumes, we allowed the data to determine the threshold for distinguishing between measurement noise and a likely change in health status. In this sample, a change in the sum of tumor volumes that was greater than 25% carried only a 2.5% risk of incorrectly diagnosing PD. This seems like an acceptable level of risk for patients with other options for treatment. Using a threshold of 30% would not have changed the results. Using a threshold of 40% would not have changed the conclusions. It seems plausible to assume that the threshold for having adequate confidence in the volume measurements will need to be reanalyzed in many different contexts and may vary with the type of tumor being treated. The thresholds we found for changes in tumor volume in this sample do not challenge the thresholds described by RECIST for LDs. These target lesions were geometrically complex. Simple extrapolations between three dimensions and one dimension are known to work well in some situations where tumor morphology is simple but would be precarious in settings like this. In fact, using the cubic root of the volumes to suggest a new threshold for LDs would be fallacious in the context we studied because the new value would fall well within the range for noise of auto-SLDs.

Volumes outperformed auto-SLD despite the fact that image resolution was not high by current standards. Had we strived for higher image resolution and tighter quality control, we would have, according to mounting evidence, reduced variance even further [12]. Because image

quality in many standard-of-care settings has already surpassed what was produced in the parent multinational trial from which we drew these cases, VIA should be effective in many new clinical trials and, eventually, in many clinical practice settings.

The variance in this project is probably not as low as what can be achieved in many research settings. This is not surprising, as few instructions were given to the image analysts, and there was no trial-specific training. Indeed, the dendrogramatic analysis of systematic bias suggests that human judgment is a major contributor to variance. Nevertheless, the level of consistency in rank order among image analysis teams was highly similar, regardless of whether the end point was whole tumor volume or auto-SLD. This finding is encouraging because it suggests that biases can be characterized before trials begin with training cases like these. Pretrial identification of bias can increase our understanding of the causes of variance, especially the components due to judgments by image analysts. These biases can then be minimized. Reader training that reduces bias should increase the interrater measurement reproducibility. As a consequence, this should increase the sensitivity of VIA as a key method for monitoring changes in tumor mass and their responses to therapeutic interventions in both clinical and experimental settings [31].

Although the sample size was small, our results are encouraging. They were based on over 6000 measurements of tumor volume and their corresponding auto-LDs by the seven teams, which might be enough to begin suggesting that there is a context in which multiple image analysts working independently can agree on the magnitude of relatively small-to-moderate changes in tumor volumes. Agreement might have been higher if the effects of treatment had been greater. Of course, measures of performance could have been worse if the tumor changes were smaller, if the tumors were even more complex in shape, if contrast was decreased even further by surrounding pathology, and for other reasons. Nevertheless, the challenge these cases presented suggests that even advanced lung cancer can be a favorable disease setting for quantifying volumes. Although consistent replication in larger samples and whole clinical trials will be required to qualify the methodology, the findings suggest that volumes might be more robust than auto-SLDs as inputs for RECIST in other neoplastic diseases as well, especially when tumor morphology and contrast are favorable.

This pilot study suggests that value is reproducible with a variety of software tools. Using volumes as the basis for RECIST should improve decision making, both in the care of individual patients and in the management of clinical trials. The increased sensitivity provided by whole tumor volumes could lead to an important paradigm shift, especially in settings where patients have multiple options for alternative treatments. When relying on manually measured SLDs, relatively large sample sizes are sometimes needed to determine whether a new treatment is biologically active and deserves to be advanced in development. VIA could reduce the number of subjects required, as well as the time-on-trial per subject and, as a result, either speed the advancement of promising treatments toward the market or hasten the elimination of not-so-promising treatments. VIA could also enhance RECIST-based assessment strategies in clinical settings because the greater sensitivity for diagnosing disease progression could benefit individual patients who have treatment alternatives available to them. In these situations, VIA could well serve all stakeholders in the treatment of cancer.

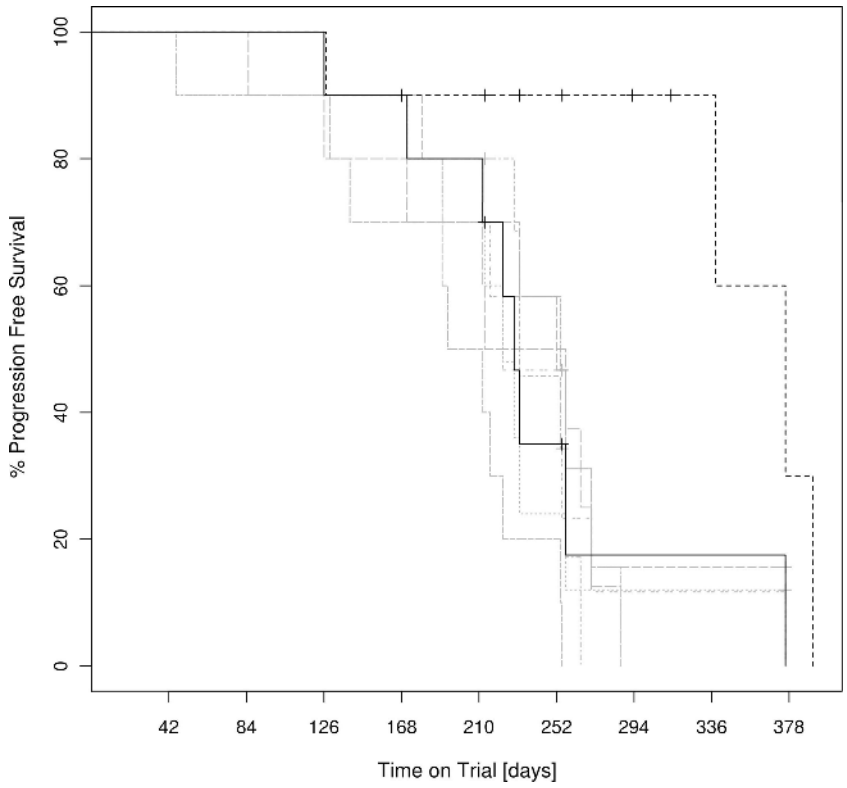

Figure 3. Kaplan-Meier curves of PD. Kaplan-Meier curves show that average changes in volume (black solid lines) perform significantly better than average changes in auto-SLD (black dashed lines) for classifying PD. For volumes, a response of PD represents an increase of +25%; for auto-SLDs, the standard RECIST value of +20% is shown. Kaplan-Meier curves in gray represent the response assessments for PD produced by the individual image analysis teams. Censoring events are shown as (+).

Acknowledgments

This study was made possible by a voluntary workforce from the Quantitative Imaging Biomarker Alliance of the Radiological Society of North America. The authors thank Dr Jeremy Z. Fields for editing the article.

References

[1] Woodcock J and Woosley R (2008). The FDA Critical Path Initiative and its influence on new drug development. *Ann Rev Med* **59**, 1–12.

[2] Petrick N, Brown DG, Suleiman O, and Myers KJ (2008). Imaging as a tumor biomarker in oncology drug trials for lung cancer: the FDA perspective. *Clin Pharmacol Ther* **84**, 523–525.

[3] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz L, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* **45**, 228–247.

[4] Therasse P, Eisenhauer EA, and Verweij J (2006). RECIST revisited: a review of validation studies on tumour assessment. *Eur J Cancer* **42**, 1031–1039.

[5] Prasad SR, Jhaveri KS, Saini S, Hahn PF, Halpern EF, and Sumner JE (2002). CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques—initial observations. *Radiology* **225**(2), 416–419.

[6] Schwartz LH, Curran S, Trocola R, Randazzo J, Ilson D, Kelsen D, and Shah M (2007). Volumetric 3D CT analysis—an early predictor of response to therapy. *J Clin Oncol* **25**(18S). ASCO Annual Meeting Proceedings Part I. Abstract 4576.

[7] Suzuki C, Jacobsson H, and Hatschek T (2008). Radiologic measurements of tumor response to treatment: practical approaches and limitations. *Radiographics* **28**, 329–344.

[8] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz L, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* **45**, 228–247. Appendix II. Measurement of lesions; p. 243.

[9] US Food and Drug Administration (2011). *Guidance for Industry. Standards for Clinical Trial Imaging Endpoints*. Available at: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm. Accessed October 4, 2011.

[10] Moertel CG and Hanley JA (1976). The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer* **38**, 388–394.

[11] Quivey JM, Castro JR, Chen GT, Moss A, and Marks WM (1980). Computerized tomography in the quantitative assessment of tumour response. *Br J Cancer Suppl* **4**, 30–34.

[12] Munzenrider JE, Pilepich M, Rene-Ferrero JB, Tchakarova I, and Carter BL (1977). Use of body scanner in radiotherapy treatment planning. *Cancer* **40**, 170–179.

[13] van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, van Iersel CA, van den Bergh KAM, van't Westeinde S, van der Aalst C, et al. (2009). Management of lung nodules detected by volume CT scanning. *N Engl J Med* **361**, 2221–2229.

[14] Mozley PD, Schwartz LH, Bendtsen C, Zhao B, Petrick N, and Buckler AJ (2010). Change in lung tumor volume as a biomarker of treatment response: a critical review of the evidence. *Ann Oncol* **21**, 1751–1755.

[15] Petrou M, Quint LE, Nan B, and Baker LH (2007). Pulmonary nodule volumetric measurement variability as a function of CT slice thickness and nodule morphology. *Am J Radiol* **188**, 306–312.

[16] Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ, de Bock GH, Gietema HA, Xu DM, Leusveld ALM, de Koning HJ, Scholten ET, Verschakelen J, et al. (2008). Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. *Radiology* **248**, 625–631.

[17] Bogot NR, Kazerooni EA, Kelly AM, Quint LE, Desjardins B, and Nan B (2005). Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. *Acad Radiol* **12**, 948–956.

[18] Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, and Munden RF (2003). Interobserver and intraobserver variability in measurement of non–small-cell carcinoma lung lesions: implications for assessment of tumor response. *J Clin Oncol* **21**, 2574–2582.

[19] Winer-Muram HT, Jennings SG, Meyer CA, Liang Y, Aisen AM, Tarver RD, and McGarry RC (2003). Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations. *Radiology* **229**, 184–194.

[20] Mulshine JL and Jablons DM (2009). Volume CT for diagnosis of nodules found in lung-cancer screening. *N Engl J Med* **361**, 2281–2282.

[21] Goodman LR, Gulsun M, Washington L, Nagy PG, and Piacsek KL (2006). Inherent variability of CT lung nodule measurements *in vivo* using semiautomated volumetric measurements. *Am J Radiol* **186**, 889–994.

[22] Bendtsen C, Kietzmann M, Korn R, Mozley PD, Schmidt G, and Binnig G (2011). X-ray computed tomography: semiautomated volumetric analysis of late-stage lung tumors as a basis for response assessments. *Int J Biomed Imaging* **2011**, 361589.

[23] Zhao B, Schwartz LH, Steve M, and Larson SM (2009). Imaging surrogates of tumor response to therapy: anatomic and functional biomarkers. *J Nucl Med* **50**, 239–249.

[24] Ramalingam SS, Parise RA, Ramananthan RK, Lagattuta TF, Musguire LA, Stoller RG, Potter DM, Argiris AE, Zwiebel JA, Egorin MJ, et al. (2007). Phase I and pharmacokinetic study of vorinostat, a histone deacetylase inhibitor, in combination with carboplatin and paclitaxel for advanced solid malignancies. *Clin Cancer Res* **13**, 3605–3610.

[25] Committee for Medicinal Products for Human Use (CHMP) (2009). *Guideline on Clinical Evaluation of Diagnostic Agents*. European Medicines Agency. Available at: http://www.emea.europa.eu. Accessed September 7, 2009.

[26] Buckler AJ, Mozley PD, Schwartz L, Petrick N, McNitt-Gray M, Fenimore C, O'Donnell K, Hayes W, Kim HJ, Clarke L, et al. (2010). Volumetric CT in lung cancer: an example for the qualification of imaging as a biomarker. *Acad Radiol* **17**, 107–115.

[27] Buckler AJ, Mulshine JL, Gottlieb R, Zhao B, Mozley PD, and Schwartz L (2010). The use of volumetric CT as an imaging biomarker in lung cancer. *Acad Radiol* **17**, 100–106.

[28] Radiological Society of North America. Available at: http://qibawiki.rsna.org/index.php?title=Main_Page. Accessed September 7, 2009.

[29] Available at: http://en.wikipedia.org/wiki/Dendrogram. Accessed April 18, 2010.

[30] Lin LI (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268.

[31] Bolte H, Jahnke T, Schäfer FK, Wenke R, Hoffmann B, Freitag-Wolf S, Dicken V, Kuhnigk JM, Lohmann J, Voss S, et al. (2007). Interobserver-variability of lung nodule volumetry considering different segmentation algorithms and observer training levels. *Eur J Radiol* **64**, 285–295.