Published in final edited form as:

Eur J Radiol. 2013 June ; 82(6): . doi:10.1016/j.ejrad.2013.02.018.

Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals

Binsheng Zhao^{a,*}, Yongqiang Tan^{a,2}, Daniel J. Bell^{b,3}, Sarah E. Marley^{c,4}, Pingzhen Guo^{a,1}, Helen Mann^{c,5}, Marietta L.J. Scott^{c,6}, Lawrence H. Schwartz^{a,7}, and Dana C. Giorghiu^{c,8}

Binsheng Zhao: bz2166@columbia.edu; Yongqiang Tan: yt2314@columbia.edu; Daniel J. Bell: doctordj@gmail.com; Sarah E. Marley: sarahthepurplelady@hotmail.com; Pingzhen Guo: jn@sina.com; Helen Mann: helen.mann@astrazeneca.com; Marietta L.J. Scott: marietta.scott@astrazeneca.com; Lawrence H. Schwartz: lhs2120@columbia.edu; Dana C. Giorghiu: dana.giorghiu@astrazeneca.com

^aDepartment of Radiology, Columbia University Medical Center, New York, NY, USA

^bNorth Middlesex University Hospital, London, United Kingdom

^cAstraZeneca, Alderley Park, Macclesfield, Cheshire, United Kingdom

Abstract

Objective—Understanding magnitudes of variability when measuring tumor size may be valuable in improving detection of tumor change and thus evaluating tumor response to therapy in clinical trials and care. Our study explored intra- and inter-reader variability of tumor uni-dimensional (1D), bi-dimensional (2D), and volumetric (VOL) measurements using manual and computer-aided methods (CAM) on CT scans reconstructed at different slice intervals.

Materials and methods—Raw CT data from 30 patients enrolled in oncology clinical trials was reconstructed at 5, 2.5, and 1.25 mm slice intervals. 118 lesions in the lungs, liver, and lymph nodes were analyzed. For each lesion, two independent radiologists manually and, separately, using computer software, measured the maximum diameter (1D), maximum perpendicular diameter, and volume (CAM only). One of them blindly repeated the measurements. Intra- and inter-reader variability for the manual method and CAM were analyzed using linear mixed-effects models and Bland–Altman method.

Results—For the three slice intervals, the maximum coefficients of variation for manual intra-/inter-reader variability were 6.9%/9.0% (1D) and 12.3%/18.0% (2D), and for CAM were 5.4%/9.3% (1D), 11.3%/18.8% (2D) and 9.3%/18.0% (VOL). Maximal 95% reference ranges for the

© 2013 Elsevier Ireland Ltd. All rights reserved.

*Corresponding author at: Department of Radiology, Columbia University Medical Center, 710 West 168th Street, New York, NY 10032, USA. Tel.: +1 212 342 1393.

¹Current address: The Second Hospital of Hebei Medical University, Shijiazhuang, Hebei 050000, China.

²Department of Radiology, Columbia University Medical Center, 710 West 168th Street, New York, NY 10032, USA.

³North Middlesex University Hospital, London, Greater London N18 1QX, United Kingdom.

⁴Flat 3, 113 Old Tiverton Road, Exeter, Devon EX4 6LD, United Kingdom. Tel.: +44 01392 21539/625 516730.

⁵AstraZeneca, Statistics and Informatics, Clinical Information Science, 90F169-1, East Wing, Parklands, Alderley Park, Macclesfield, Cheshire SK10 4TG, United Kingdom.

⁶AstraZeneca, Personalised Healthcare and Biomarkers, Alderley Park, Macclesfield SK10 4TG, United Kingdom.

⁷Department of Radiology Columbia, University Medical Center, 180 Fort Washington Avenue, New York, NY 10032, USA.

⁸AstraZeneca, Oncology Global Medicine Development - 41, 11G Mereside, Alderley Park, Macclesfield, Cheshire, SK10 4TG, United Kingdom.

Conflict of interest

No conflict of interest has been declared.

percentage difference in intra-reader measurements for manual 1D and 2D, and CAM VOL were (−15.5%, 25.8%), (−27.1%, 51.6%), and (−22.3%, 33.6%), respectively.

Conclusions—Variability in measuring the diameter and volume of solid tumors, manually and by CAM, is affected by CT slice interval. The 2.5 mm slice interval provides the least measurement variability. Among the three techniques, 2D has the greatest measurement variability compared to 1D and 3D.

Keywords

Computed tomography (CT); Response evaluation criteria in solid tumors (RECIST); World Health Organization (WHO); Volume measurements; Measurement variability

1. Introduction

World Health Organization (WHO) criteria for response assessment [1,2] and, more recently, the Response evaluation criteria in solid tumors (RECIST) [3,4] are imaging-based guidelines that are widely implemented in clinical trials and also used in clinical practice to determine patient outcomes. In these criteria, tumor response assessment is based upon measured change in tumor diameter(s) determined on cross-sectional modalities such as computed tomography (CT) and magnetic resonance imaging (MRI). However, the response cut-off values (e.g., diameter decreases 30% are considered as partial responses by RECIST) were developed by evaluating the measurement error of antiquated response assessment modalities used during the 1970s and early 1980s (i.e., physical palpation or plain X-ray measurements) [5,6]. For 30 years, despite the ubiquitous use of sectional imaging techniques in response assessment and the remarkable progress that has been made in these technologies and image analysis algorithms, limited data have been published on quantifying measurement variability of these new technologies and techniques for response assessment.

Recently, a same-day repeat CT study reported the measurement variability in non-small cell lung cancer (NSCLC) on thin slice CT images and its potential impact on response assessment [7,8]. The study found much lower magnitudes of measurement variability/error compared to the cut-off values set by the RECIST and WHO criteria to define tumor response. Indeed, with the latest imaging techniques and tumor size measurement tools, it should be possible to quantify much smaller tumor changes earlier during the treatment, and thus provide earlier assessment of patient outcome.

Volumetric techniques have begun to demonstrate advantages over uni-dimensional (1D) measurement in the response assessment of NSCLC tumors treated with targeted therapies and in the discovery of tumor biomarkers [9,10]. With the growing number of computer algorithms available for facilitating tumor volume (VOL) calculation, and as clinical trials begin to incorporate tumor VOL as an exploratory endpoint, it is now critical to optimize and validate the use of computer-aided volumetric technique in such trials.

Understanding the magnitude of measurement variability is essential for establishing a new volumetric response method and re-evaluating conventional 1D and 2D methods to better detect tumor changes with therapies, especially novel targeted therapies that may not produce tumor changes at the same magnitude or speed on radiographic images as observed in those treated with classic cytotoxic chemotherapies. Another critical issue for clinical trials that are incorporating volumetric response method is the proper (optimal) use of imaging acquisition/reconstruction parameters (e.g., slice interval). Can we use the current standard slice interval (thickness) of 5 mm to adequately measure tumor volumes? Would a thinner slice interval of 2.5 mm or 1.25 mm provide more reliable uni-, bi-dimensional and

volumetric measurements for assessing therapy response? Determining optimal/uniform CT imaging acquisition parameters is an urgent task that would help improve interpretation of measurement results obtained especially from those early phase multicenter clinical trials in which a small number of patients are enrolled.

This study was therefore designed to (1) quantify the level of intra-/inter-reader variability in 1D, 2D, and VOL measurements of solid tumors in the lungs, liver and lymph nodes, the three most frequently involved metastatic sites, on the three most commonly used CT slice intervals (5, 2.5 and 1.25 mm), and (2) assess the agreement between manual and computer-assisted methods (CAM) for tumor measurement.

2. Materials and methods

2.1. Clinical imaging data

This study was conducted as a prospective exploratory project utilizing data acquired according to protocol-scheduled tumor assessments. This ancillary, HIPAA compliant study was approved by the institutional review board, with informed patient consent waived.

CT imaging scan data collected from 30 patients, enrolled between May 2007 and June 2009 in several Phase I or II oncology clinical trials evaluating systemic therapies in melanoma, lymphoma, lung, pancreatic, and colorectal cancer patients at a single institution, were included in this study. Routine clinically acquired contrast-enhanced (single phase imaging, i.e., portal-venous phase, was performed), diagnostic CT scans (MDCT; LightSpeed 16 or 64-detector row; GE Healthcare, Milwaukee, WI) with a slice collimation of 5 mm were used. Raw data were reconstructed at 5 mm slice interval using the standard algorithm, defined in the clinical trial protocol, and also at 2.5 and 1.25 mm as per this ancillary study protocol. Only one scan per patient (baseline or the first follow-up scan where the reconstructed images of the three slice intervals were all available) was included in the study.

2.2. Lesion selection and lesion size measurement

According to RECIST (version 1.0) criteria for target lesion selection and measurement, an clinical trial expert participating radiologist (LHS) selected a total of 118 measurable lesions in the lungs, liver and lymph nodes from the imaging data of the 30 patients. The number of lesions per patient ranged from 1 to 10 (median: 3.5), with a maximum of 5 lesions per organ. Lesions were selected based on the measurability and reproducibility criteria, as described by RECIST [3]. The lesion distribution per anatomical location is shown in Table 1. Although RECIST suggests that the target lesions at baseline should not be less than 10 mm in longest diameter, some target lesions, especially those smaller than 2 cm on baseline study, can become smaller than 10 mm at follow-up examinations due to treatment effect. We included a small fraction of lesions (19/118) that were less than 10 mm in longest diameter at CT scan in order to assess the measurement variability for these lesions.

Lesion measurements were performed by: (1) manual measurement by a radiologist using electronic calipers on an imaging viewing workstation and (2) using a CAM. Three in-house algorithms for semi-automated segmentation of lesions in lung, liver and lymph nodes were integrated into an image viewing and manipulation system developed using Interface Description Language (IDL Solutions, Germantown, WI) which facilitated both manual and CAM lesion measurements. The algorithms were developed mainly based on the marker-controlled watershed segmentation, followed by the boundary smoothing techniques [11].

2.2.1. Manual method—For each lesion, the radiologist selected the most appropriate in-plane slice and measured the maximal diameter (1D) and the maximal perpendicular

diameter according to RECIST and WHO criteria. 2D measurement was the product of the maximal diameter and its maximal perpendicular diameter. Standard window/level (Hounsfield Unit) settings of lung (1500/−500), liver (150/90), and chest/abdomen (340/60) were chosen for manually measuring lesions. Manual inter-reader variability was assessed using the independent measurements of reader 1 (PG, radiologist with 25 years' experience of interpreting CT images) and reader 2 (DB, radiologist with 6 years' experience of interpreting CT images). Manual intra-reader variability was evaluated when reader 1 blindly repeated all measurements using the same methodology in a separate session.

2.2.2. CAM method—For each lesion, an operator applied the appropriate tumor-specific computer algorithm to segment the lesion in three dimensions. Subsequently, the computer-generated lesion contours were uploaded and overlaid on the lesion on original CT images for inspection by the radiologist using the standard window/level settings. If the computer result was unsatisfactory, the radiologist corrected the contour and saved this as the final result. Based on the final contours, the maximal diameter (1D), maximal perpendicular diameter, 2D and VOL of the lesion were automatically computed.

Intra-CAM variability was assessed when one operator applied the segmentation algorithms to all of the selected lesions. Reader 1 then reviewed and corrected the computer results as necessary and saved them as the final CAM results. Over a month later, this same radiologist, blinded, reviewed and corrected all of the computer-generated lesion contours. Inter-CAM variability was evaluated when another operator independently applied the segmentation algorithms to all lesions. The computer-generated lesion contours were reviewed and edited by reader 2, who was blind to the first radiologist's results.

For both methods, measurements were performed for all lesions, first on 5 mm, then 2.5 mm and finally 1.25 mm images. The time interval between one radiologist's measurements on two different slice images was longer than one week.

Fig. 1 shows an example of a liver metastasis depicted and independently measured by two radiologists on the three different slice interval images.

3. Statistical analysis

Analyses were performed with SAS software (version 8.1; SAS Institute, Cary, NC). Separate analyses were performed to obtain the CAM and manual, intra- and inter-reader components of variability for 1D, 2D and CAM VOL measurements at each slice interval. Linear mixed-effects analysis of variance (ANOVA) models [12] were fitted to the natural logarithm of the tumor lesion measurements for each patient. Patient and lesion were included as random effects in the model to account for the repeated measurements. The inter-patient, inter-lesion (intra-patient), and residual components of variation were obtained from the model and presented as a standard deviation on the log scale (s) and a coefficient of variation (COV) [13], calculated as $100 \times \sqrt{(\exp(s^2) - 1)}$ [14]. For the intra-reader analyses the residual component of variability incorporated the variability between repeat reads for reader 1; for the inter-reader analyses the residual component of variability incorporated the variability between reader 1 and reader 2's duplicate measurements.

In addition to the lesion-level analyses described above, analyses were also performed at the patient level; models were fitted to the logarithm of the tumor burden for each patient (sum of the measurements of individual patient lesions). Patient was included as a random effect in the model for these analyses.

For each slice interval, at both lesion and patient-level, Bland–Altman analyses [15] were used to assess agreement in the manual and CAM measurements, between the two blinded readings by the same radiologist, between the readings by the two independent radiologists, and between manual and CAM measurements. Ninety-five percent reference ranges (i.e., limits of agreement) were obtained using log transformed measurements and calculated as $d \pm 1.96(2)s$, where d is the mean of the difference between the measurements and s is the residual standard deviation from the relevant ANOVA model described above. The reference ranges were back transformed to give limits for the percentage difference between the measurements.

4. Results

One hundred eighteen (118) lesions in 30 patients were selected for the study. Overall, lesion diameters (averaged over the three slice intervals) ranged from 6.1 mm to 80.1 mm, with a median of 18.4 mm.

Reader 2 carried out 708 measurements (118 lesions $\times$ 3 slice intervals $\times$ 2 manual measurements) and reader 1, who performed the repeat measurements, carried out 1416 measurements (708 measurements $\times$ 2). Four out of 708 (0.6%) of reader 2's manual measurements were missing. Given their small number and that they were missing at random, missing values were not imputed in the statistical analyses.

4.1. Lesion-level analyses

4.1.1. Manual intra- and inter-reader variability—Table 2(a) shows the components of variability for the manual intra-reader analysis. As expected, due to the variable shape and location of different types of tumors, the largest component of variability was between lesions from the same patient. The intra-reader components of variability were small and similar across different slice intervals.

Similar to the intra-reader, the largest component of variability for the manual inter-reader analysis was between lesions from the same patient (Table 2(b)). For all measurements, the inter-reader variability was higher than the intra-reader variability and fluctuated across different slice intervals. The 2.5 mm images showed the smallest inter-reader component and comparable intra-reader component of variability compared with the other two slice intervals (1.25 mm and 5 mm). The reasons could be lower signal-to-noise ratio on 1.25 mm images, especially the abdominal images, and larger partial volume effect on 5 mm images, especially for smaller lesions, making 1.25 mm and 5 mm slice intervals less optimal for measuring tumors spread over the body.

For all measurements, there was greater intra-reader agreement than inter-reader agreement (Table 3 and Fig. 2). 1D changes exceeding the 95% reference range of -15.5% to 25.8% and -19.4% to 36.3% were considered outside the underlying intra- and inter-reader's measurement errors, respectively, across all slice intervals. For both 1D and 2D measurements, mean differences between the intra-reader measurements increased with increasing slice interval, whereas mean differences between the inter-reader measurements decreased with increasing slice interval.

4.1.2. CAM intra- and inter-reader variability—For the CAM intra- and inter-reader analyses, the largest component of variability was again between nodules from the same patient (Table 4). The intra-reader components of variability were small and similar across different slice intervals; the COV was smallest for the 1D measurements, followed by the VOL and 2D measurements (Table 4(a)). Compared with the intra-reader variability, the inter-reader variability was larger for all measurements and was less consistent across

different slice intervals. For all measurements, the inter-reader variability was highest on 5 mm and lowest on 2.5 mm slice interval images (Table 4(b)). For CAM VOL measurements, changes exceeding the 95% reference ranges of -22.3% to 33.6% and -36.3% to 72.7% could be considered as outside the underlying intra- and inter-variability, respectively, across all slice intervals (Table 5 and Fig. 3).

For 1D, 2D and 3D, smaller lesions (<10 mm) had larger intra- and inter-reader variability compared to larger lesions (≥ 10 mm) on the thick slice interval images (Figs. 2 and 3).

4.1.3. Manual vs CAM techniques—For 1D and 2D measurements on all slice intervals, the intra-reader variability was lower than the inter-reader for both CAM and manual techniques (Table 6). The intra-reader variability was slightly lower for CAM than manual measurements on all slice intervals. However, the CAM and manual inter-reader variability did not show the same trend: the inter-reader variability was improved by the CAM on 1.25 mm images, but worsened on 5 mm images, and was almost unaffected on 2.5 mm images.

The assessment of the level of agreement between the manual and CAM measurements (1D and 2D) for each of the two readers showed, on average, manual 1D measurements were generally smaller than CAM 1D measurements (Table 7). It is possible that the computer always identified and connected two boundary pixels that had the absolute maximal distance, whereas the radiologists were not able to do so. On the other hand, however, computer measurements are more easily affected by imperfect segmentation as radiologists can sometimes more efficiently exclude attached (normally small) non-lesion parts from the lesion. Better agreement between manual and CAM was found for the more experienced vs less experienced radiologist.

4.2. Patient-level analyses

Compared with the lesion-level analyses, the patient-level data showed considerably lower intra- and inter-reader COVs and narrower 95% reference ranges (Tables 8 and 9).

5. Discussion

Variability in tumor VOL measurements due to CT scanner, dose, imaging acquisition and reconstruction parameters and segmentation algorithm has previously been studied mainly with a focus on pulmonary nodules detected in CT lung cancer screening programs and routine clinical care for the purpose of non-invasive diagnosis [16–19]. However, in those clinical settings, nodules are usually small (<2 cm with an average size <10 mm) and image spatial resolutions are high (≤ 1 mm in *z*-axial direction). The findings may not be applicable to the clinical trial setting where solid tumors are often large in size (average approximately 40 mm) and can originate from, and spread to, any organ. This is particularly true for lesions in the abdomen or pelvis, where images at a high resolution can be very noisy and lesion-to-surrounding contrast can be low.

While limited studies have addressed the issue of tumor measurement variability in response assessment settings, few have put the results in the context of metrics such as RECIST. Punnen et al. assessed the inter- and intra-reader variability in the linear measurements on 2.5 mm slice interval using renal tumors (<40 mm) and found that error ranges of the linear measurement of renal tumors were ± 3.1 mm for multiple readers and ± 2.3 mm for a single reader [20]. Hopper et al. reported inter- and intra-observer variability of up to 15% and 6%, respectively, for maximum diameter when interpreting thoracic and abdominal/pelvic contrast CT scans on 8 mm slice interval [21]. Using RECIST and WHO response category systems, Erasmus et al. found that the highest misclassification rates could be 43% (2D) and

30% (1D) for progression when non-small-cell lung cancer (NSCLC) tumors were measured by multiple radiologists, whereas the lowest misclassification rates could be 2.5% (1D) and 3.0% (2D) for response if the tumors were measured by the same radiologist [22]. Recently, Zhao et al. reported a same-day repeat CT study, revealing the level of the minimum detectable change in measuring NSCLC on 1.25 mm thin section MDCT [7]. According to their study, changes in CAM tumor 1D, 2D, and VOL measurements of 8%, 20%, and 14% or greater exceeded the measurement variability and could be considered real tumor changes.

In this study, we designed and conducted a series of measurement experiments that allowed us to explore the magnitude of tumor measurement variability in a contemporary clinical and clinical trial setting. The measurements were carried out both manually and utilizing three in-house image segmentation methods developed for lung lesions, liver and lymph node metastasis (the three most common metastatic tumor types) and evaluated at three slice intervals, including the current clinical standard (5 mm) and future potentials (2.5 mm and 1.25 mm).

For all measurements, with or without the use of computer algorithms, intra-reader variability was lower than inter-reader variability on all slice interval images studied. This confirms the previous finding that in order to reduce measurement variability a single radiologist should follow up or measure the tumor(s) in the same patient throughout the course of therapy [22]. Our study extends the finding to the CAM 1D, 2D and true VOL measurements.

Our data further suggest that all three measurements made either manually (only 1D and 2D) or by CAM have similar and low intra-reader variability across the three different slice thicknesses, but have the lowest inter-reader variability on 2.5 mm slice images compared with 1.25 mm and 5 mm images. This suggests that the 2.5 mm slice interval may provide the best viewing condition for radiologists to measure tumor diameters and the most robust condition for radiologists to use CAMs to measure tumor volumes. The small impact that slice thickness has on intra-reader variability is further support for having the same reader review all of a single patient's scans or even all patients' scans.

In order to establish the criteria for volumetric response assessment, setting proper response and progression cut-off values is essential. Based on the intra-reader variability in VOL measurement (the 95% reference range), our data suggest that the cut-off values for recognizing tumor VOL changes by CAM should not be set lower than 23% for decrease and 34% for increase (Table 5). The VOL change cut-off levels of 65% for decrease and 73% for increase as suggested by the RECIST criteria for assessing tumor response and progression (assuming spherical tumors and symmetric tumor changes) may be too high from the perspective of reader measurement variability alone. Compared with the upper limits of the manual intra-reader 95% reference ranges observed in this study (1D: 25.8%; 2D: 51.6%; Table 3), the cut-off values used for reliably detecting disease progression in RECIST (i.e., 20%) and especially in the WHO criteria (i.e., 25%), may be set low. This could result in an early call of disease progression and the premature dismissal of a new drug. However, compared with the lower limit of the 95% reference range (1D: -15.5%; 2D: -27.1%; Table 3), the cut-off values used in both RECIST (-30%) and WHO (-50%) for identifying tumor response may be set too conservatively. This could result in a lack of recognition of tumor responses, especially when assessing newer-generation targeted therapies, where tumor shrinkages may not be as easily detectable due to lower magnitudes of change even in successfully treated tumors.

We used RECIST version 1.0 to select the target lesions (up to 10 lesions per patient) and found the decreased coefficients of variation for all three measurements for all conditions (manual, CAM, and different slice intervals) at the patient level compared with those at the nodule level. Across the three slice intervals, the manual intra-reader 95% reference ranges were tightened and CAM intra-reader 95% reference ranges were further narrowed for the VOL measurements. This suggests that measuring more than one lesion per patient would help reduce measurement variability and that lesion level findings can therefore be considered as representing the maximum variability and minimum agreement.

Since this study began before the publication of RECIST version 1.1 in 2009, RECIST version 1.0 was used. However, only 4 out of 30 patients had more than five (5) lesions—one had 6, two had 8, and one had 10. We would thus expect that the variability at the patient level obtained from this study should be similar to that if lesions had been selected using RECIST 1.1.

There are several limitations in our study. First, we analyzed a relatively small number of patients, 30 in total; there were, however, multiple target lesions per patient resulting in a moderate number of actual lesions analyzed. Additionally, both the number of patients analyzed and the actual number of target lesions were the same as an average phase I or phase II clinical trial. We also utilized the services of only two radiologists. We acknowledge that a greater number of radiologists could add different information to the analysis, but again we attempted to simulate both clinical practice where a follow-up exam may be interpreted either by the same or a different radiologist and clinical trial practice where frequently two radiologists are asked to analyze imaging data [23]. The recommendation of the appropriate CT slice interval was made based on the levels of the measurement variability rather than measurement accuracy. This was because the true lesion sizes (uni-, bi-dimensional and volume) were unknown. The assumption that a thinner CT slice interval correlates with a more accurate volume measurement was based on previous phantom studies [24]. Furthermore, all volume measurements were obtained using our in-house computer segmentation algorithms. Although the results were supervised by radiologists to ensure measurement accuracy, the results reported in this work warrant further validation by using different segmentation algorithms and more radiologists.

6. Conclusions

We assessed the intra- and inter-reader variability in 1D, 2D, and VOL measurements of solid tumors in the lungs, liver and lymph nodes made both manually and by CAM on CT scans reconstructed at three slice intervals of 5, 2.5 and 1.25 mm that cover a potential slice interval range of today's and future clinical trials and clinical practice. Our findings showed that, compared to 5 mm and 1.25 mm slice intervals, 2.5 mm may provide the best viewing and the most robust computation conditions for measuring tumor size, especially volume, in the setting of therapy response assessment. Our findings also showed that repeated intra-reader manual 1D and 2D, and CAM VOL lesion measurements must decrease >16%, 27%, 23% or increase >26%, 52%, 34%, respectively, before the change can be considered statistically significant (5% level). Although real tumor changes must exceed the magnitudes of measurement variability, response cut-off thresholds for the three measurement techniques need to be jointly determined by correlating tumor response with patient outcomes including survival and/or tumor mutation status data [9].

Acknowledgments

Role of the funding source

The study sponsors provided funding for the research; they did not have any role in study design, in the collection, analysis, and interpretation of data, in the writing of the report, or in the decision to submit the paper for publication.

This work was in part supported by Grant R01CA125143 from the National Cancer Institute (NCI) and a research grant from AstraZeneca. The content is solely the responsibility of the authors and does not necessarily represent the funding sources. The authors would like to thank Caroline Shepherd for her help in editing the manuscript.

References

1. WHO handbook for reporting results of cancer treatment. Geneva (Switzerland): World Health Organization; 1979. Offset Publication No. 48
2. Miller AB, Hoogstraten B, Staquet M, et al. Reporting results of cancer treatment. *Cancer*. 1981; 47(1):207–14. [PubMed: 7459811]
3. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada Journal of the National Cancer Institute. 2000; 92(3):205–16.
4. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1. 1). *European Journal of Cancer*. 2009; 45(2):228–47. [PubMed: 19097774]
5. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976; 38(1):388–94. [PubMed: 947531]
6. Lavin PT, Flowerdew G. Studies in variation associated with the measurement of solid tumors. *Cancer*. 1980; 46(5):1286–90. [PubMed: 7214310]
7. Zhao B, James LP, Moskowitz C, et al. Evaluating variability in tumor measurements from same-day repeat CT scans in patients with non-small cell lung cancer. *Radiology*. 2009; 252(1):263–72. [PubMed: 19561260]
8. Oxnard GR, Zhao B, Sima CS, et al. Variability of lung tumor measurements on repeat computed tomography (CT) scans taken within 15 minutes: implications for care and clinical research. *Journal of Clinical Oncology*. 2011; 29(23):3114–9. [PubMed: 21730273]
9. Zhao B, Oxnard GR, Moskowitz CS, et al. A pilot study of volume measurement as a method of tumor response evaluation to aid biomarker development. *Clinical Cancer Research*. 2010; 16(18):4647–53. [PubMed: 20534736]
10. Altorki N, Lane ME, Bauer T, et al. Phase II proof-of-concept study of pazopanib monotherapy in treatment-naïve patients with stage I/II resectable non-small-cell lung cancer. *Journal of Clinical Oncology*. 2010; 28(19):3131–7. [PubMed: 20516450]
11. Yan J, Zhao B, Wang L, et al. Marker-controlled watershed for lymphoma segmentation in sequential CT images. *Medical Physics*. 2006; 33(7):2452–60. [PubMed: 16898448]
12. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine*. 1998; 17(11):1261–91. [PubMed: 9670414]
13. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets. *Translational Oncology*. 2009; 2(4):231–5. [PubMed: 19956383]
14. He X, Oyadiji SO. Application of coefficient of variation in reliability-based mechanical design and manufacture. *Journal of Materials Processing Technology*. 2001; 119(1–3):374–8.
15. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 1(8476):307–10. [PubMed: 2868172]
16. Wormanns D, Kohl G, Klotz E, et al. Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. *European Radiology*. 2004; 14:86–92. [PubMed: 14615902]
17. Goodman LR, Gulsun M, Washington L, et al. Inherent variability of CT lung nodule measurements in vivo using semiautomated. *American Journal of Roentgenology*. 2006; 186:989–94.

18. Petrou M, Quint LE, Nan B, et al. Pulmonary nodule volumetric measurement variability as a function of CT slice thickness and nodule morphology. *American Journal of Roentgenology*. 2007; 188:306–12. [PubMed: 17242235]
19. Wang Y, de Bock GH, van Klaveren RJ, et al. Volumetric measurement of pulmonary nodules at low-dose chest CT: effect of reconstruction setting on measurement variability. *European Radiology*. 2010; 20:1180–7. [PubMed: 19921204]
20. Punnen S, Haider MA, Lockwood G, et al. Variability in size measurement of renal masses smaller than 4 cm on computerized tomography. *Journal of Urology*. 2006; 176(6):2386–90. [PubMed: 17085106]
21. Hopper KD, Kasales CJ, Van Slyke MA, et al. Analysis of interobserver and intraobserver variability in CT tumor measurements. *American Journal of Roentgenology*. 1996; 167(4):851–4. [PubMed: 8819370]
22. 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. *Journal of Clinical Oncology*. 2003; 21(13):2574–82. [PubMed: 12829678]
23. Dodd LE, Kom EL, Freidlin B, et al. Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense? *Journal of Clinical Oncology*. 2008; 26(22):3791–6. [PubMed: 18669467]
24. Prionas ND, Ray S, Boone JM. Volume assessment accuracy in computed tomography: a phantom study. *Journal of Applied Clinical Medical Physics*. 2010; 11(2):3037–52. [PubMed: 20592693]

Fig. 1.

Effects of slice thickness on the lesion appearance and measurement. From left to right, a liver lesion depicted on 1.25 mm, 2.5 mm and 5 mm slice thickness images (a). Overlays of the greatest diameter (Uni) line and the greatest perpendicular diameter (Perp) line of the lesion that were drawn by two independent radiologists (blinded to the slice interval information) on the three slice thickness images (b and c). Corresponding Uni, Perp and Bi (product of Uni and Perp) measurements of the lesion (d).

Fig. 2.

Bland–Altman plots: manual (a) intra- and (b) inter-reader agreement for 1D and 2D measurements. The solid line shows the mean percentage difference in measurements and the dashed lines show the 95% reference range. *Note:* measurements within the 95% reference range can be considered as intrinsic measurement errors (or variations) that are associated with the given measurement tools and imaging techniques. Therefore, a narrower reference range indicates a lower measurement error/variation.

Fig. 3. Bland–Altman plots: CAM (a) intra- and (b) inter-reader agreement for VOL measurements. The solid line shows the mean percentage difference in measurements and the dashed lines show the 95% reference range. *Abbreviations:* CAM, computer-aided method; VOL, volumetric.

Table 1

Primary and metastatic lesions from imaging scan data.

<i>N</i> = 118 lesions		<u>Size category (1D)</u>		Total
Lesion site	<10 (mm)	10 (mm)		
Liver	7	32		39
Lung	8	31		39
Lymph	4	36		40
Total	19	99		118

Components of variability: manual (a) intra- and (b) inter-reader analyses.

N = 118 lesions		(a) Intra-reader analysis; manual reader 1					(b) Inter-reader analysis; manual reader 2 vs 1				
Measurement	Slice interval (mm)	Coefficient of variation (%)					Coefficient of variation (%)				
		Inter-patient	Intra-patient	Inter-lesion	Intra-reader + residual	Total	Inter-patient	Intra-patient	Inter-lesion	Intra-reader + residual	Total
1D	1.25	21.2		44.6	6.1	50.8	23.5		44.5	9.0	52.4
	2.5	22.5		46.2	6.7	53.0	23.7		45.4	6.8	52.9
	5	23.3		45.8	6.9	53.1	25.6		44.5	8.6	53.5
2D	1.25	42.9		100.7	12.2	119.1	46.0		101.6	18.0	124.2
	2.5	45.5		103.7	12.3	124.3	47.3		103.3	14.1	125.7
	5	44.8		103.7	12.1	123.7	49.1		102.1	16.5	126.7

Table 3

Bland–Altman analyses: manual (a) intra- and (b) inter-reader agreement.

<i>N</i> = 118 lesions		(a) Intra-reader analysis; manual reader 1		(b) Inter-reader agreement; manual reader 2 vs 1	
Measurement	Slice interval (mm)	% Difference		% Difference	
		Mean difference	95% reference range	Mean difference	95% reference range
1D	1.25	0.0	(−15.5, 18.3)	6.1	(−17.4, 36.3)
	2.5	2.9	(−14.5, 23.8)	4.7	(−13.4, 26.4)
	5	3.9	(−14.2, 25.8)	2.4	(−19.4, 30.1)
2D	1.25	2.2	(−27.1, 43.2)	11.5	(−31.9, 82.8)
	2.5	7.4	(−23.5, 50.8)	6.7	(−27.6, 57.3)
	5	8.5	(−22.4, 51.6)	3.4	(−34.4, 62.9)

Table 4

Components of variability: CAM (a) intra- and (b) inter-reader analyses.

<i>N</i> = 118 lesions		(a) Intra-reader analysis; CAM reader 1				(b) Inter-reader analysis; CAM reader 2 vs 1			
Measurement	Slice interval (mm)	Coefficient of variation (%)				Coefficient of variation (%)			
		Inter-patient	Intra-patient	Inter-lesion	Intra-reader + residual	Total	Inter-patient	Intra-patient	Total
1D	1.25	23.7	47.2	5.4	54.3	54.3	25.0	47.8	6.4
	2.5	23.8	47.2	4.3	54.2	54.2	24.6	48.0	6.2
	5	24.6	48.6	4.5	56.0	56.0	24.5	46.0	9.3
2D	1.25	45.3	106.9	11.3	127.1	127.1	49.0	110.2	16.4
	2.5	46.4	106.4	10.0	127.1	127.1	49.0	108.4	14.5
	5	47.8	111.2	11.0	133.5	133.5	46.4	103.5	18.8
3D	1.25	70.0	192.0	8.2	245.5	245.5	72.8	202.2	15.1
	2.5	68.4	182.8	9.3	233.0	233.0	70.4	192.8	13.5
	5	66.1	172.2	9.0	217.8	217.8	65.6	163.9	18.0

Abbreviations: CAM, computer-aided method; VOL, volumetric.

Table 5

Bland–Altman analyses: CAM (a) intra- and (b) inter-reader agreement.

<i>N</i> = 118 lesions		(a) Intra-reader analysis; CAM reader 1		(b) Inter-reader agreement; CAM reader 2 vs 1	
Measurement	Slice interval (mm)	% Difference		% Difference	
		Mean difference	95% reference range	Mean difference	95% reference range
1D	1.25	0.0	(−13.8, 16.0)	2.8	(−14.0, 22.8)
	2.5	−0.4	(−11.7, 12.3)	2.7	(−13.5, 22.0)
	5	1.3	(−10.5, 14.6)	7.0	(−17.1, 38.2)
2D	1.25	0.5	(−26.5%, 37.5)	0.3	(−36.2, 57.6)
	2.5	−0.1	(−24.2, 31.8)	2.9	(−31.1, 53.5)
	5	1.8	(−24.8, 37.9)	14.9	(−31.5, 92.6)
VOL	1.25	1.5	(−19.0, 27.2)	−3.3	(−36.3, 46.8)
	2.5	0.5	(−22.3, 30.2)	−1.0	(−31.7, 43.6)
	5	4.2	(−18.8, 33.6)	5.2	(−35.9, 72.7)

Abbreviations: CAM, computer-aided method; VOL, volumetric.

Table 6
Components of variability: summary of CAM and manual, intra- and inter-reader analyses.

Measurement	Slice interval (mm)	Coefficient of variation (%)			
		CAM		Manual	
		Intra-reader + residual	Inter-reader + residual	Intra-reader + residual	Inter-reader + residual
1D	1.25	5.4	6.4	6.1	9.0
	2.5	4.3	6.2	6.7	6.8
	5	4.5	9.3	6.9	8.6
2D	1.25	11.3	16.4	12.2	18.0
	2.5	10.0	14.5	12.3	14.1
	5	11.0	18.8	12.1	16.5
VOL	1.25	8.2	15.1	—	—
	2.5	9.3	13.5	—	—
	5	9.0	18.0	—	—

Abbreviations: CAM, computer-aided method; VOL, volumetric.

Bland–Altman analyses: agreement between manual and CAM techniques for 1D and 2D measurements.

Table 7

<i>N</i> = 118 lesions		Manual vs CAM agreement		
Measurement	Slice interval (mm)	Reader	<i>N</i>	% Difference
		Mean difference		
		95% reference range		
1D	1.25	1	118	−2.4 (−19.5, 18.4)
		2	117	0.7 (−17.6, 23.1)
	2.5	1	118	−3.0 (−17.3, 13.8)
		2	118	−1.2 (−18.4, 19.7)
	5	1	118	−1.6 (−17.1, 16.9)
		2	117	−5.7 (−27.2, 22.1)
2D	1.25	1	118	−2.4 (−32.0, 40.1)
		2	117	8.4 (−26.2, 59.0)
	2.5	1	118	−3.3 (−31.1, 35.8)
		2	118	0.3 (−34.0, 52.4)
	5	1	118	−0.5 (−33.2, 48.1)
		2	117	−10.4 (−43.9, 43.2)

Note: two sets of manual measurements (1D and 2D) for two different lymph nodules from different patients (one on 1.25 and one on 5 mm slice interval) were missing for reader 2.
Abbreviation: CAM, computer-aided method.

Table 8

Tumor burden components of variability: summary of patient-level, CAM and manual, intra- and inter-reader analyses.

Measurement	N = 30 patients	Slice interval (mm)	Coefficient of variation (%)			
			CAM		Manual	
			Intra-reader + residual	Inter-reader + residual	Intra-reader + residual	Inter-reader + residual
1D		1.25	2.5	4.4	4.9	6.8
		2.5	2.0	4.7	4.8	4.4
	5	5	2.6	6.6	5.7	7.1
2D		1.25	6.2	8.0	7.1	13.4
		2.5	4.6	8.5	7.8	9.7
	5	5	6.7	11.1	8.7	10.9
VOL		1.25	4.2	7.7	-	-
		2.5	4.2	7.6	-	-
	5	5	4.8	10.3	-	-

Abbreviations: CAM, computer-aided method; VOL, volumetric.

Tumor burden Bland–Altman analyses: summary of patient-level, CAM and manual, intra- and inter-reader agreement.

Table 9

Measurement	N = 30 patients	Slice interval (mm)	% Difference 95% reference range			
			CAM		Manual	
			Intra-reader	Inter-reader	Intra-reader	Inter-reader
1D	1.25		(-6.0, 7.7)	(-8.7, 16.3)	(-14.2, 12.4)	(-13.1, 26.4)
	2.5		(-5.4, 5.5)	(-8.9, 18.2)	(-11.3, 15.5)	(-7.7, 17.8)
	5		(-5.6, 8.9)	(-12.6, 26.2)	(-12.4, 20.4)	(-17.9, 21.9)
2D	1.25		(-13.7, 21.5)	(-18.1, 27.8)	(-17.7, 21.6)	(-22.5, 62.0)
	2.5		(-9.8, 16.1)	(-19.3, 29.3)	(-15.5, 30.3)	(-18.8, 38.7)
	5		(-16.2, 21.5)	(-17.7, 52.3)	(-16.8, 34.7)	(-25.4, 36.6)
VOL	1.25		(-8.3, 15.7)	(-19.1, 23.6)		
	2.5		(-8.3, 15.7)	(-18.1, 24.6)		
	5		(-9.6, 18.0)	(-21.9, 38.3)		

Abbreviations: CAM; computer-aided method; VOL; volumetric.