Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Tumor volume measurement error using computed tomography imaging in a phase II clinical trial in lung cancer

Claudia I. Henschke
David F. Yankelevitz
Rowena Yip
Venice Archer
Gudrun Zahlmann
Karthik Krishnan
Brian Helba
Ricardo Avila

SPIE.

Claudia I. Henschke, David F. Yankelevitz, Rowena Yip, Venice Archer, Gudrun Zahlmann, Karthik Krishnan, Brian Helba, Ricardo Avila, "Tumor volume measurement error using computed tomography imaging in a phase II clinical trial in lung cancer," *J. Med. Imag.* **3**(3), 035505 (2016), doi: 10.1117/1.JMI.3.3.035505.

Tumor volume measurement error using computed tomography imaging in a phase II clinical trial in lung cancer

Claudia I. Henschke,[a,b,*] David F. Yankelevitz,[a] Rowena Yip,[a] Venice Archer,[c] Gudrun Zahlmann,[d] Karthik Krishnan,[e] Brian Helba,[e] and Ricardo Avila[f]
[a]Icahn School of Medicine at Mount Sinai, One Gustave Levy Place, Box 1234, New York, New York 10029, United States
[b]Early Diagnosis and Treatment Research Foundation, PO Box 1609, New York, New York 10021-0044, United States
[c]Roche Products Limited (Pharmaceuticals), Hexagon Place, 6 Falcon Way, Shire Park, Welwyn Garden City AL7 1TW, United Kingdom
[d]F.Hoffmann-La Roche, Grenzacherstrasse 124, Basel 4070, Switzerland
[e]Kitware Inc., 28 Corporate Drive, Clifton Park, New York 12065, United States
[f]Accumetra LLC, 7 Corporate Drive, Clifton Park, New York 12065, United States

Abstract. To address the error introduced by computed tomography (CT) scanners when assessing volume and unidimensional measurement of solid tumors, we scanned a precision manufactured pocket phantom simultaneously with patients enrolled in a lung cancer clinical trial. Dedicated software quantified bias and random error in the *X*, *Y*, and *Z* dimensions of a Teflon sphere and also quantified response evaluation criteria in solid tumors and volume measurements using both constant and adaptive thresholding. We found that underestimation bias was essentially the same for *X*, *Y*, and *Z* dimensions using constant thresholding and had similar values for adaptive thresholding. The random error of these length measurements as measured by the standard deviation and coefficient of variation was 0.10 mm (0.65), 0.11 mm (0.71), and 0.59 mm (3.75) for constant thresholding and 0.08 mm (0.51), 0.09 mm (0.56), and 0.58 mm (3.68) for adaptive thresholding, respectively. For random error, however, *Z* lengths had at least a fivefold higher standard deviation and coefficient of variation than *X* and *Y*. Observed *Z*-dimension error was especially high for some 8 and 16 slice CT models. Error in CT image formation, in particular, for models with low numbers of detector rows, may be large enough to be misinterpreted as representing either treatment response or disease progression. [DOI: 10.1117/1.JMI.3.3.035505]

Keywords: volumetry; measurement error; response evaluation criteria in solid tumors; calibration; computed tomography.

Paper 16007RRR received Jan. 14, 2016; accepted for publication Aug. 23, 2016; published online Sep. 20, 2016.

1 Introduction

Change in tumor size on computed tomography (CT) imaging is commonly used to assess treatment response, both in the context of routine clinical practice as well as in clinical trials. The use of volume instead of length or response evaluation criteria in solid tumors (RECIST) as a measure of tumor size has inherent theoretical advantages, as the former reflects changes in three-dimensions (3-D) while the latter only reflects changes in one-dimension (1-D).[1] Despite this theoretical advantage, volume measures have not been widely accepted. Indeed, the revised RECIST (version 1.1) reports that there is currently insufficient standardization and widespread availability to recommend adoption of alternative assessment methods.[2]

A major challenge in introducing a new standard for assessment of tumor size is to understand the underlying measurement error and to define what constitutes a "meaningful change" in the assessment criteria. The greater the measurement error, the greater the observed change must be to be certain that there is a genuine change. CT scanner performance is typically assessed by periodic imaging of a standard calibration device[3] according to predefined imaging protocols. These protocols, however, do not necessarily reflect the acquisitions needed for clinical assessment of patients or account for the many sources of variation that occur in the context of a large multicenter trial. Even when using the identical scanners and imaging protocols, image production is influenced by the particular patient and the particular lesion.

To address and to personalize the measurement error attributable to the CT scanner for simple measurements of length and CT attenuation, calibration devices or "phantoms" have been developed, which can be simultaneously scanned with the patient.[4,5] A precision phantom, called the "pocket phantom," was designed to assess the fundamental imaging performance of the CT scanner[6] and to quantify the measurement error of CT imaging in a clinical trial.

2 Materials and Methods

2.1 *Clinical Trial*

ABIGAIL was a multicenter, randomized, phase II study to explore the correlation between biomarkers and RECIST assessed response to first-line carboplatin-based chemotherapy in combination with bevacizumab in patients with advanced or recurrent NSCLC (ClinicalTrials.gov identifier: NCT00700180; 6), which randomly assigned 303 patients from 49 clinical sites in 15 countries to one of the two treatment regimens. All patients

*Address for correspondence: Claudia I. Henschke, E-mail: claudia.henschke@mountsinai.org

provided written informed consent, and the study protocol was approved by the Independent Review Boards and/or ethics committees of each site. The results of the primary endpoint (RECIST response rate) have been reported.[6] Tumor volume was an additional exploratory endpoint.

Standardized imaging protocols for the chest and abdomen were developed for the different CT scanners at the investigational sites. Intravenous contrast material was mandated, unless there was a medical contraindication. CT scans were obtained for all patients at baseline (maximum of 14 days prior to treatment) and every 12 weeks (the end of every second treatment cycle) until disease progression was documented. The imaging protocol specification allowed data from a single CT data acquisition to be used to reconstruct images at 5.0 mm (thick-sections) for RECIST assessment as well as images at 2.0 mm or less (thin-sections) for volume assessment. Each participating site was able to choose its own protocol within certain set limits of protocol-set parameters of pitch, reconstruction kernel, tube rotation, tube potential, and field of view. Pitch was selected to allow for scanning of the entire chest in a single breath at 2.0 mm slice thickness or less. A nonedge-enhancing reconstruction kernel was acquired and different kernels were also allowed for optional additional reconstructions. Tube rotation was set to 1 s or less. Tube potential was in the range of 120 to 140 kVp, with the tube current being adjusted to either fixed or automatic dose modulation. Field of view was adjusted according to patient size. All sites followed standard CT calibration procedures. The anonymized imaging data were sent to the Early Diagnosis and Treatment Research Foundation where imaging data were stored and standard analyses were performed.

2.2 *Pocket Phantom*

Twenty-one pocket phantoms (see Appendix) were distributed to participating sites and placed on the sternum of patients while they were undergoing the CT scan. The Teflon sphere embedded in the phantom had a specified diameter of 15.875 mm (maximum tolerance $\pm$ 0.05 mm), which corresponds to a specified volume of 2094.79 mm^3 (maximum tolerance of $-$ 19.73 to 19.86 mm^3).

A total of 77 patients had at least one eligible scan, i.e., a CT scan with a slice thickness of 2.0 mm or less, slice spacing not greater than the slice thickness, and the phantom fully included in the field of view. Fully automated software detected and measured the Teflon sphere in the CT scans for each patient: a to c) 1-D maximum orthogonal length (henceforth referred to as simply "length"), separately in the

a. X,
b. Y,
c. and Z imaging dimensions,
d. the maximum RECIST measure,
e. 3-D volume measure.

The RECIST measure is the longest diameter in the X and Y planes in a single CT image. Volumes were obtained based on the segmented boundary of the Teflon sphere using a constant threshold, which uses the midpoint of the expected CT Hounsfield Unit (HU) value between the Teflon sphere and the surrounding urethane material as the threshold value for boundary segmentation. In addition, adaptive thresholding segmentations were also obtained for which estimates of the foreground Teflon HU density and the background urethane HU density were calculated from homogeneous regions within each pocket phantom. Figure 1 shows calculated sphere boundaries and spatial measurements on axial, sagittal, and coronal slices of the phantom when scanned with a patient at different time points using constant thresholding. The systematic error (bias), bias percent error, random error, and coefficient of variation were calculated for each of the five measurements (see Appendix). The Pearson correlation coefficient between the Z length and volume measurements was calculated.

2.3 *Analysis*

All CT scans obtained on the 77 patients were used for statistical analysis and are referred to as the "study" dataset. The slice thickness distributions for the study dataset are provided in Table 1.

A second set of CT scans, designated as the "longitudinal" dataset, was created from the study dataset so that the variability of the scans of the same patient could be tracked over time. For each patient, a set of longitudinal scans was identified, where all scans were acquired with the same reconstruction kernel, slice thickness, and slice spacing as the first scan. Of the 77 patients, 43 patients had more than one scan that met the aforementioned criteria, 17 patients had a set of two such scans, 15 had three, 7 had four, 3 had five, and 1 had six. Thus, the analysis of the longitudinal scan dataset was performed on a total of 128 scans.

To examine the impact of the different CT scanners on measurement consistency, the same analyses described for the study dataset were performed on the 43 longitudinal patient scan series for the three manufacturers, designated as A, B, and C, and for scanner models that were generationally categorized by their number of detector rows: either 16 or fewer, or more than 16. Therefore, six different combinations of manufacturers and model generations were analyzed.

To graphically illustrate the systematic and random errors, we provide the quartile values (the nonparametric equivalents of the mean and standard deviation) for the x, y, and z measurements using adaptive thresholding. The second quartile value is the median and the third minus the first quartile value is a measure of the precision of the measurement, as is the standard deviation (Fig. 2).

3 Results

Table 2 summarizes the data observed for the study dataset. The sample mean (and bias percent error) for X, Y, and Z length measurements in the study dataset were 15.65 mm ($-$1.41%), 15.56 mm ($-$1.99%), and 15.66 mm ($-$1.36%) for constant thresholding and 15.84 mm ($-$0.25%), 15.72 mm ($-$0.95%), and 15.88 mm (0.02) for adaptive thresholding, respectively. The random error of the X, Y, and Z length measurements is given by the standard deviations (and CV). These were 0.10 mm (0.65), 0.11 mm (0.71), and 0.59 mm (3.75) for constant thresholding and 0.08 mm (0.51), 0.09 mm (0.56), and 0.58 mm (3.68) for adaptive thresholding, respectively. The Z length computed using adaptive thresholding had a sixfold (0.58/0.09) higher standard deviation and sixfold (3.68/0.56) higher CV than that of X and Y while the same Z length computed using constant thresholding had a fivefold (0.59/0.11) higher standard deviation and a fivefold (3.75/0.71) higher CV than that of X and Y. This is illustrated by the quartile plots shown in Fig. 2.

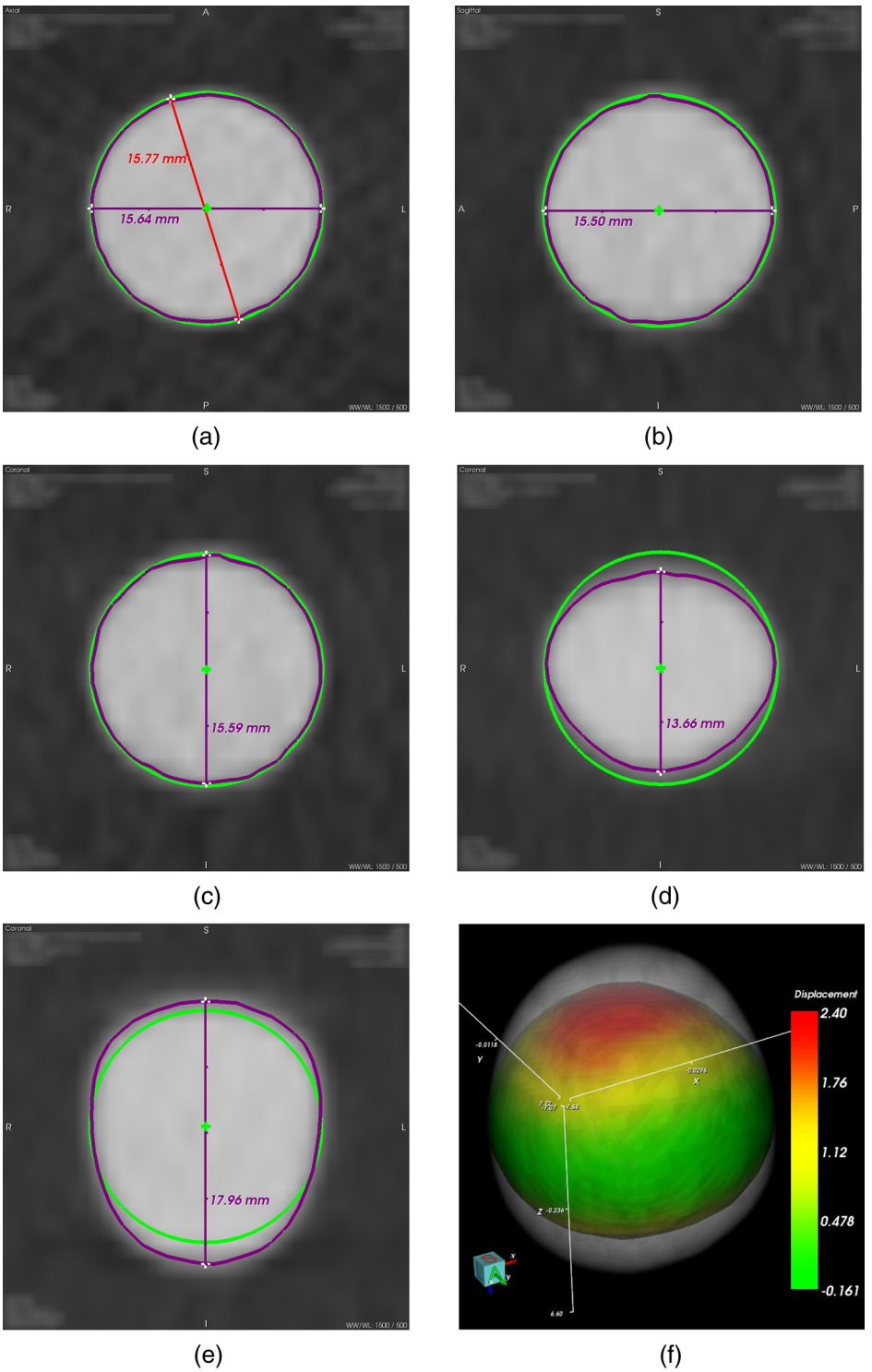

Fig. 1 The top row shows (a) axial, (b) sagittal, and (c) coronal images of a Teflon sphere within a CT pocket phantom deployed in the ABIGAIL study. The bottom row demonstrates the amount of variation over two time points observed in the patient longitudinal dataset with the greatest volumetric change in size. (d, e) Coronal images show measurements of the Teflon sphere from the two time points with the largest observed change in volume. For images (a)–(e), estimated position of the ideal sphere boundary is shown in green, the constant threshold sphere boundary is shown in purple, and the RECIST measurement is shown in red. For image (f), the inner sphere shows the surface of the earlier and smaller volume time point (d), which has been color coded to illustrate the minimum radial distance displacement needed to reach the boundary of the next and larger volume time point (e). The outer gray object shows the extent of the outer surface of the Teflon sphere in the scan with the larger volume time point (e).

Table 1 Slice thickness distribution for the study data set of 77 patients with 162 CT scans.

Slice thickness (mm)	*n*	(%)
0.625	18	(11.1%)
0.8	4	(2.5%)
1	46	(28.4%)
1.25	28	(17.3%)
1.5	54	(33.3%)
2	12	(7.4%)

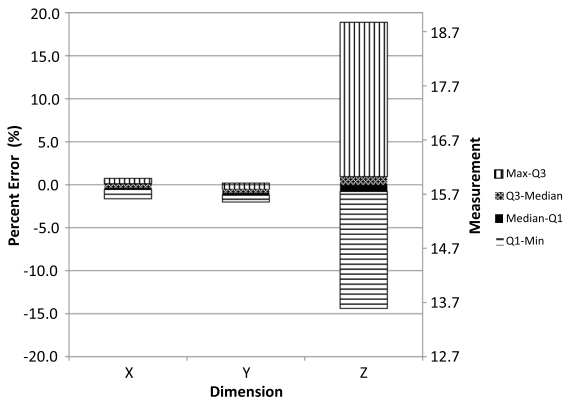

Fig. 2 Teflon sphere diameter measurements for *X*-, *Y*-, and *Z*-dimensions for 77 patients showing the entire range and quartiles of values for each dimension. Percent error (left axis) with respect to manufactured sphere specification and actual measurement values (right axis) are both shown. The measurement error for the *Z*-dimension is about six times higher than that of the *X*- and *Y*-dimensions.

The mean of the RECIST measurements using constant thresholding was 15.59 mm with a negative bias percent error of −1.81%, similar to the percent error for the *X*-, *Y*-, and *Z*-dimensions. The standard deviation of the constant thresholding RECIST measurement was 0.09, and the CV was 0.60, also in line with the *X*- and *Y*-dimension coefficients but much less than that of *Z*-dimension. RECIST measurements using adaptive thresholding showed lower levels of bias percent error and CV. For the volume measurement, the corresponding values were 1978.80 mm^3 (−5.54%) and 97.80 mm^3 (4.94) for constant thresholding and 2049.19 mm^3 (−2.18%) and 96.85 mm^3 (4.73) for adaptive thresholding. The bias percent error of the volume measurement was also negative and larger than that of the RECIST for both constant and adaptive thresholding because it reflects the bias in all three-dimensions (*X*, *Y*, and *Z*). The standard deviation of the volume measurement was also much larger for the same reason; thus, the CV was useful for comparison purposes. The CVs of the *X*-, *Y*-, and *Z*-dimension measurements of length with respect to the RECIST constant thresholding measure were 1.08, 1.17, and 6.21 times, respectively, while the volume measurement was 8.18 times that of the RECIST measure. The 9.46 CV ratio value relative to RECIST for adaptive thresholding was similar to constant thresholding. Observing that the degree of random error in the *Z*-dimension length and volume measurements as measured by the coefficient of variation are similar, the correlation coefficient was calculated and found to be 0.95 for both constant and adaptive thresholding.

The *Z*-dimension measurements for the six combinations of manufacturers and number of detector rows showed more variability in the *Z*-dimension than the *X* and *Y*. In particular, Manufacturer A scanners with 16 or less detector rows had greater variability in the *Z*-dimension than all the other five combinations of manufacturer and detector row scanners (Fig. 3). The maximum change from scan to scan for an individual patient was observed when using scanner manufacturer A with a 16 detector row or less. Figure 4 compares the results when Manufacturer A scanners with a 16 or less detector rows were compared to Manufacturer A with more than 16 detector rows using adaptive thresholding. The maximum change from one time point to the next for volume was 734.9 mm^3 (1718.16 to 2453.04), a 42.77% increase as shown in Fig. 4. At the same points, the *Z* unidimensional measurement changed slightly less, 4.24 mm (13.94 to 18.18), a 30.42% increase.

Table 3 shows the status of scanner performance with the removal of data from 16 or less detector rows from scanner manufacturer A. Although systematic error remains nearly identical with the exclusion of this data, CVs are reduced by more than a factor of 2 whether using constant or adaptive thresholding. Despite this reduction, variability of measurements that involve the *Z*-dimension remains higher than those involving only the *X* and *Y*-dimensions.

4 Discussion

The goal of this paper was to address the measurement error resulting from CT image production. To isolate the error, a calibration device was developed and scanned with the patients undergoing CT scans in a clinical trial. The results clearly illustrate that the measurement error is lower in the axial plane (*X*- and *Y*-dimensions) as compared with that of the *Z*-dimension for all the scanners used in the study. These differences are also reflected in the unidimensional RECIST and the 3-D volume measures. The analysis also demonstrates the differences among different CT makes and types of scanners as defined by the number of detector-rows. Note that measurement of the Teflon sphere in the pocket phantom represents a best case scenario where contrast and object simplicity result in much lower levels of measurement error than those presented by a tumor in a particular patient.

Both the systematic errors as measured by the bias and the random error as measured by the standard deviation were studied. There was an underestimation bias for all measurements except the *Z* length measurement under adaptive thresholding. Under ideal sampling and noise conditions, the combined interaction of the imaging resolution of a typical CT scanner and a sphere will produce a negative bias.[7] The bias was not critical for assessing change between two measurements, although it is in determining the actual volume at any one time point. Critical for assessment of change is the random error, which is measured by the standard deviation or quartile values as this determines the capability of assessing a genuine change over time.

The high correlation of 95% between the random error in the *Z*-dimension and in the volume measurements can be better understood using a simplified model. The volume for an ideal ellipsoid, *V*, is equal to $[\frac{4}{3} \cdot \pi \cdot r_x \cdot r_y \cdot r_z]$ where $r =$ radius of

Table 2 Summary of measurements of the *X*-, *Y*-, and *Z*-dimension maximum orthogonal lengths, RECIST measure, and volume for the Teflon sphere (specified diameter 15.88 mm, volume of 2094.79 mm^3) using constant and adaptive threshold segmentation for the study data set of 77 patients with 162 CT scans.

				Systematic error		Random error		
	Measure	Mean	Std. error of the mean	Mean bias	Bias % error	Standard deviation	CV (×100)	CV(×100) relative to RECIST
Constant	*X* length	15.65	0.01	−0.22	−1.41	0.10	0.65	1.08
	Y length	15.56	0.01	−0.32	−1.99	0.11	0.71	1.17
	Z length	15.66	0.05	−0.22	−1.36	0.59	3.75	6.21
	RECIST	15.59	0.01	−0.29	−1.81	0.09	0.60	1.00
	Volume	1978.80	7.68	−116.00	−5.54	97.80	4.94	8.18
Adaptive	*X* length	15.84	0.01	−0.04	−0.25	0.08	0.51	1.02
	Y length	15.72	0.01	−0.15	−0.95	0.09	0.56	1.13
	Z length	15.88	0.05	0.00	0.02	0.58	3.68	7.36
	RECIST	15.75	0.01	−0.13	−0.79	0.08	0.50	1.00
	Volume	2049.19	7.61	−45.61	−2.18	96.85	4.73	9.46

the ellipsoid. If there were little variability in the *x*- and *y*-dimensions, r_x and r_y, would be virtually unchanged, and the variability for volume measure would be directly proportional to the variability in the radius along the *z*-dimension, r_z. Our calibration phantom showed that there was minimal error in *X* and *Y* and so the volume error was proportional to the variability in *Z*, although slightly larger due to the minimal additional contributions from the *X* and *Y* dimensions.

The effect of the variability in the *Z*-dimension becomes more evident when considering the longitudinal dataset. Here again there is minimal random error in the *X*- and *Y*-dimension lengths, but the random error in the *Z*-dimension is much larger and thus, also in the volume measure. Similarly, the measurement of the change in volume over time reflects the error in the *Z*-dimension (Figs. 2 and 3). When comparing the three manufacturers, the *X*- and *Y*-dimension length measures are quite similar; however, there were substantial differences in measurements in the *Z*-dimension and thus, also of the volume. This difference is most apparent for scanners with 16 or fewer detector-rows. In particular, the 8- and 16-slice scanner family of one manufacturer consistently had the largest measurement error for volume change assessment, being as large as 43%.

To better understand the cause of variability in the *Z*-dimension associated with the 16 slicer scanner model with the highest

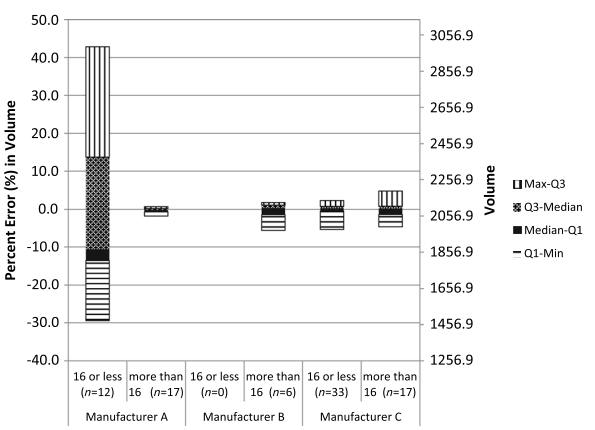

Fig. 3 For the 43 patients with more than one scan, measurements made on CT scans at the subsequent visits were compared to their measurements at baseline and the percentage change in measurements was calculated. The entire range and quartiles of percentage change in volume measurements relative to the baseline value for these 43 patients (i.e., 85 subsequent visits) by the six combinations of manufacturers and detector row scanners. The largest measurement error for volume change was 43% from manufacturer A's 16 or less detector row scanner.

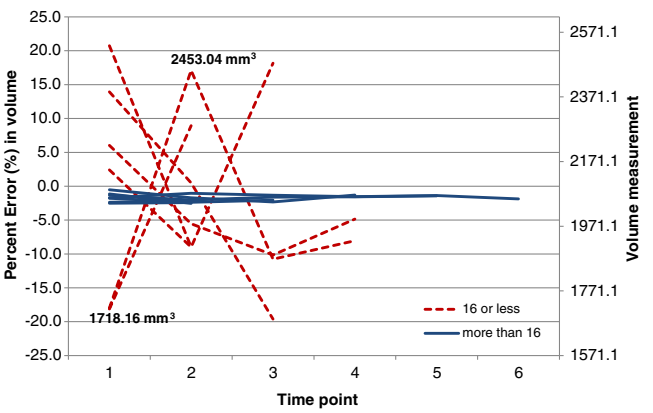

Fig. 4 Longitudinal change in volume measurement in patients whose CT scans were performed using manufacturer A with more than 16 detector row scanners (dashed red line) and with 16 or less detector row scanners (solid blue line). The largest measurement error for volume change was seen in one case where the volume of the Teflon sphere measured 1718 mm^3 at baseline and 2453 mm^3 on the next visit, a 43% increase in volume.

Table 3 Summary of measurements of the *X*-, *Y*-, and *Z*-dimension maximum orthogonal lengths, RECIST measure, and volume for the Teflon sphere (specified diameter 15.88 mm, volume of 2094.79 mm^3) using constant and adaptive threshold segmentation for the study data set after removal of the problematic scans from manufacturer A scanners, resulting in 70 patients with 143 CT scans.

				Systematic error		Random error		
	Measure	Mean	Std. error of the mean	Mean bias	Bias % error	Standard deviation	CV (×100)	CV (×100) relative to RECIST
Constant	*X* length	15.65	0.01	−0.22	−1.40	0.11	0.68	1.13
	Y length	15.56	0.01	−0.31	−1.96	0.11	0.72	1.19
	Z length	15.66	0.02	−0.21	−1.33	0.23	1.50	2.48
	RECIST	15.59	0.01	−0.28	−1.78	0.09	0.60	1.00
	Volume	1979.36	3.44	−115.44	−5.51	41.19	2.08	3.45
Adaptive	*X* length	15.84	0.01	−0.04	−0.22	0.08	0.52	1.09
	Y length	15.73	0.01	−0.14	−0.91	0.09	0.55	1.14
	Z length	15.88	0.02	0.00	0.03	0.22	1.36	2.82
	RECIST	15.75	0.01	−0.12	−0.76	0.08	0.48	1.00
	Volume	2049.65	2.41	−45.15	−2.16	28.83	1.41	2.92

volume variability, we identified the same scanner model at an institution in the US and scanned three second-generation CT pocket phantoms with the identical clinical trial CT scanning protocol. These three pocket phantoms are very similar to the ones used in the clinical trial in that they contain the same size Teflon spheres surrounded by Urethane; however, they differ in that they have a smaller form factor. The coronal image of these phantoms in Fig. 5 shows similar spatial variation along the *Z*-dimension as observed at multiple sites with the same scanner in the clinical trial. The slice thickness and spacing for this scan is 1.25 mm. The spatial warping in this coronal image appears to be periodic allowing for both positive and negative displacements depending upon position along the *Z*-dimension. Given that no patient is present in the scan, this periodic spatial warping along the *Z*-dimension cannot be attributed to any patient factors including patient motion.

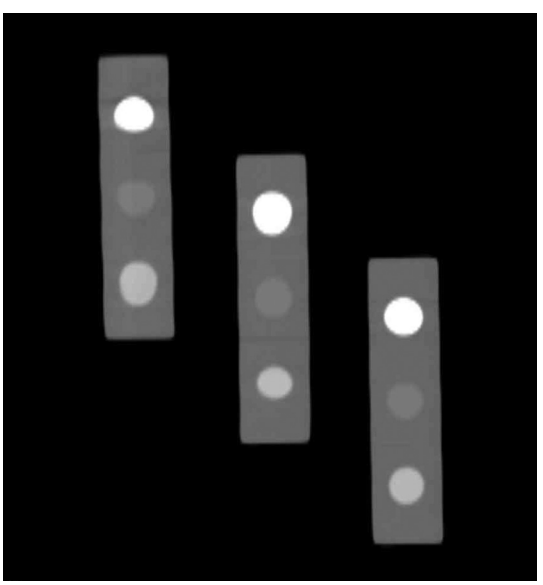

Fig. 5 A CT scan performed in the United States of three second-generation CT pocket phantoms using the same model CT scanner observed to have the highest variability in the clinical trial. Slice thickness and spacing for this scan was 1.25 mm. No patient is present in this scan and spatial warping is evident along the *Z* direction (top to bottom), particularly for the phantom that is furthest from scanner iso-center (left side).

The implications of our findings are far reaching. They demonstrate that CT imaging results in precise measurements of the RECIST measure as this measure utilizes only the *X*- and *Y*-dimensions. However, for any measurement that utilizes the *Z*-dimension, the standard deviation is typically more variable, mainly due to the CT scanner itself.

Currently, there is no standard definition of "meaningful change" (i.e., disease response or progression) for a volume measure, although the Quantitative Imaging Biomarkers Alliance has previously suggested that an empirical figure of 30% would be reasonable for any nodule greater than one centimeter.[8] The NELSON investigators have recommended the use of 25% or larger as representing a true volume increase.[9] In that study, all but one site used a particular model of CT scanner made by the same manufacturer. In the context of CT screening, where volume changes of even smaller nodules are followed, it is prudent to assume that the measurement error for these small nodules would be even larger. Our results show that simply on the basis of measurement error introduced by the CT scanner, the volume change may be large enough to be considered as meaningful change according to the above criteria when in fact there is no change at all. The implications of basing treatment decisions on such measurements are obvious and profound.

While unidimensional RECIST continues to be the standard measure for tumor response, the new RECIST 1.1 criteria allow for the *Z* axis to be used for this unidimensional measurement.

Here again, our results suggest caution, as we found that large variation in measurements attributable to the Z axis for some scanners, as high as 30%.

We have focused solely on the measurement error introduced by the CT acquisition device itself. This is certainly an underestimate of the overall error that will occur when measuring tumor size in a clinical trial. Additional errors can be introduced by measuring lower contrast objects and using software with a more sophisticated algorithm to define the border of the tumor, particularly when the border is ill-defined and complex with various vascular and other attachments.[10] Petrick et al., e.g., have observed advantages of volumetric measurement over RECIST measurement when considering more complex nodule shapes.[11]

All of this implies that, when moving from RECIST 1.0 to RECIST 1.1 and volume-based measures, deeper understanding of the various contributions to the overall measurement error need to be understood. Scanner performance is quite variable and it may be that certain scanners cannot be used for measuring volumetric change. This study also demonstrates how calibration devices can be used to monitor a trial, potentially in real time, so that the scanners can be checked and image error corrected for a particular patient being scanned with a given protocol.

This study has several limitations including that the data are from a single study using a single calibration device. It is also acknowledged that there were a relatively limited number of cases and investigational sites from which we obtained longitudinal data compared to the total number of sites and patients enrolled in the clinical trial as a whole. However, we believe our results convincingly demonstrate differences in measurements performed in the X- and Y-dimensions compared with those involving the Z-dimension. We also were able to demonstrate that one particular family of CT scanners produced a consistently larger error in the Z-dimension. As this was demonstrated at five different sites, and reproduced at a sixth site under tight constraints so as to eliminate any source of confusion in terms of determining that the cause of the spatial warping came from the scanner itself, it suggests that when considering using volume measurements that CT scanners should undergo a qualification process. In addition, the linear measurement results shown in Table 3 suggest that large differences in image variability performance along the Z-dimension can remain for low number of detector row CT scanners even after avoidance of scanner models with the highest levels of Z variability.

In conclusion, we found that volumetric measures were subject to measurement errors introduced by the production of the CT images. In some instances, these measurement errors were sufficiently large to be considered as meaningful change in volume when in fact there was no change at all. While the largest errors were limited to a certain class of scanners in our study, the full extent of how different scanners perform given all of the interacting scanning parameters, including those that are inherent to the scanner as well as those that can be varied by the user remains unknown, and implies a need for improved scanner calibration, including consideration of the necessity of calibrating on an individual scan basis. In addition, the implications from our findings extend well beyond tumor response assessment and into a vast range of medical applications that already require accurate spatial (and likely attenuation) measurements such as prosthetic implants, emphysema,[12] and coronary artery calcifications[13] where the need for measurement accuracy may even be greater.

Appendix: Assessment of Systematic and Random Error of Images Obtained on CT Scanners Using the Pocket Phantom

A major challenge in introducing a new standard for assessment of tumor size is to understand the underlying measurement error and to define what constitutes a "meaningful change" in the assessment criteria. The greater the measurement error, the greater the observed change must be to be certain that there is a genuine change.

Errors in measuring tumor volume on CT scans can be separated into two categories: errors due to the production of the images and errors due to the definition of the tumor boundary. Regarding image production, errors are influenced by the particular make and model of the CT scanner (e.g., the geometry of the detector arrays and performance of the scintillators) and the particular choice of imaging parameters (e.g., slice thickness, field of view, dose, and pitch). With regard to tumor boundary definition, errors are influenced by the characteristics of the tumor and patient (e.g., lesion complexity, tumor–nontumor interface, lesion location, and patient size) and the software algorithm used to define the tumor characteristics.

To address the measurement error for a given person undergoing imaging using a CT scanner, a precision manufactured phantom, called the CT pocket phantom, was designed (RA) and manufactured by The Phantom Laboratory (Salem, New York). The "pocket phantom" can be used as a reference standard to quantify error due to CT scanner image production; an essential step in formulating meaningful assessment criteria.

The phantom consisted of three precision-manufactured spheres made from Teflon, Delrin, and Acrylic materials,

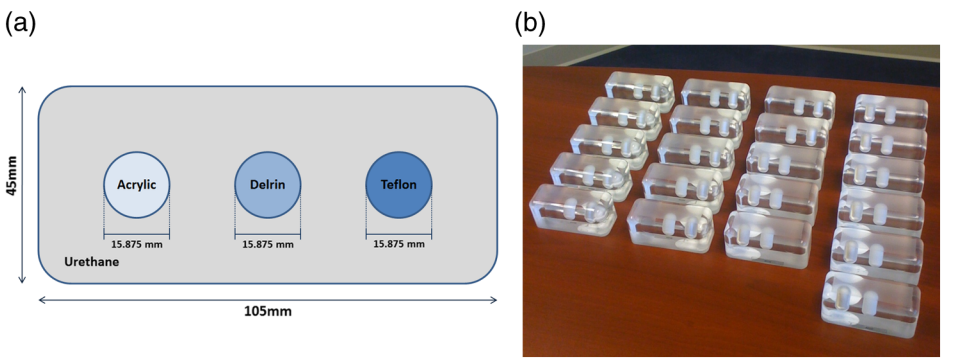

Fig. 6 (a) CT pocket phantom design and (b) 21 manufactured CT pocket phantoms.

embedded in a 45 mm × 45 mm × 105 mm urethane block (Fig. 6). The analysis is limited to the Teflon sphere in this report because the attenuation difference between the Teflon sphere and the urethane is closest to the difference between a solid lung nodule and surrounding lung parenchyma, and the Teflon sphere has the maximal attenuation difference between it and the urethane in which it is embedded and therefore, allows for more precise differentiation of surface boundary of the sphere.

The systematic error (bias) is the difference between the sampled mean and true known mean. The true diameter of the Teflon sphere was the specified diameter, 15.88 mm. The systematic error (bias) can also be expressed as percent error, i.e., the systematic error (bias) divided by the true mean. A negative value represents an underestimation of the true mean (a positive value an overestimation).

The random error for a measure is provided by the calculated sample standard deviation.

To compare the degree of random error of the length measurement (in mm) to those of the volume measurements (in mm^3), the dimensionless coefficient of variation (CV) is provided.

Acknowledgments

Funding Sources: Roche, Kitware Inc., Accumetra, LLC, Early Diagnosis and Treatment Research Foundation.

References

1. D. F. Yankelevitz et al., "Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation," *Radiology* **217**, 251–256 (2000).
2. E. A. Eisenhauer et al., "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)," *Eur. J. Cancer* **45**, 228–247 (2009).
3. C. H. McCollough et al., "The phantom portion of the American College of Radiology (ACR) computed tomography (CT) accreditation program: practical tips, artifact examples, and pitfalls to avoid," *Med. Phys.* **31**, 2423–2442 (2004).
4. Z. H. Levine et al., "A low-cost fiducial reference phantom for computed tomography," *J. Res. Nat. Inst. Stand. Technol.* **113**, 335–340 (2008).
5. Z. H. Levine et al., "A low-cost density reference phantom for computed tomography," *Med. Phys.* **36**, 286–288 (2009).
6. T. Mok et al., "A correlative biomarker analysis of the combination of bevacizumab and carboplatin-based chemotherapy for advanced non-squamous non-small-cell lung cancer: results of the phase II randomized ABIGAIL study (BO21015)," *J. Thorac. Oncol.* **9**(6), 848–855 (2014).
7. P. Mendonça et al., "Bias in the localization of curved edges," in *European Conf. on Computer Vision (ECCV)*, 2004.
8. CT Volumetry Technical Committee, "CT tumor volume change profile, quantitative imaging biomarkers alliance. Version 2.2. Reviewed draft," QIBA, 2012, http://www.rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA-CT%20Vol-TumorVolumeChangeProfile_v2.2_ReviewedDraft_08AUG2012.pdf, (9 September 2016).
9. R. J. Van Klaveren et al., "Management of lung nodules detected by volume CT scanning," *N. Engl. J. Med.* **361**, 2221–2229 (2009).
10. K. Krishnan et al., "An open-source toolkit for the volumetric measurement of CT lung lesions," *Opt. Express* **18**(14), 15256–15266 (2010).
11. N. Petrick et al., "Comparison of 1D, 2D, and 3D nodule sizing methods by radiologists for spherical and complex nodules on thoracic CT phantom images," *Acad. Radiol.* **21**(1), 30–40 (2014).
12. B. Keller et al., "Quantitative assessment of emphysema from whole lung CT scans: comparison with visual grading," *Proc. SPIE* **7260**, 726008 (2009).
13. J. Shemesh et al., "Ordinal scoring of coronary artery calcifications on low-dose CT scans of the chest is predictive of death from cardiovascular disease," *Radiology* **257**, 541–548 (2010).

Claudia I. Henschke, PhD, MD, is a thoracic radiologist at the Icahn School of Medicine. She is the director of the Early Lung and Cardiac Program. She has been the principal investigator for the International Early Lung Cancer Action Program (I-ELCAP), since its inception. She also holds a doctorate degree in mathematical statistics. She has over 300 peer-reviewed publications.

David F. Yankelevitz, MD, is a thoracic radiologist. His main areas of interest are lung cancer screening and the evaluation of pulmonary nodules. He has been the coprincipal investigator for the I-ELCAP, since its inception. He has over 200 peer-reviewed publications.

Rowena Yip, MPH, is a researcher with academic training in biology and public health with emphases in biostatistics and epidemiology. She has been working for the International Early Lung Cancer and Cardiac Action Program for over 10 years, and her research interests include statistical and epidemiologic methodology as applied to critical diagnostic and therapeutic topics involved in lung cancer screening.

Venice Archer, MD, is an oncologist in late stage clinical development at Roche Products Limited UK.

Gudrun Zahlmann is a biomedical engineer and computer scientist working in early drug development at F-Hoffmann-La Roche Ltd. Switzerland.

Karthik Krishnan, MS, obtained his bachelor's in electrical engineering from Birla Institute of Technology and Science, Pilani, India, in 2002 and a master's in electrical and computer engineering from the University of Arizona in 2004. His interests lie in medical image analysis, visualization, and biomedical computing on GPUs. He has over 20 peer-reviewed publications and contributes to widely used open source libraries, such as the visualization toolkit and the insight segmentation and registration toolkit.

Brian Helba has worked for the past 5 years as an R&D engineer at Kitware, Inc., developing systems for the management of biomedical image data. He is particularly interested in promoting the principals of open and reproducible science to the curation, sharing, and analysis of data. He is currently leading the development of several large publicly accessible image archives, which fuse clinically relevant radiology, digital microscopy, and surface imaging data with expert annotations and state-of-the-art algorithmic analysis.

Ricardo Avila is the CEO of Accumetra, LLC. He has extensive experience developing imaging detection and measurement algorithms with an emphasis on early lung cancer applications. Throughout his over 20-year career, he has contributed over 20 publications and supported a wide range of open science projects, including VTK, ITK, Give a Scan, and the Open Source Electronic Health Record Alliance. He holds a MS degree in computer science from SUNY Stony Brook, specializing in 3-D biomedical imaging and visualization.