Comparison of 1D, 2D and 3D nodule sizing methods by radiologists for spherical and complex nodules on thoracic CT phantom images

Original Research

Nicholas Petrick[a], Hyun J. Grace Kim[b], David Clunie[c], Kristin Borradaile[c], Robert Ford[d], Rongping Zeng[a], Marios A. Gavrielides[a], Michael F. McNitt-Gray[b], Z. Q. John Lu[e], Charles Fenimore[e], Binsheng Zhao[f] Andrew J. Buckler[g]

[a]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD

[b]David Geffen School of Medicine at UCLA, Los Angeles, CA

[c]CoreLab Partners, Inc, Princeton, NJ

[d]Princeton Radiology, Princeton, NJ

[e]National Institute of Standards and Technology, Gaithersburg, MD

[f]Department of Radiology, Columbia University Medical Center, New York, NY

[g]Buckler Biomedical LLC, Wenham, MA, USA

Comparison of 1D, 2D and 3D nodule sizing methods by radiologists for spherical and complex nodules on thoracic CT phantom images

Original Research

Advances in Knowledge:

1. 3D volume-based sizing under radiologist control provided almost unbiased estimates of true size (maximum bias -3.6%) while uni-dimensional (maximum -39.8%) and bi-dimensional (maximum -63.9%) could have substantial bias, especially for complex nodules.
2. 3D volume-based sizing had larger relative variability (10.7% to 29.5%) compared with uni-dimensional (2.6% to 20.2%) and bi-dimensional (6.1% to 21.5%), however, the variabilities among techniques were comparable in thin slice CT imaging.
3. Uni-dimensional and bi-dimensional in-plane size measurements were sensitive to changes in nodule orientation while 3D volume-based sizing was not.

Implications for Patient Care:

Patients need access to efficacious treatments based on accurate and reliable lesion assessment techniques. This study characterizes and compares the technical performance of various lesion sizing techniques including volumetry. If volumetry is shown to be an improvement over linear diameter measurements, it may speed the development of new treatments and improve patient management.

Summary Statement:

Our data and analysis show that radiologist 3D volume estimates have the potential to provide low bias/low variance estimates of lesion size, especially with thin slice imaging, and that each sizing method has its own unique bias/variance tradeoff.

ABSTRACT

Purpose: To estimate the bias and variance of radiologists measuring the size of spherical and complex synthetic nodules.

Methods: This study did not require IRB approval. Six radiologists estimated the size of 10 synthetic nodules embedded within an anthropomorphic thorax phantom from CT scans at 0.8 and 5 mm slice thicknesses. The readers measured the nodule size using three sizing techniques (1D longest in-slice dimension; 2D area from longest in-slice and longest perpendicular dimension; 3D semi-automated volume). Inter-comparisons of bias (difference between average and true size) and variance among methods were performed.

Results: The relative biases of radiologists with the 3D tool were -3.6%, -0.5%, -1.6%, 1.9%, -2.6% for 10 mm spherical, 20 mm spherical, 10 mm elliptical, 10 mm lobulated and 10 mm spiculated nodules compared with 1.4%, -0.1%, -26.5%, -7.8%, -39.8% for 1D. 3D was significantly less biased than 1D for elliptical and spiculated nodules. The relative standard deviation for 3D was 21.9%, 11.9%, 10.7%, 29.4%, 24.0% compared with 5.7%, 2.6%, 20.2%, 5.3%, 16.3% for 1D. 1D sizing was significantly less variable than 3D for spherical and lobulated nodules and significantly more variable for the ellipsoid. 2D results were similar to 1D. 3D bias and variability were smaller for thin 0.8 mm slice data compared with thick 5.0 mm data.

Conclusion: Radiologists' 3D sizing had low bias across all nodule shapes, while 1D and 2D sizing underestimated the true largest diameter and area of complex nodules. Radiologists were generally less variable in their 1D and 2D size measurements.

INTRODUCTION

Multi-detector CT imaging is a critical clinical tool in lung cancer evaluation. The recently reported results from the National Lung Screening Trial (NLST) of subjects at high risk for lung cancer indicated that low-dose CT screening reduced lung cancer mortality by 20% compared with screening with planar chest radiography (1). These results along with the results from the International Early Lung Cancer Action Program (I-ELCAP) (2) provide strong evidence that CT screening has potential as an effective tool for detecting early, more survivable lung cancers. CT imaging has also had an impact on the staging of lung cancers (3) and is one of the factors that led to the updated guidelines on the TNM Classification of Malignant Tumours stage groupings in 2009 (4, 5). Likewise, CT imaging has become a critically important tool for monitoring lung cancer patients undergoing therapy.

The Response Evaluation Criteria in Solid Tumors (RECIST) is currently the quantitative standard used to assess disease progression in patients with lung cancer in clinical trials (6, 7). Although, it was originally developed for use only in clinical trials, clinicians routinely ask radiologists to provide RECIST measurements as an objective evaluation of patient response in daily practice (8). The measurement standard for tumor sizing used as part of RECIST is the longest, in-plane diameter of a tumor. This measurement standard, while simple to implement, is also problematic for complex cancers because tumors do not generally expand or contract uniformly (6). In an effort to address the limitations of RECIST and to potentially improve the sensitivity of the measurement to true anatomical changes, volumetric sizing has been proposed as an alternative quantitative approach for measuring anatomical changes in a lesion over time.

However, questions have been raised about whether volumetric analysis will add value or only increase the costs of patient care and the complexity of running clinical trials (6).

A number of studies have tried to compare and contrast 1D sizing approaches with 2D and 3D volumetric sizing methods using both clinical images and phantom data (9-14). Zhao *et al*. evaluated the measurement reproducibility of *in vivo* non–small cell lung cancer tumors on same-day repeat CT scans (11). They showed that the limits of agreement for computer-aided 1D, 2D and volumetric sizing measurements between the two repeat scans were (-7.3%, 6.2%), (-17.6%, 19.8%), and (-12.1%, 13.4%), respectively. The study was limited to only assessing the precision (not the accuracy) in tumor sizing because of its use of clinical images. Chen *et al*. recently reported on the impact of image acquisition and reconstruction parameters on the bias and variability of 3D volume measurements for simple spherical lesions embedded within an anthropomorphic chest phantom for a single reader (12).

The Radiological Society of North America (RSNA) created the Quantitative Imaging Biomarker Alliance (QIBA) at its annual meeting in 2007 to investigate the role of quantitative imaging methods as potential biomarkers in evaluating disease and responses to treatment (15). Specifically, the QIBA Volumetric CT Technical Subcommittee is investigating the technical feasibility of volumetric image acquisition and analysis as a biomarker for treatment response. One of that group's efforts is reported on in this manuscript. The goal of our study was to estimate the bias and variance of radiologists measuring the size of spherical and complex (non-spherical) synthetic nodules. Preliminary results from this study are described in (16).

MATERIALS AND METHODS

Institutional Review Board review was not required to perform this research since it involved only the collection and analysis of anthropomorphic phantom image data. No human subjects were involved or at risk.

Database Description: The data set consisted of CT image data of an anthropomorphic thorax phantom containing synthetic nodules of varying shapes, densities and sizes. The anthropomorphic thorax phantom and vascular insert (("Lungman" N1 Phantom, Kyotokagaku, Kyoto, Japan) are shown in Figure 1. The phantom and insert were both designed to mimic the x-ray attenuation properties of the complex structures within the thorax and lung region (17). A total of 10 nodules (five different types [shape/size combinations] at two different CT densities) were attached to the vasculature within the thorax phantom and imaged. Figure 2 shows the four nodule shapes that were selected to be representative of both simple (spherical) and more complex clinical lesions (elliptical, lobulated and spiculated). Table 1 summarizes the characteristics of the nodules used in our reader study. The phantom was imaged on a Philips Mx8000 IDT 16-row scanner (Philips Healthcare, Andover, MA) with scan acquisition parameters summarized in Table 2. A total of 40 CT datasets (10 nodules x 2 slice thicknesses x 2 repeat exposures) were evaluated by the reviewing radiologists during the reader study.

Reading Protocol: Six radiologists familiar with evaluating lesion response in drug trials participated as readers in this study. They measured the size of all nodules using three different measurement techniques in each of two reading sessions. Each reading session was separated by at least 3 weeks. The sizing methods were: (1) a manual 1D uni-dimensional measurement using electronic calipers measuring the largest in-slice diameter for the lesion; (2) a manual bi-dimensional measurement using electronic calipers with the reader first performing a uni-

dimensional measurement and then measuring the largest perpendicular diameter within the same slice; and (3) a 3D volumetric measurement using a semi-automated 3D volumetric tool.

The 1D size measure was based on the RECIST sizing standard (7), but it excluded summing over multiple nodules and did not incorporate recommendations for assessing change. The in-plane and perpendicular linear measurements were multiplied to provide an area-based size estimate for the 2D measurement. This measure is based on the World Health Organization (WHO) criteria (18) but again limited to the single lesion measurement standard. The manual 1D and 2D size measurements were made using Medstudio v.4.6 (Megasoft Limited, South Riding, VA). The 3D volumetric measurements were made using a prototype proprietary semi-automated tool (Oncocare Prototype, Siemens Corporate Research, Princeton, NJ) (19) which included a lesion segmentation component (19, 20). The 3D measurement process was as follows: (1) the reader defined a seed stroke across the lesion (i.e., a RECIST-like line across the perceived maximum diameter of the lesion), (2) applied the segmentation tools, (3) evaluated the quality of the segmentation, and (4) refined or added seeds strokes and reapplied the segmentation tool until satisfied with the 3D nodule segmentation. The software then provided the estimate of nodule volume, which was recorded.

All reading sessions took place at a central facility (CoreLab Partners Inc., Princeton, NJ) using proprietary software and consumer color LCD monitors calibrated to the Digital Imaging and Communications in Medicine (DICOM) Grayscale Standard Display Function. The size measurements were made using a fixed lung window/level display setting of 1200HU (window) and -600HU (level). All measurement techniques were independently applied (i.e., the manual 2D technique involved a separate estimate of the longest diameter instead of relying on the 1D

estimate) and readers, cases, and measurement techniques were randomized between sessions to reduce potential biases.

Reference Standard: The longest diameter, largest perpendicular dimension and volume were measured on the physical nodules and used as the reference standard. The longest diameter was measured using calipers multiple times with the average of these measurements used as the 1D reference standard for the nodule. A similar approach was used to measure the perpendicular dimension as well with the average longest diameter multiplied by the average largest perpendicular dimension used as the 2D area reference standard. The 1D or 2D reference standards were measured without regard to any CT reconstructed slice planes so they represent the true longest dimensions of the nodule. The reference standard volume was determined using a mass-density approach. The density of the nodule was measured by the manufacturer. The mass of each nodule was the average of multiple weight measurements made using a precision scale. The reference standard volume was then estimated using the formula:

$$\boldsymbol{Volume_{Nod}} = \frac{\boldsymbol{Mass_{Nod}}}{\boldsymbol{Density_{Nod}}}, \qquad (1)$$

Statistical Methods: The primary objective was to compare the bias (difference between the average and true lesion size) and variance among the three sizing methods. Secondary analysis comparing the sizing methods for the subset of thin 0.8 mm and thick 5.0 mm slice thickness data was also performed. In order to facilitate the comparisons, a normalized nodule size error for each measurement technique was calculated, thus producing a unitless size error estimate for each technique. Some type of normalization was necessary because each of the sizing methods produced estimates in different dimensional spaces (i.e., 1D: mm, 2D: mm^2, 3D: mm^3). The

percent size errors were used to calculate and compare the relative bias and relative standard deviation among the three sizing techniques. Percent size error, relative bias and relative standard deviation for sizing method i are defined as follows:

$$PE^{i}_{Size} = \frac{Size^{i}_{Est} - Size^{i}_{True}}{Size^{i}_{True}} \times 100\%, \quad (2)$$

$$Bias^{i}_{Rel}\left(PE^{i}_{Size}\right) = E[PE^{i}_{Size}] - 0, \text{ and} \quad (3)$$

$$Std^{i}_{Rel}\left(PE^{i}_{Size}\right) = std\left(PE^{i}_{Size}\right). \quad (4)$$

where $Size^{i}_{Est}$ and $Size^{i}_{True}$ are the estimated and reference standard sizes, respectively, for sizing method i ($i = 1, 2, 3$ corresponds to the 1D, 2D or 3D sizing methods).

We applied ANOVA and a goodness-of-fit statistic defined by the correlation coefficient R^2 to the various study factors to identify the most important contributing factors and interactions to include in our analysis. We then performed multiple comparisons of the difference in relative bias among the sizing methods as well as comparisons of variability among the methods using the relative standard deviations as the analysis metric. Error bars for the relative biases and all bias comparisons were determined using a t-test comparison applied to the data within each subgroup. This analysis accounted for the possibility of different variances within each subgroup. Error bars on the relative standard deviation comparisons where determined using a bootstrap approach and 2000 bootstrap realizations. A Bonferroni correction (21) for multiple comparisons was included in both the bias and variances analyses.

All analyses were performed using Matlab 7.12.0.635 with Statistical Toolbox 7.5 (R2011a). Numbers are reported as a percentage of the reference standard size with 95% confidence intervals.

RESULTS

ANOVA and Goodness-of-Fit analysis: ANOVA analysis (no interactions) and a Goodness-of-Fit comparison were applied starting with Nodule Type, Nodule Density, Nodule Set, Sizing Method, Readers, Slice Thickness, and Reading Session as factors. Nodule Type refers to the five shape/equivalent diameter combinations from Table 1; Nodule Set refers to the two repeat reconstructed scans evaluated; and Reading Session refers to the two different sessions in which nodule sizes were estimated. The ANOVA analysis identified five significant factors but two of these significant factors, Nodule Density and Slice Thickness, accounted for only a small percentage of the overall error based on the Goodness-of-Fit comparison. Because of their small contributions, these factors were eliminated from the model and were subsequently lumped in with the unexplained error resulting in a reduced pool of factors (Nodule Type, Sizing Method and Readers). The ANOVA (two-way interactions) and Goodness-of-Fit analyses were repeated for these three remaining factors. Table 3 and Figure 3 show the ANOVA table and the Goodness-of-Fit comparison among these individual and interacting factors. It is clear from Figure 3 that terms involving Readers as a factor explained the least amount of error (each factor involving the Reader explaining ≤5% of the total error and combined they explained only about 10% of the error) resulting in Readers also being removed as a factor in our subsequent analyses. This initial analysis identified Nodule Type and Sizing Method as the most important contributing factors to include in our analysis. Therefore, a comparison among the three sizing methods as a function of Nodule Type was performed.

Comparison of Relative Bias among the sizing methods: The relative biases of the readers as a function of Nodule Type and Sizing Method are given in Table 4 along with the corresponding 95% confidence intervals for each estimate. Figure 4 shows the data summarized as boxplots. In this figure, a low relative bias would result in a median error near zero. The results indicate that radiologists using the 3D semi-automated sizing tool had a relative bias of -3.6%, -0.5%, -1.6%, 1.9%, -2.6% for a 10 mm spherical, 20 mm spherical, 10 mm elliptical, 10 mm lobulated and 10 mm spiculated nodules compared with 1.4%, -0.1%, -26.5%, -7.8%, -39.8% for 1D longest in-slice diameter sizing. The 2D relative biases were similar to 1D except for the 10 mm spiculated nodules where 2D had substantially more bias. Figure 5 shows comparisons of relative bias among the sizing methods as a function of Nodule Type. The radiologists measuring nodule volume with the 3D semi-automated tool had less bias for the complex nodules and a statistically significant lower bias compared with either their 1D or 2D sizes for the 20 mm elliptical and 10 mm spiculated nodules. The 1D method had a statistically significant lower bias compared with the 2D method for the 10 mm spiculated nodules. No other comparison reached statistical significance.

Comparison of variability among the sizing methods: The comparison on variability among the methods was performed by comparing relative standard deviations ($\boldsymbol{Std}_{\mathbf{Rel}}$). The relative standard deviations as a function of nodule shape and sizing method are given in Table 5 along with the corresponding 95% confidence intervals for each estimate. Figure 6 provides boxplots for the relative standard deviations where the distributions were generated through bootstrap resampling. In this figure, a smaller median value indicates low variability for the measurement method. The variability, as measured by the relative standard deviation, for the radiologists measuring nodule volume was 21.9%, 11.9%, 10.7%, 29.4%, 24.0% compared with 5.7%, 2.6%,

20.2%, 5.3%, 16.3% for 1D. The variability for the radiologists making 2D bi-directional size estimates was similar to 1D. Figure 7 shows comparisons of relative standard deviation among the sizing methods as a function of Nodule Type. 3D sizing had a significantly higher relative standard deviation compared with 1D sizing for the 10 mm spherical, 20 mm spherical, and 10 mm lobulated nodules. 3D sizing had a significantly higher relative standard deviation compared with 2D for the 10 mm lobulated and 10 mm spiculated nodules, and a significantly lower relative standard deviation compared with either the 1D or 2D sizing methods for the 20 mm ellipsoid nodules. No other comparisons with 3D reached statistical significance.

Comparison of bias and variability between thin and thick slice CT data: As part of our secondary analyses, a comparisons of biases and variances for each nodule type and sizing method combination were performed for the subset of radiologist size measurements made on thin 0.8 mm and thick 5.0 mm slice data. Figure 8 and Figure 9 summarize the comparisons of relative bias and relative standard deviation, respectively. The relative bias data was generally consistent between the thin and thick slice data for 1D and 2D sizing with statistical significance achieved only for 1D sizing of the 10 mm spiculated nodules and for 2D sizing of the 10 mm lobulated and 10 mm spiculated nodules. The trend was different for 3D volume estimates where thin slice volume estimates were consistently less biased compared to thick slice volume estimates, although none of the differences achieved statistical significance. Different trends were observed in the variance analysis of 3D compared with 1D and 2D. In these analyses, the 3D relative standard deviations were smaller for thin slice data compared with thick for all nodule types while 1D and 2D generally had slightly smaller relative standard deviations for thick slice data compared with thin. None of the thin/thick relative standard deviation comparisons achieved statistical significance.

DISCUSSION

The goal of this study was to determine how radiologists' 1D, 2D and 3D lesion size measurements are affected by differences in nodule characteristics and CT acquisition parameters, especially for more complex nodules. Our results show that radiologists using the semi-automated volume sizing tool were able to achieve close to unbiased estimates of lesion size for all nodule types used in this study. Unbiased is defined as the average nodule size being equal to the actual reference standard size. 1D and 2D radiologist size measurements tended to be unbiased for spherical nodules but systematically underestimated true size for the more complex nodule shapes. The relative biases are based on a reference standard for the physical nodules (i.e., not on specific in-slice CT measurements that would account for nodule orientation and slice location) leading to large 1D and 2D biases for the complex nodule shapes. The idea that 3D volume estimates, even under the idealized conditions in our phantom experiments, are close to unbiased is surprising since the measurements are being made on voxelized representations of the object transformed by the CT imaging process instead of the actual structure. Thus, CT imaging may preserve the ability to estimate nodule volume with properly selected acquisition parameters. On the other hand, 1D and 2D in-plane measurements by the radiologists are much more likely to underestimate the nodule's true longest dimensions. This has major implications for both absolute volume as well as change over time measurements.

In general, having unbiased, or at least low bias, quantitative CT measures allows for the measurements to accurately reflect *in vivo* reality on average (e.g., CT image-based size accurately reflects the true lesion size in the patient) assuming the variability is controlled. This is critical for absolute measurements but it is also crucial for change over time measurements when the bias is not constant between scans (i.e., bias uncertainty). Bias uncertainty propagates

as additional variability in the change measurement, thereby reducing the measurement's ability to detect true change. The 1D and 2D in-plane sizing methods suffered from bias uncertainty between scans because they were found to be sensitive to orientation differences and differences in CT hardware and acquisition parameters, especially for the complex nodule shapes.

Bias is only one factor in the quality assessment for a quantitative measure. Variability in the measurement must also be assessed. Our results indicate that the variability is lower for the 1D and 2D sizing methods compared with 3D based on the relative standard deviation comparisons. The only exception was the 20 mm elliptical lesion where variability was higher with the low dimensional measurements. However this anomaly in our data is not attributable to an actual increase in reader measurement variability but instead to differences in orientation between the -10 HU and +100 HU nodules within the CT scans. Figure 10 shows cross-sections of the -10 HU and +100 HU elliptical nodules. It is clear that the in-plane longest diameters and perpendicular diameters are very different because of the orientation difference. Not surprisingly, this difference did not impact the volume estimates because the 3D measurement process generally accounts for orientation variation. This highlights an important limitation of low dimensional 1D and 2D sizing. Changes in lesion orientation, either due to physical changes in the patient or a change in patient alignment within the CT scanner, adversely affect absolute size and change of time measurements. Again, this is not just random variability in the radiologist measurements but is specifically due to bias uncertainty among the imaged nodules. The 1D and 2D variability associated with bias uncertainty may be mitigated to some extent by summing across a set of lesions as recommended in RECIST (7). An observation from our secondary analysis is that bias and variability in the radiologist estimated volumes tended to be smaller with thin 0.8 mm slice data compared with thick 0.5 mm slice data. The analysis found

3D variability to be more consistent across nodules types for thin slice data compared with 1D and 2D variability, as depicted in Figure 9, and had a different trend in that 3D variability fell with thin slices while 1D and 2D variability tended to increase. This suggests that radiologist volume estimates with thin slice imaging may produce a low bias/low variance estimate of nodule size, at least under the ideal conditions evaluated within this study, although this conclusion should be taken as preliminary since only a few relative biases and none of the relative standard deviation thin/thick comparisons achieved statistical significance.

The study did have limitations. One is that the study utilized synthetic nodules with well-circumscribed boundaries imaged within an anthropomorphic phantom. The advantage of this approach is that it allowed for both the bias and variance of the radiologists' measurements to be estimated. Determining the true size of a nodule is very difficult in a clinical patient scan so estimating bias in clinical datasets is extremely challenging. The disadvantage of our approach is that the synthetic nodules and the phantom do not match the true complexity of clinical lesions and lung fields so our results may be more of a lower bound, with clinical performance expected to have higher variability and potentially additional bias. In addition, indistinct boundaries in clinical lesions may be somewhat more problematic for 3D segmentation-based sizing because the entire lesion boundary must to be segmented compared with 1D and 2D where reasonable boundaries on the largest lesion diameter slice may be sufficient to make a measurement. A second limitation is the method selected for comparing the sizing techniques. We chose to compare relative bias and relative standard deviation across sizing methods applied in different dimensional spaces. This may not be completely appropriate. As an example, a 10% uncertainty in a 1D in-plane diameter for a spherical nodule propagates to a 20% uncertainty in cross-sectional area and a 30% uncertainty in volume. We chose this direct comparison because these

metrics would be natural for evaluating an individual sizing tool and for developing standard of use for a specific tool. However, to better understand our comparisons we scaled the 1D size estimates to 3D $\left(Size^{1}_{Scale} = Size^{1}_{Est}{}^{3} \right)$ and compared the scaled 1D biases and variances with the radiologists original 3D volume estimates. The relative biases for the scaled 1D sizes (scaled to 3D) were 5.4%, -0.1%, -50.9%, -20.8%, -72.8% for the 10 mm spherical, 20 mm spherical, 10 mm elliptical, 10 mm lobulated, and 10 mm spiculated nodules. These scaled biases were larger than the unscaled 1D biases and were statistically larger than the radiologists' 3D volume measurements for the ellipsoid, lobulated, and spiculated lesions. The variabilities (relative standard deviations) for the scaled 1D measurements were 17.7%, 7.8%, 33.9%, 13.6%, 23.9% for the same set of lesions. The scaled 1D variabilities were larger than the unscaled 1D variabilities and were more comparable to the 3D variabilities such that only the 10 mm ellipsoid was statistically different (statistically more variable) than the radiologists original 3D measurements. It is not clear to us yet whether scaled or unscaled comparisons are most appropriate for comparing sizing tools applied within different dimension spaces.

Our data and analysis show that radiologist 3D volume estimates have the potential to provide low bias/low variance estimates of lesion size, especially with thin slice imaging, and that each sizing method has its own unique bias/variance tradeoff. This result emphasizes the need to consider both the bias and variance properties of a technique when assessing absolute or change over time in nodule size and when developing clinical standards for evaluating, applying, and interpreting image-based lesion size changes.

ACKNOWLEDGEMENTS

CoreLab Partners Inc. conducted the reader study component of this investigation. They provided the reading facility, review workstations, software, and logistical support. CoreLab Partners radiologists also participated as readers. We would like to acknowledge Corelab Partners for their strong support of this effort as well as the substantial contributions of Lisa M. Kinnard (Medical Research Program, Department of Defense, Fort Detrick, MD) in the design and conduct of the reader study. We would also like to acknowledge CoreLab Partners radiologists Ruth Feldman, MD, Steven Kaplan, MD, George Edeburn, MD, Kevin Byrne, MD, Julie Barudin, MD and Joyce Sherman, MD for participating as readers in this study. Finally, we acknowledge the members of the QIBA Volumetric CT Technical Committee and especially the members of the QIBA Volumetric CT Part 1A subcommittee for making substantial contributions to this work.

The phantom data collection was funded though a Critical Path grant from the U.S. Food and Drug Administration. The intramural research program of the National Institute of Biomedical Imaging and Bioengineering, the National Cancer Institute through IAG no. 224-07-6030, the Center for Interventional Oncology at the National Institutes of Health (NIH) and an Interagency Agreement between the NIH and the United States Food and Drug Administration (FDA) also provided partial support for this project.

The mention of commercial entities, or commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such entities or products by the Department of Health and Human Services.

REFERENCES

1. National Lung Screening Trial Research T, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395-409.
2. International Early Lung Cancer Action Program I, Henschke CI, Yankelevitz DF, et al. Survival of patients with stage I lung cancer detected on CT screening.[see comment]. N Engl J Med. 2006;355(17):1763-71.
3. Buckler AJ, Mulshine JL, Gottlieb R, Zhao B, Mozley PD, Schwartz L. The use of volumetric CT as an imaging biomarker in lung cancer. Acad Radiol. 2010;17(1):100-6.
4. Goldstraw P. The 7th Edition of TNM in Lung Cancer: what now? J Thorac Oncol. 2009;4(6):671-3.
5. Goldstraw P, Crowley J, Chansky K, et al. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours.[Erratum appears in J Thorac Oncol. 2007 Oct;2(10):985]. J Thorac Oncol. 2007;2(8):706-14.
6. Mozley PD, Schwartz LH, Bendtsen C, Zhao B, Petrick N, Buckler AJ. Change in lung tumor volume as a biomarker of treatment response: a critical review of the evidence. Ann Oncol. 2010;21(9):1751-5.
7. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228-47.
8. Suzuki C, Jacobsson H, Hatschek T, et al. Radiologic measurements of tumor response to treatment: practical approaches and limitations. Radiographics. 2008;28(2):329-44.
9. Gavrielides MA, Kinnard LM, Myers KJ, et al. Noncalcified lung nodules: volumetric assessment with thoracic CT. Radiology. 2009;251(1):26-37.
10. Das M, Muhlenbruch G, Heinen S, et al. Performance evaluation of a computer-aided detection algorithm for solid pulmonary nodules in low-dose and standard-dose MDCT chest examinations and its influence on radiologists. Br J Radiol. 2008;81(971):841-7.
11. Zhao B, James LP, Moskowitz CS, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology. 2009;252(1):263-72.
12. Chen B, Barnhart H, Richard S, Colsher J, Amurao M, Samei E. Quantitative CT: technique dependence of volume estimation on pulmonary nodules. Phys Med Biol. 2012;57(5):1335.
13. Das M, Ley-Zaporozhan J, Gietema HA, et al. Accuracy of automated volumetry of pulmonary nodules across different multislice CT scanners. Eur Radiol. 2007;17(8):1979-84.
14. Tran LN, Brown MS, Goldin JG, et al. Comparison of treatment response classifications between unidimensional, bidimensional, and volumetric measurements of metastatic lung lesions on chest computed tomography. Acad Radiol. 2004;11(12):1355-60.
15. Quantitative Imaging Biomarkers Alliance. Radiological Society of North America (RSNA); 2010 [cited 2012 7/12/2012]; Available from: www.rsna.org/research/qiba.cfm.
16. Petrick N, Kim HJG, Clunie D, et al. Evaluation of 1D, 2D and 3D nodule size estimation by radiologists for spherical and non-spherical nodules through CT thoracic phantom imaging. In: Summers RM, van Ginneken B, eds. SPIE Medical Imaging. 1 ed. Lake Buena Vista, Florida, USA: SPIE, 2011; p. 79630D1-7963D7.
17. Gavrielides MA, Kinnard LM, Myers KJ, et al. A resource for the development of methodologies for lung nodule size estimation: database of thoracic CT scans of an anthropomorphic phantom. Optics Express. 2010;18(4):15244-55.

18. World Health Organization.WHO Handbook for Reporting Results of Cancer Treatment. Geneva, Switzerland, 1979.
19. Jolly MP, Grady L. 3D general lesion segmentation in CT. Biomedical Imaging: From Nano to Macro, 2008 ISBI 2008 5th IEEE International Symposium on2008; p. 796-9.
20. Grady L. Random Walks for Image Segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2006;28(11):1768-83.
21. Altman DG. Practical Statistics for Medical Research. Boca Raton, FL: Chapman and Hall/CRC Press, 1991.

Table 1: Technical characteristics of the 10 synthetic nodules (five nodules shape/size combinations at two different x-ray densities) evaluated in the study.

Shape	Equivalent Diameter*	CT Densities
Spherical	10 mm	-10 HU, +100 HU
Spherical	20 mm	-10 HU, +100 HU
Elliptical	20 mm	-10 HU, +100 HU
Lobulated	10 mm	-10 HU, +100 HU
Spiculated	10 mm	-10 HU, +100 HU

*Equivalent diameter is defined as the diameter of a sphere with the same volume as that of the nodule.

• Table 2: Table of scan parameters used in CT data acquisition. Data acquired on a Philips 16-slice Mx8000 IDT scanner (Philips Healthcare, Andover, MA).

Acquisition Parameter	Value
Tube Voltage	120 kVp
Exposure	100 mAs/slice
Pitch	1.2
Reconstructed Slice Thickness*	0.8 mm (0.4 mm interval, 16x0.75 mm collimation) 5.0 mm (2.5 mm interval, 16x1.5 mm collimation)
Reconstruction Kernel	Detail
Repeat Exposures	2 repeat scans of each nodule

*0.8 mm and 5.0 mm reconstructions were acquired at the same radiation dose.

Table 3: ANOVA table (1 and 2-way interactions) limited to Nodule Type, Sizing Method and Readers as fixed factors (ordered by statistical significance).

Source	Sum of Squares	d.f.	Mean Squared Error	F	Prob>F
Nodule Type†	264214.6	4	66053.7	301.6	<0.0001
Nodule Type*Sizing Method†	149435	8	18679.4	85.3	<0.0001
Sizing Method†	80064.5	2	40032.2	182.8	<0.0001
Readers†	48535	5	9707	44.3	<0.0001
Readers*Sizing Method†	33604	10	3360.4	15.3	<0.0001
Nodule Type*Readers†	11108.4	20	555.4	2.5	0.0002
Unexplained Error	304456.5	1390	219		
Total	891417.9	1439			

†Indicates statistical significance at the 0.05 level

Table 4: Relative bias estimates (95% confidence intervals in brackets) as a function of nodule shape and sizing method.

Nodule Type	Sizing Method	Relative Bias (%)*		Standard Error
10 mm Spherical	1D	1.41	[-0.60,3.41]	0.62
	2D	2.61	[-1.74,6.97]	1.35
	3D	-3.58	[-11.20,4.05]	2.36
20 mm Spherical	1D	-0.12	[-1.03,0.79]	0.28
	2D	1.72	[-0.39,3.83]	0.65
	3D	-0.53	[-4.76,3.69]	1.31
20 mm Elliptical	1D	-26.45	[-33.26,-19.63]	2.11
	2D	-23.77	[-30.81,-16.72]	2.18
	3D	-1.61	[-5.32,2.09]	1.15
10 mm Lobulated	1D	-7.79	[-9.60,-5.99]	0.56
	2D	-10.6	[-15.61,-5.60]	1.55
	3D	1.87	[-8.30,2.04]	3.14
10 mm Spiculated	1D	-39.84	[-45.42,-34.26]	1.72
	2D	-63.86	[-67.37,-60.35]	1.08
	3D	-2.55	[-10.76,5.66]	2.54

*95% confidence intervals based on t-distribution within each subgroup and adjusted using a Bonferroni correction for 15 comparisons are shown in brackets.

Table 5: Relative variability as defined by relative standard deviation (95% confidence intervals in brackets) as a function of nodule type and sizing method.

Nodule Type	Sizing Method	Relative Std (%)*	
10 mm Spherical	1D	5.66	[2.53,8.73]
	2D	12.22	[5.36,19.29]
	3D	21.86	[11.05,33.35]
20 mm Spherical	1D	2.59	[1.14,3.96]
	2D	6.05	[3.25,8.45]
	3D	11.85	[4.65,9.47]
20 mm Elliptical	1D	20.20	[16.99,21.50]
	2D	21.05	[17.47,24.20]
	3D	10.71	[5.29,15.68]
10 mm Lobulated	1D	5.29	[3.79,6.78]
	2D	14.58	[9.20,18.96]
	3D	29.40	[16.87,38.27]
10 mm Spiculated	1D	16.33	[8.32,20.15]
	2D	10.23	[6.45,13.28]
	3D	24.04	[13.40,34.43]

*95% confidence intervals based on bootstrap resampling and adjusted using a Bonferroni correction for 15 comparisons are shown in brackets.

FIGURE CAPTIONS

Figure 1: (a) Anthropomorphic thorax phantom and its (b) vascular insert. The synthetic phantom nodules were directly attached to the vascular insert and the vascular insert was placed within the thorax phantom before scanning.

Figure 2: The four basic nodule shapes evaluated by the clinicians in the study included (a) spherical, (b) elliptical. (c) lobulated, and (d) spiculated.

Figure 3: Plot comparing the Goodness-of-Fit as defined by R^2. Note the Readers terms explained the least amount of error (each factor explained$\leq$5% of the total error).

Figure 4: Boxplot of percent size error as a function of sizing method and nodule type.

Figure 5: Comparison plots of relative bias among the sizing methods as a function of nodule shape. The 95% confidence intervals and significance are based on the t-distribution applied within each subgroup and adjusted using a Bonferroni correction for 15 comparisons.

Figure 6: Boxplot of relative standard deviations as a function of sizing method and nodule type.

Figure 7: Comparison plots of relative standard deviation among the sizing methods as a function of nodule shape. The 95% CI's and significance are determined using 2-way

bootstrap resampling (readers by all other factors) and adjusted using a Bonferroni correction for 15 comparisons.

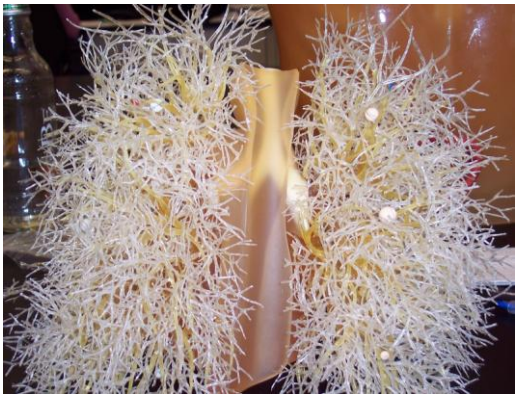
Figure 8: Comparison plots of relative bias between thin 0.8 mm and thick 5.0 mm slice data for each nodule type, sizing method combination. The 95% confidence intervals and significance are based on the t-distribution applied within each subgroup and adjusted using a Bonferroni correction for 15 comparisons.

Figure 9: Comparison plots of relative standard deviation between thin 0.8 mm and thick 5.0 mm slice data for each nodule type, sizing method combination. The 95% CIs and significance are determined using 2-way bootstrap resampling (readers by all others) and adjusted using a Bonferroni correction for 15 comparisons.

Figure 10: Central cross-sections of the (a) -10 HU and (b) +100 HU ellipsoid nodules. The figure depicts the substantial orientation difference between the two ellipsoid nodules. This resulted in very different in-plane longest dimension and longest perpendicular dimension measurements between the scans.

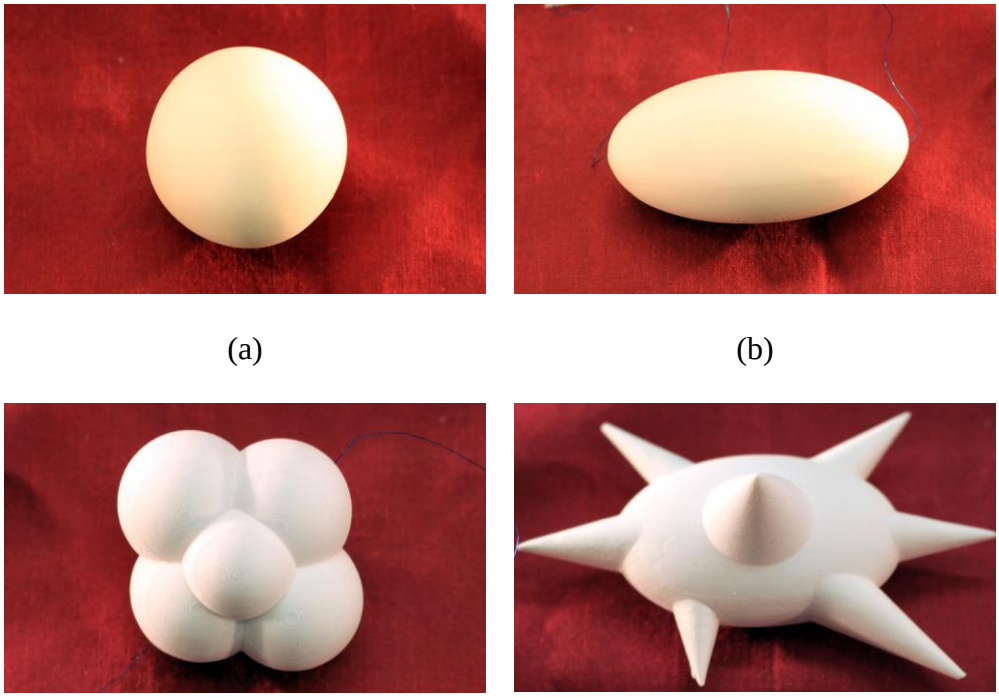
FIGURES

(a)



(b)

Figure 1: (a) Anthropomorphic thorax phantom and its (b) vascular insert. The synthetic phantom nodules were directly attached to the vascular insert and the vascular insert was placed within the thorax phantom before scanning.



(a) (b)

(c) (d)

Figure 2: The four basic nodule shapes evaluated by the clinicians in the study included (a) spherical, (b) elliptical. (c) lobulated, and (d) spiculated.

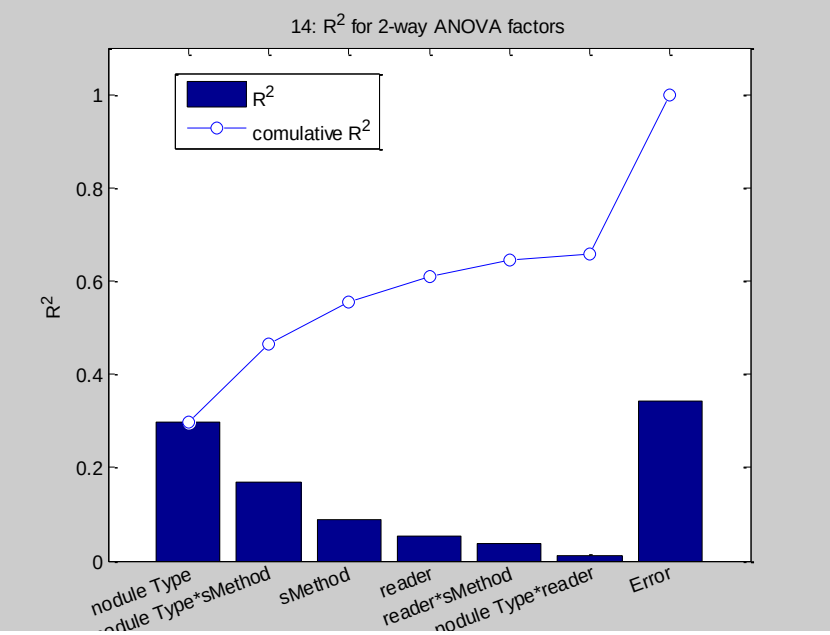
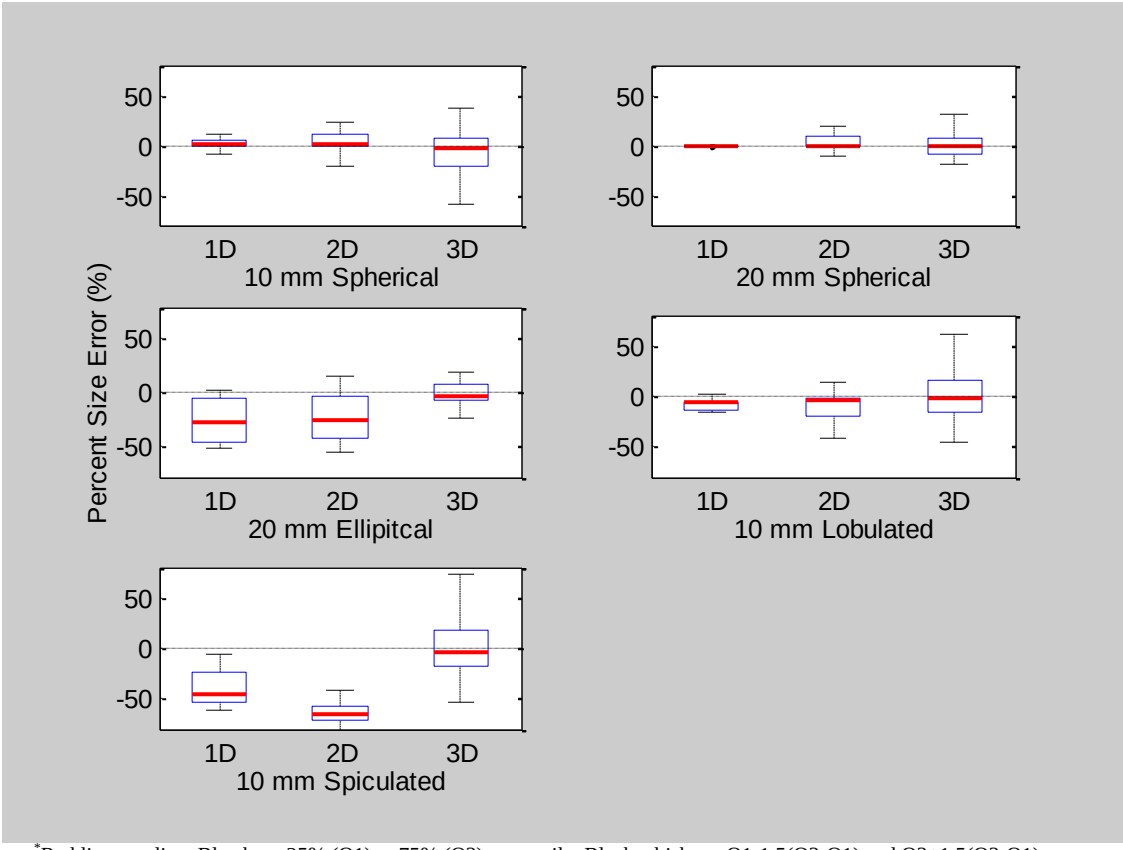
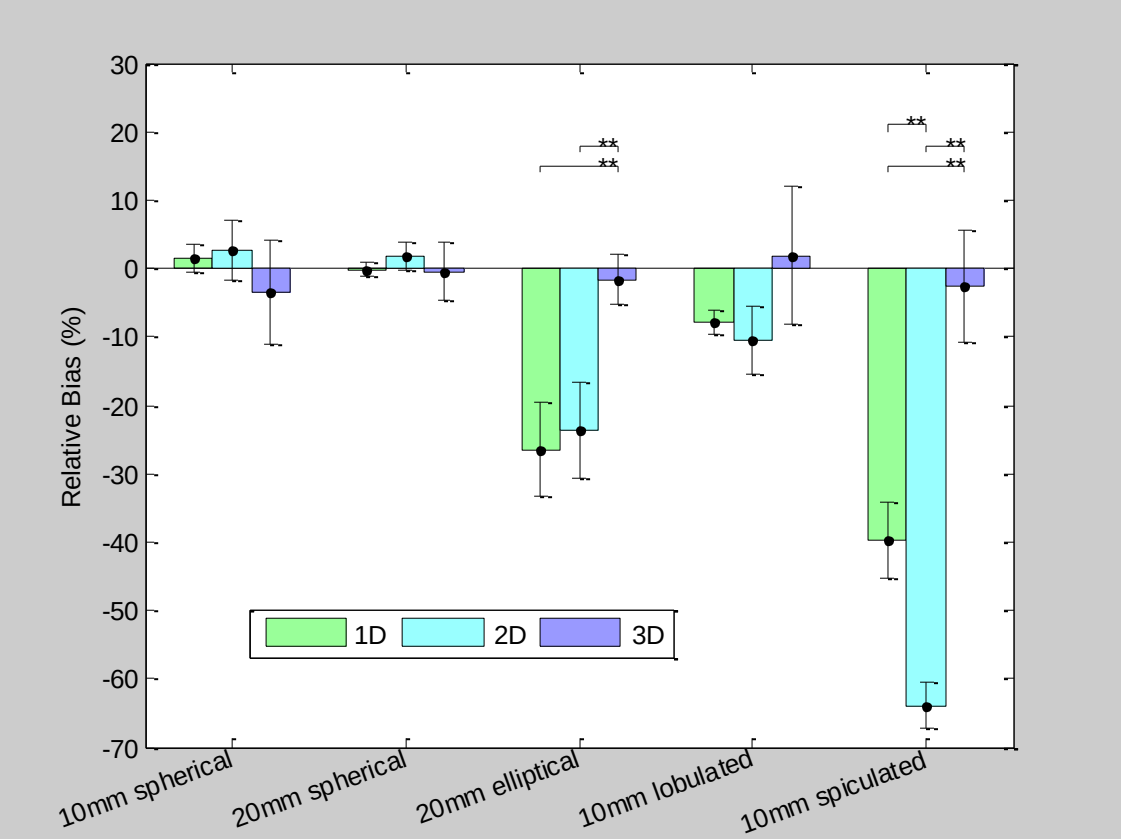


Figure 3: Plot comparing the Goodness-of-Fit as defined by R^2. Note the Readers terms explained the least amount of error (each factor explained≤5% of the total error).



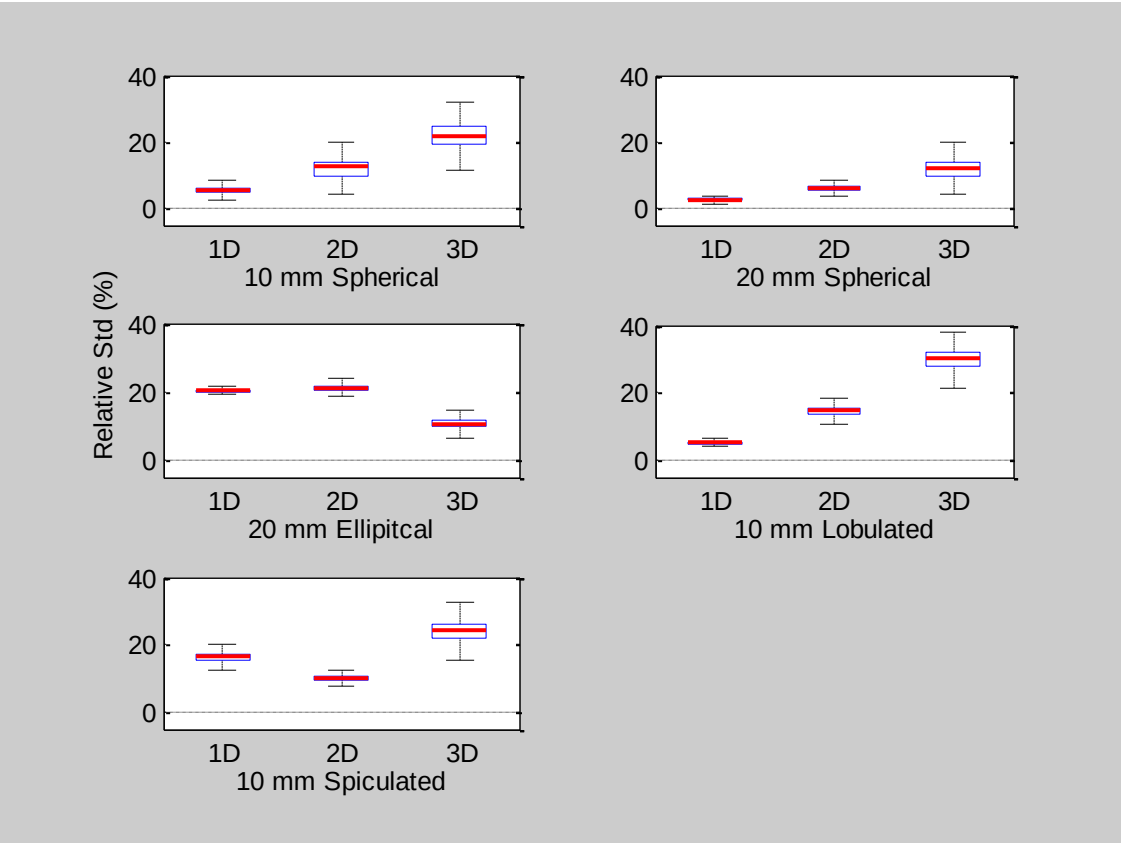
*Red line: median; Blue box: 25% (Q1) to 75% (Q3) percentile; Black whiskers: Q1-1.5(Q3-Q1) and Q3+1.5(Q3-Q1).

Figure 4: Boxplot of percent size error as a function of sizing method and nodule type.



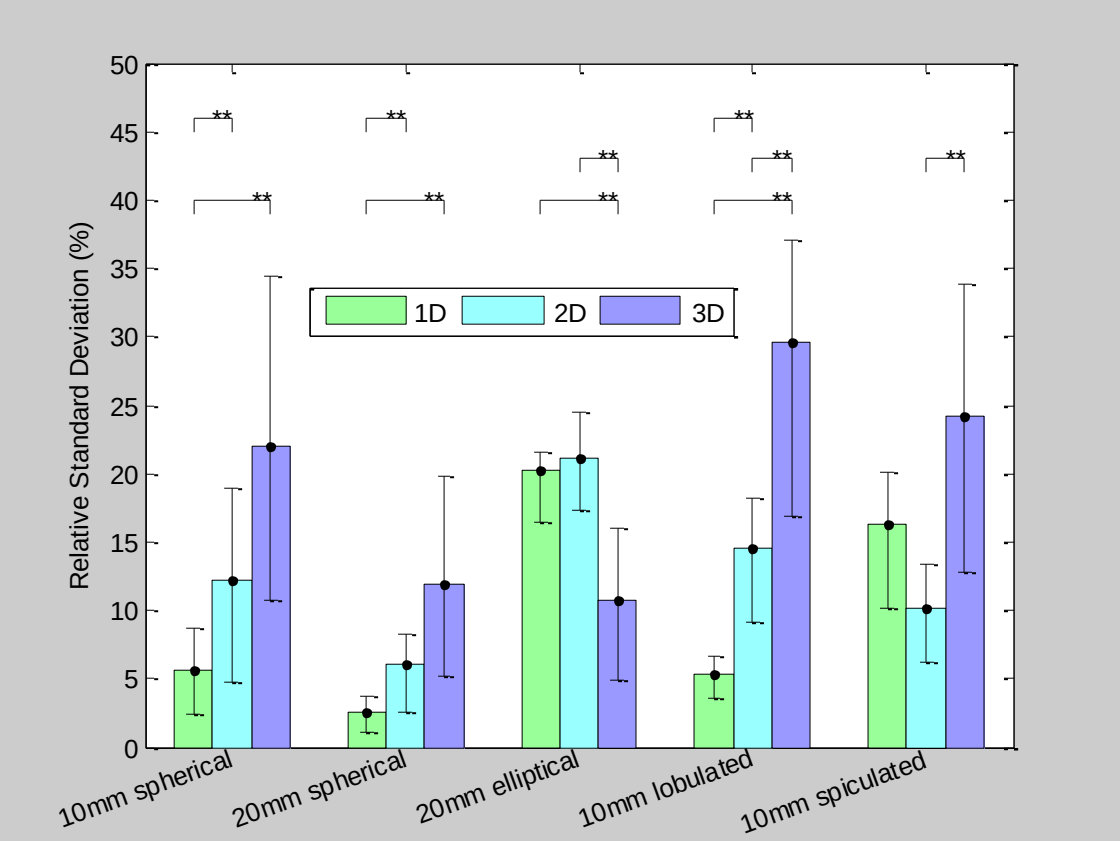
**Indicates a statistically significant difference between the two sizing methods represented by the connecting line.

Figure 5: Comparison plots of relative bias among the sizing methods as a function of nodule shape. The 95% confidence intervals and significance are based on the t-distribution applied within each subgroup and adjusted using a Bonferroni correction for 15 comparisons.



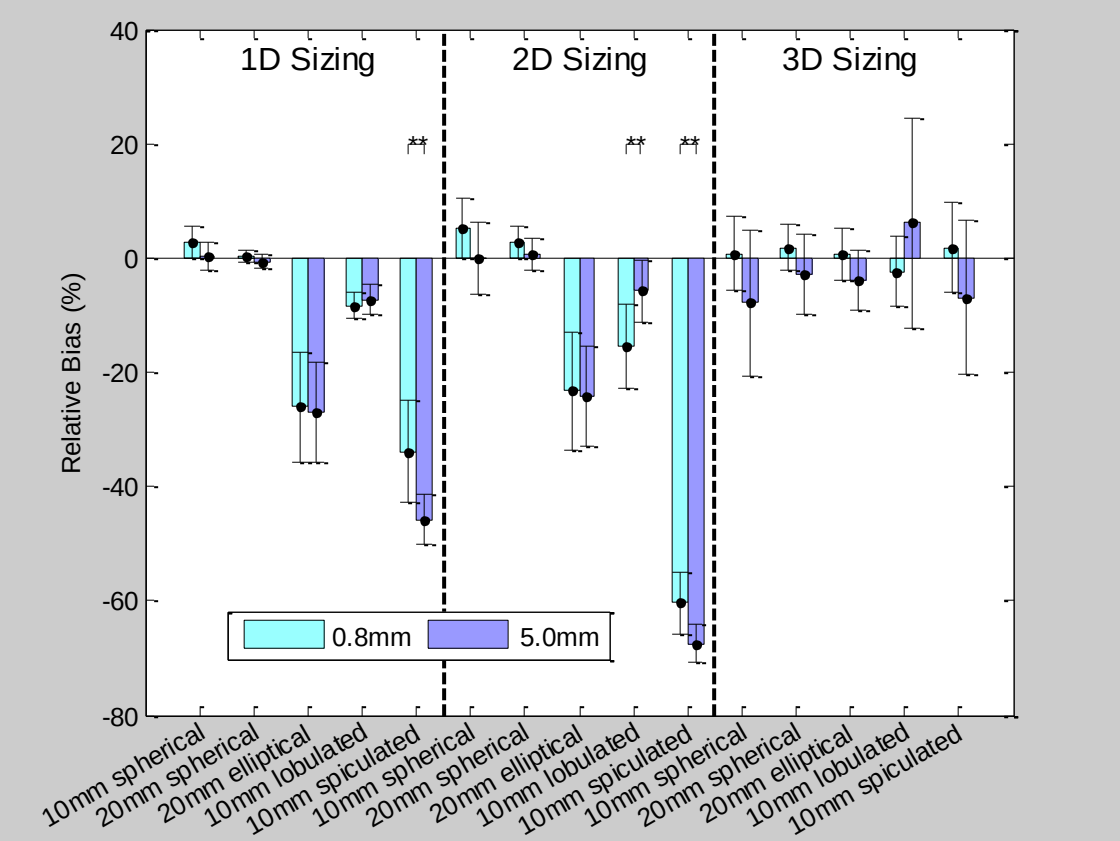
*Red line: median; Blue box: 25% (Q1) to 75% (Q3) percentile; Black whiskers: Q1-1.5(Q3-Q1) and Q3+1.5(Q3-Q1).
†Boxplot distributions are generated using 2-way bootstrap resampling (readers by all others).

Figure 6: Boxplot of relative standard deviations as a function of sizing method and nodule type.



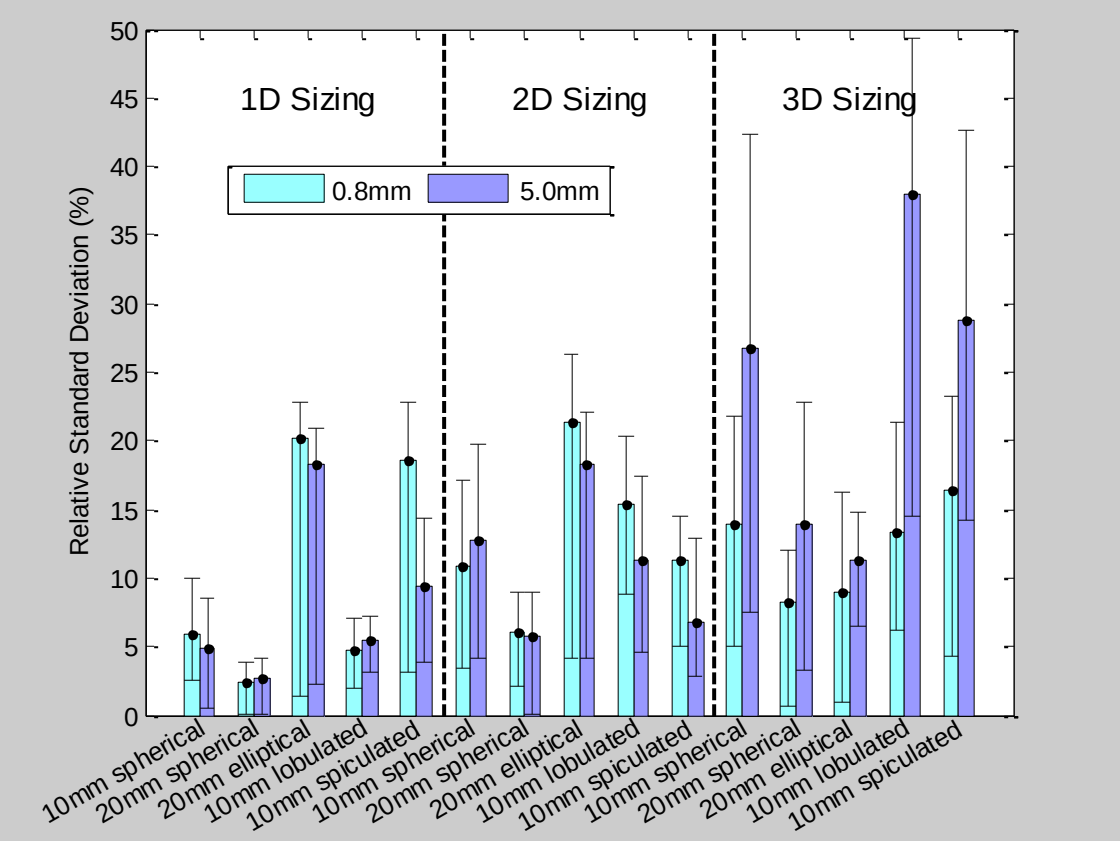
**Indicates a statistically significant difference between the two sizing methods represented by the connecting line.

Figure 7: Comparison plots of relative standard deviation among the sizing methods as a function of nodule shape. The 95% CI's and significance are determined using 2-way bootstrap resampling (readers by all other factors) and adjusted using a Bonferroni correction for 15 comparisons.



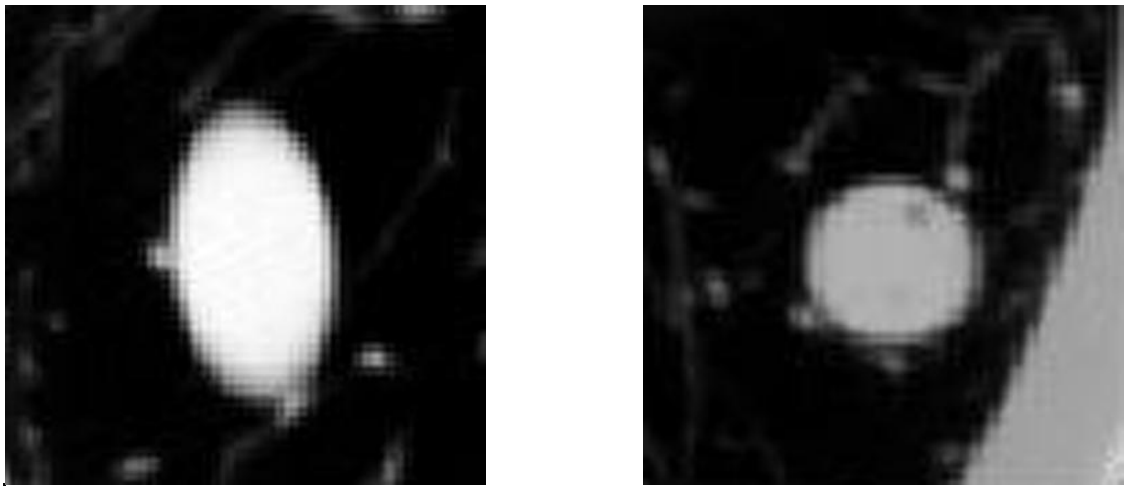
**Indicates a statistically significant difference between the two sizing methods represented by the connecting line.

Figure 8: Comparison plots of relative bias between thin 0.8 mm and thick 5.0 mm slice data for each nodule type, sizing method combination. The 95% confidence intervals and significance are based on the t-distribution applied within each subgroup and adjusted using a Bonferroni correction for 15 comparisons.



**Indicates a statistically significant difference between the slice thicknesses represented by the connecting line (no comparisons achieved statistical significance in this plot).

Figure 9: Comparison plots of relative standard deviation between thin 0.8 mm and thick 5.0 mm slice data for each nodule type, sizing method combination. The 95% CIs and significance are determined using 2-way bootstrap resampling (readers by all others) and adjusted using a Bonferroni correction for 15 comparisons.



(a) (b)

Figure 10: Central cross-sections of the (a) -10 HU and (b) +100 HU ellipsoid nodules. The figure depicts the substantial orientation difference between the two ellipsoid nodules. This resulted in very different in-plane longest dimension and longest perpendicular dimension measurements between the scans.