Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example

Nancy A Obuchowski, Huiman X Barnhart, Andrew J Buckler, Gene Pennello, Xiao-Feng Wang, Jayashree Kalpathy-Cramer, Hyun J (Grace) Kim, Anthony P Reeves and for the Case Example Working Group

Stat Methods Med Res published online 11 June 2014
DOI: 10.1177/0962280214537392

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–34

DOI: 10.1177/0962280214537392
smm.sagepub.com
SAGE

Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example

Nancy A Obuchowski,[1] Huiman X Barnhart,[2] Andrew J Buckler,[3] Gene Pennello,[4] Xiao-Feng Wang,[1] Jayashree Kalpathy-Cramer,[5] Hyun J (Grace) Kim,[6] Anthony P Reeves[7] for the Case Example Working Group

Abstract

Quantitative imaging biomarkers are being used increasingly in medicine to diagnose and monitor patients' disease. The computer algorithms that measure quantitative imaging biomarkers have different technical performance characteristics. In this paper we illustrate the appropriate statistical methods for assessing and comparing the bias, precision, and agreement of computer algorithms. We use data from three studies of pulmonary nodules. The first study is a small phantom study used to illustrate metrics for assessing repeatability. The second study is a large phantom study allowing assessment of four algorithms' bias and reproducibility for measuring tumor volume and the change in tumor volume. The third study is a small clinical study of patients whose tumors were measured on two occasions. This study allows a direct assessment of six algorithms' performance for measuring tumor change. With these three examples we compare and contrast study designs and performance metrics, and we illustrate the advantages and limitations of various common statistical methods for quantitative imaging biomarker studies.

Keywords

bias, repeatability, reproducibility, agreement, limits of agreement, coverage probability, intraclass correlation coefficient

[1]Cleveland Clinic Foundation, Cleveland, OH, USA
[2]Duke University, Durham, NC, USA
[3]Elucid Bioimaging Inc., Wenham, MA, USA
[4]Food and Drug Administration/CDRH, Washington, DC, USA
[5]MGH/Harvard Medical School, Boston, MA, USA
[6]UCLA, Los Angeles, CA, USA
[7]Cornell University, Ithaca, NY, USA

Corresponding author:
Nancy A Obuchowski, Quantitative Health Sciences/JJN 3, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195, USA.
Email: obuchon@ccf.org

1 Introduction

Quantitative imaging biomarkers (QIBs) are being used increasingly in medicine in the diagnosis, prognosis, staging, and monitoring of patients' disease. For example, volumetrically determined growth rates of pulmonary nodules are helpful in differentiating benign and malignant lesions (diagnosis role), the absolute volume of a nodule is used for clinical staging, and the change in nodule volume after the initiation of treatment helps physicians assess the response to treatment. Other QIBs include coronary artery calcium scores from computed tomography (CT) for atherosclerosis screening, standardized uptake values in positron emission tomography imaging to measure cancer cell activity, and dynamic contrast-enhanced magnetic resonance imaging for measuring microvascularity of tumors. The Radiological Society of North America (RSNA) created the Quantitative Imaging Biomarker Alliance (QIBA) to investigate the role of quantitative imaging methods, including CT volumetry, in evaluating disease and responses to treatment.[1–3]

For QIBs such as CT nodule volume, there are multiple computer algorithms available for measuring the volume of a tumor and the change in nodule volume. Different QIB algorithms, however, have different technical performance characteristics. One algorithm might have less bias than another but also have less precision. The bias and precision of the algorithms may depend on characteristics of the lesion, e.g. its size and/or shape, and the pattern of bias and precision may differ among algorithms.

In this paper we compare various commonly used study designs and illustrate the analysis of QIB algorithm comparison studies. We do not develop any new statistical methods, but rather present existing metrics[4] and their analysis,[5] illustrating their relative strengths and weakness on common datasets. We focus on methods to (i) characterize the range of performance of a group of algorithms, (ii) test algorithm performance against a specified performance claim, (iii) assess agreement among algorithms, and (iv) identify the best algorithm among several competing algorithms. Performance is defined using characteristics such as bias, linearity, and precision.[4,6] Since it is unusual for any single study to address all of these characteristics, multiple studies are usually required for complete characterization. Different study designs generally complement each other to optimally evaluate these characteristics.

We use the analyses of pulmonary nodule measurement as examples in this paper due to the considerable attention that this task has received. Reasons for this attention include: the important clinical impact of size change measurements for early diagnosis of lung cancer and timely evaluation of response to therapy, and the clear image presentation of these nodules in CT images when compared to other medical image measurement tasks. This task also illustrates major issues in QIB evaluation that challenge current statistical methods. These issues include: (a) the lack of the true value of the QIB and (b) the vast range of image presentations of these nodules coupled with a very strict limit on the sample size that can be practically employed in a study. We use both the absolute volume of a nodule and the change in nodule volume as measured on CT as illustrations.

In Section 2 we discuss common study designs for comparing QIB algorithms. We describe the design of three CT volumetric studies comparing the technical performance of multiple computer algorithms. One study used phantoms to assess the bias and repeatability of an algorithm for measuring tumor volume (Section 3). The second study used anthropomorphic phantoms to compare the bias and reproducibility of different algorithms (Section 4). The third study used an in vivo test–retest design to assess the agreement and bias for measuring change in tumor volume (Section 5). In Section 6 we discuss the limitations of the analyses and future directions. Though this paper presents different statistical methods for the analysis of CT volumetric image measurements, this is not to imply that such methods are sufficient to address all the issues in the measurement of pulmonary nodules.

2 Study design comparisons

The two most common types of studies for comparing QIB algorithms are those based on synthetic data, i.e. through simulated data or anthropomorphic phantoms where truth is known and in vivo studies.

An anthropomorphic phantom is a synthetic physical model of the target of interest; the model and target of interest resemble the human body as closely as possible. The targets, e.g. pulmonary nodules, have been built to a specified shape and size; thus, the true value is known. The model is imaged using standard clinical imaging devices as used to image human subjects.

In vivo studies use human subject cases to study and compare algorithm performance. There are a number of options to assess bias in these studies. In a reference standard study, human subjects are imaged with the QIB algorithm machines as well as with a reference standard. A reference standard can be expert human measurements, another QIB algorithm which is accepted as an unbiased measure of the true value albeit with measurement error, or a synthesis of multiple algorithms computed, for example, by the Simultaneous Truth And Performance Level Estimation method.[7] For measuring tumor volumes, there is no suitable reference standard. Results from the National Cancer Institute Lung Image Database Consortium public database indicate that there is large variation in the measurement of pulmonary nodules by experienced radiologists,[8] suggesting that manual markings are not suitable as a reference standard. In this paper we will not further discuss this type of study. In the test–retest design, sometimes called "coffee-break studies," human subjects are imaged with the QIB algorithm machines at two or more time points very close in time, e.g. 5 min apart, so that there is no possibility of a true biological change occurring.[5] These studies offer a valid assessment of algorithms' reproducibility, as well as their bias for measuring the change in tumor volume when there is no true change. In vivo studies are also performed to assess agreement among algorithms where no true value is needed.

There are important advantages and disadvantages of each design (see Table 1). Clinical studies offer the best estimates of the precision of a QIB because the inherent variability in human subjects is incorporated into the estimates of precision, whereas precision may be underestimated in a phantom study. Bias in measuring the change in tumor volume can be addressed in both phantom and clinical studies, but neither design is ideal. Phantom studies provide the true value for a range of lesion sizes, shapes, and location; however, phantoms do not simulate well the wide range of human anatomy complexities. For example, a number of factors have been shown in the literature to affect the performance of pulmonary nodule change measurements. These include:

(1) Nodule size.[9]
(2) Nodule complexity: chest wall attachment, vessel attachment, and surface irregularities.[10,11]
(3) Nodule texture, e.g. nonsolid nodules[12] within which the more dense vascular structure is partially visible.
(4) Nodule margins: the transition region between a tumor and the surrounding lung parenchyma is different than the nodule center.[13]
(5) CT scanner parameters: mainly slice thickness, dose, and reconstruction method.
(6) The computer algorithm.
(7) For algorithms that involve human interaction, the skill of the operator.[14]

The in vivo studies incorporate all of these effects; however, the small study size precludes characterization of each specific effect. The phantom method currently involves homogenous, smooth surface models and does not consider effects 2–4 listed earlier.

Table 1. Comparison of anthropomorphic phantom and test–retest clinical designs.

Design features	Phantom study	Clinical study
Technical performance:		
Bias	Valid assessment for simple nodules, but limited generalizability to complex nodules often seen in humans	Valid assessment for assessing no change
Linearity	All variables, except the one being tested (e.g. size), can be held constant to assess linearity	Not usually possible to assess
Precision	Replicate measurements easily acquired for estimating repeatability and reproducibility, but limited generalizability to complexities seen in humans	Test–retest design offers valid estimates of the reproducibility of measurements
Complexity in conducting study	Phantom must be designed and built	Patients must be consented and Institutional Review Board (IRB) approval required
Adequacy of resources[a]	Large number of replicate measurements easily obtained. The same phantoms can be used repeatedly by many algorithms. Construction of a large number of unique nodules can be difficult	Number of replicate measurements is sometimes limited by the characteristics of the imaging methods (e.g. radiation with CT). Depending on the study protocol, the number of available patients can be a limiting resource

[a]In both study designs, the number of images that can be processed by the algorithms is limited by the human effort in conducting the measurements.

In vivo test–retest studies provide the opportunity for measuring bias in patients, but only under the scenario of no change. These studies cannot be used to compare algorithms' ability to estimate the magnitude of change when the change differs from zero.

For studies aimed at assessing algorithm performance for quantifying the change over time, the precision of the estimate of change is critical to both the comparison of the algorithms, as well as to the interpretation of the estimated change. To estimate the precision of an estimate of change from a phantom study, linearity is a critical issue. Linearity, in the general sense, means that the measured quantity values (Y) bear a linear relationship of the form: $Y = a + bX$, where X is the reference (or true) value of the measurand in the sample.[6] It is well recognized that many QIB measurements follow a linear relationship with X only over a narrow range of values of X; for other values of X, the QIB measurements usually adhere at least to a monotonic relationship with X such that there is a single biomarker value for each value of X.[4]

Phantom studies may be the preferred study design for assessing cross-sectional (i.e. one time-point) linearity because the location, density, and shape of the lesion can be held constant while changing only the lesion size. Linearity with respect to other features, e.g. lesion density, can be assessed with a similar design. If the linearity assumption is shown to be reasonable for the range of plausible values of X, then one can use a simple error propagation formula to estimate the precision of the estimated change from the cross-sectional estimate of precision. Let $s(Y)$ denote the estimated precision of Y at a single time point. $s(Y)$ is often expressed as the standard deviation (SD) of Y but

other measures of precision are also common. If $s(Y)$ is a constant value not related to the value of X, then an upper bound (assuming a positive correlation) on the precision of an estimate of the measured change between time t = 0 and t = t is given by

$$s(Y_0 - Y_t) = \sqrt{2 \times s(Y)^2} \tag{1}$$

For example, let wSD be the within-subject standard deviation of a QIB algorithm measuring nodule volume at a single time point. Suppose wSD is 15. Then from equation (1) the estimated wSD of the change in nodule volume, wSD_Δ, is 21.2.[15] If, on the other hand, $s(Y)$ changes in magnitude with, say, the true size of the lesion, X, then a reasonable upper bound on $s(Y_0 - Y_t)$ is given by

$$s(Y_0 - Y_t) = \sqrt{s(Y_0)^2 + s(Y_t)^2} \tag{2}$$

where $s(Y_0)$ and $s(Y_t)$ are the precision estimates of the nodule volume at baseline and time t, respectively. Note that equations (1) and (2) provide an upper bound on the precision for the change measured by the algorithm, but this may not translate directly to the true change if b in $Y = a + bX$ differs from one. In Section 4.5 we illustrate the calculations for measuring a confidence interval (CI) for the true change when $b \neq 1$.

The estimates of uncertainty in change measurements in equations (1) and (2) do not take into account the within-subject correlation. The within-subject correlation is the correlation in the measurements at the two time points due to the fact that it is the same lesion in the same patient being measured at two time points. A more appropriate formula is

$$s(Y_0 - Y_t) = \sqrt{s(Y_0)^2 + s(Y_t)^2 - 2 \times r \times s(Y_0) \times s(Y_t)} \tag{3}$$

It is not easy, however, to estimate the within-subject correlation, r, from a phantom study. One could conceive of a phantom study where a larger nodule, of the same density and shape, could be inserted at the same location as a smaller nodule to assess this correlation. However, patient orientation and patient motion are critical factors in measuring change over time in real clinical cases because they affect the magnitude of the within-subject correlation; these variables cannot be easily accounted for in phantom studies. Furthermore, the magnitude of the correlation may change over time and with the aggressiveness of the disease, often attenuating over increased time intervals and with more aggressive variants of the disease. For these reasons, investigators often take a conservative approach and set the correlation to zero, thus utilizing the formulae in equations (1) and (2).

In contrast to phantom studies, clinical studies are well suited to measuring the precision in estimates of the change over time. The within-subject correlation is inherently incorporated into the estimates of change, but the degree to which this correlation is accounted for depends on the QIB algorithm. Some algorithms measure nodule volume at each of the two time points independently, without using information from the other image. Other algorithms utilize the baseline image when making measurements at the second time point. Still other algorithms do not measure nodule volume at each of two time points; rather, these algorithms directly estimate the change in volume by measuring the difference in lesion overlap on the images at the two time points. One might expect that the precision of the estimates of change would be smallest for the latter algorithms and larger for the former.

The clinical context plays an important role in identifying the ''best'' algorithm. For monitoring purposes we may be interested in determining if *any* change in nodule volume has occurred during a short time interval, regardless of magnitude; thus, the clinical needs are for a binary decision, and

clinical studies may be sufficient. On the other hand, if the magnitude of the change impacts the decision to alter treatment or not, then an unbiased measurement of the degree of change may be required. Linearity is important in both clinical scenarios and should never be assumed.

In this paper we illustrate the analysis of three studies: a study originally published by Chen et al.[16] which is a phantom study evaluating the bias and repeatability of tumor volume measurements (Section 3), the QIBA 3a study which is a phantom study assessing bias and reproducibility (Section 4), and the CT VOLume Change Analysis of NOdules program (VOLCANO) study (Section 5) which used an in vivo test–retest design to compare the bias of tumor volume change measurements.

3 Bias and repeatability example

The data are taken from the algorithm performance study described in Chen et al.[16] The study employed an anthropomorphic thoracic phantom with inserted synthetic nodules of realistic structures such as ribs and a removable lung insert. Two acrylic spherical nodules (size diameters of approximately 5 and 10 mm) were included in this example. Eight nodules from each category were randomly attached to the vessels of the lung inserts and pleura. The true volume of each nodule category was calculated using its diameter assuming a perfect sphere and verified by liquid replacement measurements in a graduated cylinder. The images were acquired with 40-mm detector width, 120 kVp, 1.375 pitch, and 6.22 mGy computed tomography dose $\text{index}_{\text{vol}}$ and were reconstructed at three slice thicknesses (0.625, 1.25, and 2.5 mm, with equal slice spacing and thickness), using filtered backprojection (FBP) with the kernel "Standard." The acquisition was repeated 10 times for each nodule of a given size, slice thickness, and algorithm without repositioning the phantom. We denote the measured value as Y_{ijsk}, where j denotes the jth nodule size ($j = 1$ or 2 for 5 and 10 mm, respectively), i denotes the ith nodule ($i = 1, \ldots, 8$), s denotes the sth slice thickness ($s = 1$, 2, or 3 for 0.625, 1.25, and 2.5 mm, respectively), and k denotes the kth repetition ($k = 1, \ldots, 10$).

The volume of each nodule was quantified from the reconstructed CT images by a single observer using a clinical lung analysis software package in which nodules were segmented in a semiautomated fashion assuming "solid" density. In challenging cases such as those for nodules with complex vessel attachments, the segmentation procedure sometimes failed. As a result, only six of the eight 5-mm acrylic nodules, and seven of the eight 10-mm acrylic nodules were successfully segmented. Thus, the study contains the measured volumes of six small and seven large nodules with 10 repetitions for each of the three slice thicknesses used for image reconstruction under the same scanning condition. For the purpose of the analyses below, the "algorithm" includes the FBP reconstruction at the prescribed slice thickness as well as the semiautomated segmentation algorithm used to calculate the nodule volumes for those slices.

The results presented here differ from that published by Chen et al.,[16] as follows: (1) We used a subset of the full dataset; (2) rather than reporting relative measures of performance (i.e. relative to the true volume), we report estimation of performance without such scaling for better interpretation; and (3) to assess the overall agreement between the QIB and true value, instead of aggregating bias and precision, we report here the unaggregated agreement results. The data are available at www.smmr.givewebsitehere.

3.1 Assessment of bias

We start our analysis with investigating the bias of the algorithm's measurements of nodule volume. Table 2 provides a descriptive summary of the algorithm's estimated *bias* by nodule size and

Table 2. Mean (min, max) bias in mm^3 by nodule size and slice thickness.

	Nodule size			
Slice thickness (mm)	5 mm	10 mm	Difference (95% CI)	P-value
0.625	3.42 (−0.6, 10.4)	32.23 (0.5, 72.5)	−28.8 (−40.8, −16.9)	<0.001
1.25	3.58 (−0.6, 14.4)	32.71 (3.5, 69.5)	−29.1 (−38.3, −20.0)	<0.001
2.5	3.56 (−5.6, 29.4)	39.21 (−55.5, 93.5)	−35.6 (−51.5, −19.8)	<0.001

slice thickness. This is the mean of the *individual bias*[5] (i.e. individual bias is the mean over each nodule's measurements of volume minus the true volume for that nodule, i.e. $\sum_{k=1}^{10}\left(Y_{ijkl} - X_j\right)/10$). The algorithm, on average, gives larger volume measurements than the true volume.

We next tested whether the bias of the algorithm was similar for nodules of different size and for different slice thicknesses. The analyses were carried out utilizing the generalized estimating equations (GEE) approach in SAS procedure GENMOD, with observed individual bias as the response and nodule size and slice thickness as main effects and their two-way interaction. The p-values in Table 2 were obtained by using the contrast statements in GENMOD to compare the biases between the two nodule sizes within a given slice thickness based on this full model. We found that there is no significant interaction effect and no significant main effect of slice thickness after removing the interaction from the model. The main effect of nodule size combining all three slice thickness is highly significant with $p = 0.002$. Thus, we conclude that the bias for large nodules is significantly larger than the bias for small nodules for all slice thicknesses. Although there appears to be a slight increase in bias as slice thickness increases, the difference is not statistically significant.

Note that the GEE approach was used to account for the correlation between the 10 replications on the same nodule within a given slice thickness. The identity link function was used because we have a continuous outcome and want to compare the mean difference. An independent working correlation structure was used in GENMOD for estimation and model-based estimation, but the empirical standard errors, which are robust to misspecification of the working correlation matrix, are used for inference. The advantage of using the GEE approach for correlated data is that it does not make any distributional assumptions. In fact, it makes only two assumptions: (1) the marginal mean model is linear in terms of the mean function and (2) the number of clusters (number of nodules in this example) is relatively large. The first assumption is not restrictive because most statistical models make this assumption, and for our example there is no restriction at all because we have two parameters, one for each of the two groups. As with any asymptotic technique, a large total sample size is required for consistent estimation. What represents a large sample size, however, is not unanimously agreed upon. The determination depends on a number of factors including the number of parameters and the size of the clusters relative to the number of clusters. A rule of thumb for the number of clusters is no fewer than 10, and preferably more than 30.[17] Our number of cluster is 13. Sherman and le Cessie[18] showed that a bootstrap approach may perform better than GEE in small sample size situations. We also applied the bootstrap approach with 1000 bootstrap samples and obtained similar results as the GEE results (results not shown). Another alternative to GEE is hierarchical linear modeling, which allows for a nesting structure to account for the correlations within a cluster and may be the preferred approach when the number of clusters is very small.

3.2 Assessment of repeatability

Here, we assess the precision of the algorithm. Because we have 10 replications under the same scanning and reconstruction conditions, the algorithm's *repeatability*[6] can be assessed. Table 3 provides a descriptive summary of the algorithm's repeatability by nodule size and slice thickness. Four repeatability metrics are illustrated: within-subject standard deviation (wSD), the repeatability coefficient (RC), the within-subject coefficient of variation (wCV), and the intraclass correlation coefficient (ICC).[5]

A test of the equality of wSD and RC between the 5 and 10 mm nodules is equivalent to a test of the equality of the mean sample variances of replications between the two nodule sizes.[5] Such a test can be carried out using the SAS procedure GENMOD with sample variance of replications for each nodule as the response and nodule size as the independent variable; identical p-values are obtained for wSD and RC. The same SAS procedure was used to test the equality of wCV between the two nodule sizes with observed individual wCV (calculated from the 10 replications for each nodule). Again the identity link and independent working correlation structure were used in SAS procedure GENMOD.

From Table 3, the wSD is larger for larger nodules and increases with slice thickness. The RC shows the same pattern because it is equivalent to wSD, but with a different interpretation.[5] The wCV, on the other hand, does not exhibit this trend because it is a function of both the standard deviation and the nodule size (i.e. wSD/mean) and thus is proportional to the magnitude of the lesion's size. One needs to be cautious in interpreting the test on equality of wCVs because a difference of 10% wCV for a large nodule (e.g. 10% of true volume 452.5 mm^3 is 45.25 mm^3) does not have the same meaning as a difference of 10% wCV for a small nodule (e.g. 10% of

Table 3. Estimated repeatability by nodule size and slice thickness.

	Nodule size		
Repeatability parameter	5 mm	10 mm	p-value
wSD			
Slice thickness = 0.625	0.56	8.29	0.009
1.25	0.63	10.28	0.004
2.5	3.54	14.02	0.17
RC			
Slice thickness = 0.625	1.56	22.99	0.009
1.25	1.75	28.49	0.004
2.5	9.81	38.86	0.17
wCV (%)			
Slice thickness = 0.625	0.8%	1.4%	0.12
1.25	0.9%	1.8%	0.08
2.5	3.3%	2.1%	0.43
ICC			
Slice thickness = 0.625	0.98	0.80	NA
1.25	0.99	0.56	NA
2.5	0.86	0.68	NA

Note that the RC = 2.77 × wSD.

volume 56.6 mm^3 is 5.66 mm^3), particularly if we think that a 5.66 mm^3 difference in volume among replications is acceptable, while a 45.25 mm^3 volume difference is not.

The precision can also be evaluated by the ICC, which is a measure of the agreement between the 10 replicated measurements of the CT volumes. The estimated ICC values are greater than 0.8 for the measurements of the small nodules but are 0.8 or less for the measurements of the large nodules. Note that the ICC is a relative index; it depends on the between-nodule variability. Since the between-nodule variability differs in magnitude for small and large nodules, the ICCs of the small and large nodules are not comparable.

3.3 Agreement between measured and true volumes

Bias and precision provide two separate evaluations of algorithm performance, but they do not provide the aggregated impact of bias and precision on the agreement between the measured volume and the true volume.[5] There is a trade-off between bias and precision, and different agreement indices evaluate this trade-off differently. We next consider several aggregate measures of the algorithm's performance.

Two different kinds of limits of agreement (LOAs) may be considered for data with replicated measurements. The first kind of LOAs has to do with the distribution of the difference between a measured volume and the true value for a random replication for a given nodule, where the distribution is across replications and this distribution is assumed to be the same for all nodules. Specifically, consider the differences of all replicated volumes and the true volume for a given nodule, i.e.$\left(Y_{ijkl} - X_j\right)$ for all k replications of nodule i. We are interested in knowing the limits within which 95% of such replicated differences fall for a given nodule size j and slice thickness s. Assuming such limits are the same for all nodules, given nodule size j and slice thickness s, the LOAs are estimated as follows

$$\widehat{LOAs}_{95\%js}(replications) = \bar{D}_{js} \pm 1.96 \times \widehat{wSD}_{js}$$

where $\bar{D}_{js}$ and $\widehat{wSD}_{js}$ are the bias and precision estimates in Tables 2 and 3, respectively (see Appendix 1 for the derivation). Note that bias and precision contribute aggregately to the estimates of LOAs. Also note that it is much easier to show that an algorithm's mean bias is within an acceptable difference of $\pm d_0$ than it is to show 95% of the replicated differences are within $\pm d_0$.

The second kind of LOAs considers the distribution of the difference between a measured volume and the true value for a randomly chosen nodule, where the distribution is across nodules, rather than across replications. Such 95% LOAs are expressed as

$$LOAs_{95\%js}(nodules) = D_{js} \pm 1.96 \times SD_{js}(measured - truth)$$

where $SD_{js}(measured - truth) = \sqrt{wSD_{js}^2 + \sigma_{Njs}^2 + \sigma_{Tjs}^2}$, which includes variations from within (wSD_{js}^2) and between (σ_{Njs}^2) nodules, as well as variation of true volumes (σ_{Tjs}^2) across nodules and the mean bias ($D_{js} = \sum_{i=1}^{8} \bar{D}_{js}/8$) (see Appendix 1 for derivation of SD_{js} and its estimation). We consider this second kind of agreement index here.

Following Bland and Altman[19] on estimating LOAs by using data with replicates, wSD_{js}^2 can be estimated by $\widehat{wSD}_{js}^2 = s_{\bar{d}_{js}}^2 + \left(1 - \frac{1}{K}\right)\widehat{wSD}_{js}^2$, where $s_{\bar{d}_{js}}^2$ is the sample variance of mean differences

Table 4. Agreement of measured and true volumes by nodule size and slice thickness.

	LOA_80%	TDI_80% 95% CI	LOA_95%	TDI_95% 95% CI	CP_10 95% CI	CP_50 95% CI
Diameter = 5 mm						
Slice thickness						
0.625 mm	(−2.0, 8.84)	7.4 (1.4, 10.4)	(−4.89, 11.73)	10.4 (2.4, 10.4)	0.87 (0.45, 0.98)	1.00 (NA)
1.25 mm	(−3.42, 10.58)	7.4 (0.6, 13.4)	(−7.15, 14.31)	13.4 (1.4, 14.4)	0.83 (0.37, 0.98)	1.00 (NA)
2.5 mm	(−8.72, 15.84)	8.4 (3.6, 24.4)	(−15.24, 22.36)	26.4 (4.6, 26.4)	0.87 (0.45, 0.98)	1.00 (NA)
Diameter = 10 mm						
Slice thickness						
0.625 mm	(8.44, 56.02)	44.5 (27.5, 68.5)	(−4.19, 68.65)	68.5 (42.5, 70.5)	0.07 (0.02, 0.27)	0.83 (0.50, 0.96)
1.25 mm	(12.93, 52.49)	41.5 (33.5, 57.5)	(2.43, 62.99)	61.5 (41.5, 64.5)	0.07 (0.02, 0.27)	0.89 (0.73, 0.96)
2.5 mm	(7.33, 71.09)	48.5 (37.5, 82.5)	(−9.61, 88.03)	85.5 (44.5, 88.5)	0.04 (0.01, 0.23)	0.81 (0.46, 0.96)

$\bar{d}_{ijs} = \left(\bar{Y}_{ijs} - X_j\right)$ where $\bar{Y}_{ijs}$ is the mean of the 10 replicates. Table 4 presents the agreement assessment based on these LOAs, as well as two other measures of agreement: the total deviation index (TDI) with 80 and 95% coverage probability (CP), and the CP for acceptable differences of 10 and 50 mm^3.[5]

The TDI_80% and TDI_95% are computed nonparametrically by using the quantile regression with SAS procedure QUANTREG, where the response is the absolute difference of measured and true volume, with intercept term only and with quantiles set at 0.8 and 0.95. The CP_10 and CP_50 are estimated as the proportion of the absolute difference less than or equal to 10 or 50, respectively.[5] The same CP estimate can also be obtained by fitting a GEE model in SAS procedure GENMOD, where the binary response takes a value of one if the absolute difference of measured volumes and true value is less than or equal to 10 (or 50) and takes a value of 0 otherwise. The CI for the TDI estimate was obtained by the bootstrap method, where bootstrap samples were taken at the level of nodules to account for within-nodule correlation. Then, with bootstrap samples as data input, the bootstrap percentile method can be performed by SAS procedure QUANTREG. To account for within-nodule correlation, the empirical standard errors were used to construct the CI for CP where the logit link and independent working correlation were specified in the GENMOD procedure.

If there is no systematic bias and the difference is normally distributed, we expect the LOAs and (−TDI, TDI) to be similar. Since there are positive biases as indicated in Table 2, the LOAs are not symmetric around zero. The upper limit of LOA_80% is larger than the TDI_80%, while the upper limit of LOA_95% and TDI_95% is similar due to limited data at the 0.95 quantile. If we require 80% differences to be within 10 and 50 mm^3 between measured and true volumes for small and large nodules, respectively, then we would conclude that the algorithm is acceptable based on the TDI index, but not acceptable based on the LOA index due to its asymmetry. The same conclusion can be made based on CP_10 $\geq$ 80% for small nodules and based on CP_50 $\geq$ 80% for large nodules.

4 QIBA 3A project

The QIBA CT Volumetry Technical Committee is investigating the hypothesis that volumetry is an effective method for quantifying treatment-induced changes in tumor volume, and ultimately, changes in the health status of patients with cancer. The committee is conducting work to identify and evaluate volumetry methods in collaboration with Food and Drug Administration

Division of Applied Math/Office of Science and Engineering Laboratories/Center for Devices and Radiological Health, National Cancer Institute, National Institute of Standards and Technology, American College of Radiology Imaging Network, major imaging equipment manufacturers, value-added software companies, the Pharma Imaging Group, scientists from academia, and others. To do this, a variety of experimental projects including this one have been undertaken. One of these efforts, known internally as ''3A,'' has been instituted to study and report on algorithm performance.

The 3A study has been conducted to evaluate the technical performance of algorithms applied on synthetic nodules from CT scans of anthropomorphic phantoms. The study was organized as a public ''challenge.'' Twenty-two unique synthetic lung nodules varying in size (5–40 mm), shape (spherical, elliptical, lobulated, spiculated), and density (−630, −10, +100 HU) were constructed. The nodules (possibly the same nodule) were placed in various locations and in different slice thicknesses such that each unique nodule was imaged multiple times. Details of the phantoms and synthetic nodules can be found at www.qi-bench.org. Anonymized participants associated from academic and commercial algorithm developers then downloaded the study data from QI-Bench,[2] performed their volumetric algorithm on each of the data lesions, and reported the anonymized lesion volume results using the RSNA-provided ID to replace the participant organization's name. Study 3A proceeded in two phases, a pilot and a pivotal, using disjoint input data. The pilot collected 97 observations from each of 12 participants, and the pivotal phase collected 408 observations from each of 10 participants. For this paper, we randomly selected data from the pilot phase based on four randomly selected participants. This was intentionally done for purposes of providing an example, without intending to either completely analyze all the available pilot data or to represent actual participants; thus, the results do not represent the findings of the QIBA 3A study. This resulted in a total of 90 observations on 22 physically unique lesions by each of the four participants. Each participant used a different algorithm to measure lesion volume and we thus evaluate the performance of these four algorithms below.

4.1 Profile of algorithms' bias and reproducibility

There are several goals for the analysis of these data. The first goal is to characterize the range of performance of a group of algorithms in measuring nodule volume. Thus, we begin with an overall profile of algorithms' performance (bias and precision). The true volume of the nodules ranged from 282.35 to 34,389.21 with three groups: (1) 36 observations on 10 unique nodules with 8 and 10 mm nominal diameters where the true volume ranges from 282.35 to 706.75, (2) 44 observations on 10 unique nodules with 20 mm nominal diameter where true volume ranges from 4207.83 to 5279.07; and (3) 10 observations on two nodules with 40 mm nominal diameter where the true volume ranges from 33,781.46 to 34,389.21. There are multiple observations on each unique nodule because each nodule was placed in different locations of the lung and/or different slice thicknesses were used. Table 5 provides the estimated individual bias (i.e. for each unique nodule, the bias is the mean of the differences between an algorithm's measurements and the true volume of the nodule) and individual reproducibility[6] (i.e. within-nodule standard deviation) for each of the 22 physically unique nodules. Figures 1 and 2 are the plots corresponding to the data in Table 5 for individual bias and reproducibility, respectively, at the nodule level.

4.2 Comparison of algorithms' bias for measuring tumor volume

Our second goal for this study is to identify the best algorithm among the competing algorithms. We begin with a comparison of the algorithms' bias. From Table 5, the bias tends to be associated with

Table 5. Individual bias and precision at nodule level.

Nodule # (diameter)[a]	True volume	n[b]	Estimated mean bias of each algorithms				Estimated within-nodule SD of each algorithm			
			1	2	3	4	1	2	3	4
1 (8 mm)	282.35	6	16.08	34.65	−45.35	69.54	56.46	93.18	138.30	53.35
2 (10 mm)	471.30	2	9.80	124.70	35.15	105.34	26.36	98.99	207.11	51.26
3 (10 mm)	524.63	2	−24.28	−94.13	−29.63	34.86	60.31	0.71	7.07	84.14
4 (10 mm)	526.64	5	45.28	−2.24	−9.04	173.68	104.41	31.44	125.82	100.37
5 (10 mm)	527.42	2	−53.12	−101.92	22.68	65.21	12.17	7.78	141.56	12.77
6 (10 mm)	528.67	1	18.33	−26.67	−278.67	301.33	NA	NA	NA	NA
7 (10 mm)	533.69	6	36.83	−10.02	−12.97	105.80	46.45	119.20	208.36	107.98
8 (10 mm)	569.17	4	27.04	−158.92	1455.83	5.15	158.56	170.67	663.85	128.19
9 (10 mm)	679.31	5	−125.46	−119.71	122.33	−27.59	40.04	126.81	118.35	120.96
10 (10 mm)	706.75	3	−98.24	−244.42	−63.32	−10.23	92.38	23.46	32.96	61.12
11 (20 mm)	4207.83	5	143.34	−38.63	222.97	531.96	137.87	304.71	587.59	264.64
12 (20 mm)	4215.10	3	−145.84	−521.43	871.50	171.29	240.04	691.56	393.84	316.95
13 (20 mm)	4232.05	6	50.35	−174.55	−242.85	525.27	135.82	465.62	680.53	243.65
14 (20 mm)	4234.41	3	169.80	−99.41	−137.04	373.71	158.99	66.14	1002.32	165.40
15 (20 mm)	4286.76	3	−88.18	48.24	390.94	317.50	193.43	454.18	153.83	263.75
16 (20 mm)	4315.84	2	−266.41	−404.34	159.91	511.75	93.27	2.12	672.81	468.68
17 (20 mm)	4350.42	6	−34.37	−193.75	430.53	643.13	92.89	101.01	435.12	330.47
18 (20 mm)	4920.35	5	−553.57	−1002.55	468.53	−690.41	350.33	455.44	721.19	152.39
19 (20 mm)	5062.15	5	−412.40	−1301.35	591.95	−891.68	478.00	402.90	808.69	38.11
20 (20 mm)	5279.07	6	−804.29	−1490.40	884.75	−1002.34	485.99	360.13	1002.86	109.58
21 (40 mm)	33781.46	6	710.81	−239.63	615.21	1925.19	930.23	843.45	299.58	1373.31
22 (40 mm)	34389.21	4	−133.08	−5566.71	1123.49	556.83	285.31	3708.34	790.64	1088.33

[a]For each synthetic nodule, an equivalent diameter is approximated as a rounded representation of the diameter that a sphere would have if it had the same volume as the nodule. This is provided as a means to establish categories of size in a form intuitive to practicing clinicians and useful for stratifying the statistical results.
[b]n is the number of observations by each algorithm for a given nodule.

nodule size, but the bias does not seem to be proportional to the nominal diameter nor to the true volume. Thus, when comparing the algorithms, we compare the bias according to the three groups of nodule size (defined by the nominal diameter). We observe that there is still substantial variation in bias among nodules with similar nominal diameters; this variation may be due to the physical characteristics of the nodule, e.g. shape and density, or due to the image slice thickness. Table 6 provides the summary statistics of the mean bias by nodule group and algorithm. Pairwise comparisons of the algorithms were performed within each nodule group via GEEs while accounting for the correlation among multiple measurements made on the same nodule by different algorithms. This can be accomplished by SAS procedure GENMOD with the observed bias as the response variable and algorithm as the independent variable with an identity link. The pairwise p-values were obtained through the contrast statement in SAS. Table 7 presents the p-values adjusted for multiple comparisons. Sidak's method[20] that controls the family-wise error rate (FWER) was used for each column. The method performs pairwise tests on differences between means with levels adjusted according to Sidak's inequality, which is a technique slightly less conservative than the Bonferroni procedure.

Using the results in Tables 6 and 7, we make the following conclusions. For nodules with 8–10 mm diameter, algorithm 2 had significantly smaller bias than algorithm 3. For nodules with

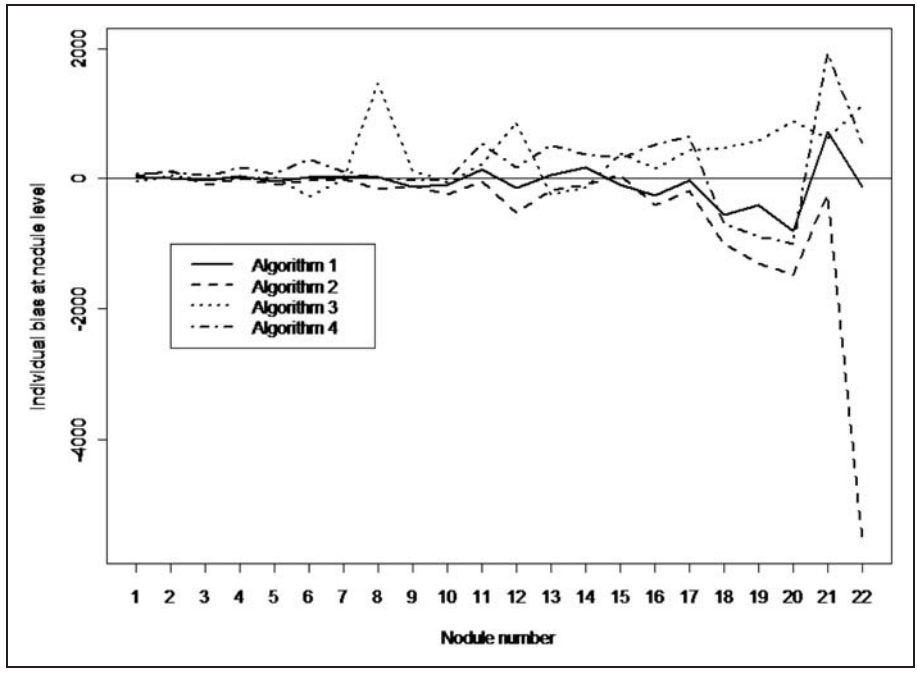

Figure 1. Plot of individual biases for 22 unique nodules ordered from smallest to largest volume (in mm^3).

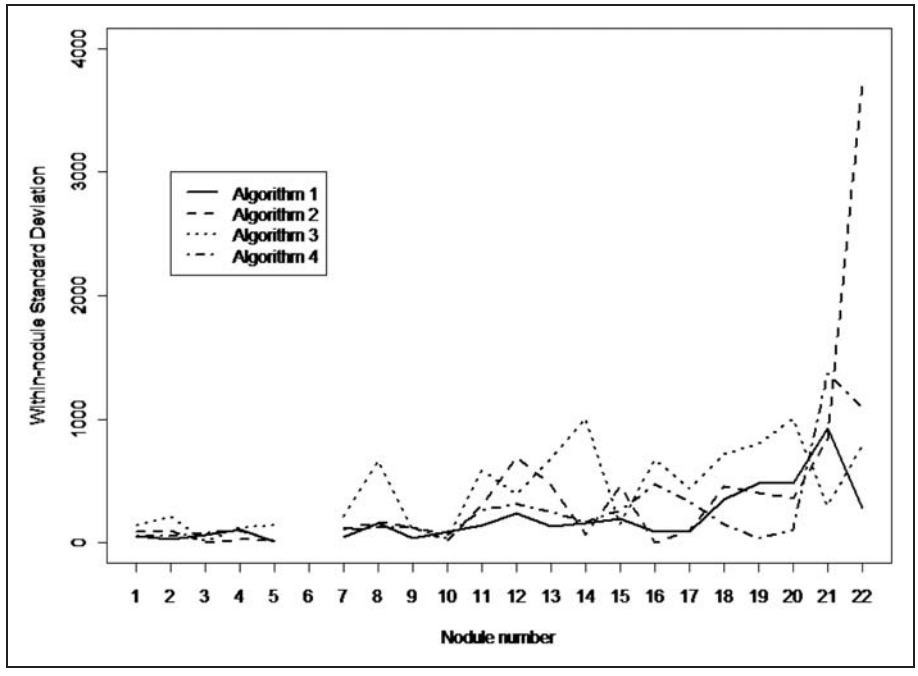

Figure 2. Plot of within-nodule SDs for 22 unique nodules ordered from smallest to largest volume (in mm^3). Note that there is no standard deviation (SD) estimate for nodule #6 because there was only one measurement for this nodule.

Table 6. Estimated bias by nodule size and algorithm.

			Estimated mean bias			
Nominal diameter (mm)	True volume	# obs	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4
8–10	282.35–706.75	36	−10.75	−55.56	156.32	69.02
20	4207.83–5279.07	44	−217.47	−577.08	376.09	−14.62
40	33,781.46–34,389.21	10	373.25	−2370.46	818.52	1377.85
Total		90	−69.14	−567.74	337.34	173.55

Table 7. Adjusted p-values of pairwise comparisons on mean bias.

	Nominal diameter (mm)			
Comparisons	8–10	20	40	Total
Algorithm 1 versus 2	0.9839	0.0544	0.0214	0.0013
Algorithm 1 versus 3	0.0719	0.0001	0.9968	0.0151
Algorithm 1 versus 4	0.7888	0.5942	0.8378	0.3542
Algorithm 2 versus 3	0.0097	<0.0001	0.0054	<0.0001
Algorithm 2 versus 4	0.3138	0.0004	0.0009	<0.0001
Algorithm 3 versus 4	0.7123	0.0283	0.9891	0.7770

20 mm diameter, algorithms 1 and 4 had significantly smaller bias than algorithms 2 and 3. For nodules with 40 mm diameter, algorithms 1, 3, and 4 had significantly smaller bias than algorithm 2. Overall, algorithm 1 seems to perform better in terms of bias across different nodule sizes.

4.3 Comparison of algorithms' reproducibility for measuring tumor volume

Continuing with our second goal for this analysis, we now compare the algorithms' precision. Reproducibility, not repeatability, can be assessed in the QIBA 3a study because the multiple measurements taken on the same nodule come from different images of various slice thicknesses and various placements of the same nodule in the lung (i.e. reproducibility conditions).[6] Similar to bias, the within-nodule reproducibility tends to be worse (within standard deviation larger) for larger nodules, and this precision may not be proportional to the nominal diameter nor to the true volume. Thus, we evaluate the algorithms' reproducibility also within the three groups of nodule size. Again there is substantial precision variation among nodules with similar nominal diameter and this variation may be due to the physical characteristics of the nodule. We averaged the precision estimates across nodules in the same nodule size group by averaging the within-nodule sample variances (rather than SDs). The group within-nodule SD is obtained as the square root of the averaged within-nodule sample variance (Table 8).

A traditional *scaled* index of reproducibility is the ICC[5] (note that the within-nodule SD is an *unscaled* index). For comparison in Table 8 we also present the estimated ICCs by nodule size and algorithm as well as the overall ICCs. Due to the ICC's dependence on the between-nodule variability, the ICC values by the same algorithm are not comparable across the three different

Table 8. Estimated precision (i.e. reproducibility) by nodule size and algorithm.

Nominal diameter (mm)	# nodules	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4
Estimated within-nodule standard deviation (wSD)					
8–10	9	78.886	93.99	257.585	87.72
20	10	276.134	388.36	693.803	263.12
40	2	688.018	2689.16	597.851	1239.04
Total	21	289.930	874.258	540.089	427.174
Reproducibility coefficient (RDC)[a]					
8–10	9	218.514	260.352	713.510	242.984
20	10	764.891	1075.76	1921.83	728.842
40	2	1905.81	7448.97	1656.05	3432.14
Total	21	803.106	2421.69	1496.05	1183.27
Estimated ICC					
8–10 mm	9	0.61	0.33	0.79	0.61
20 mm	10	0.07	0.042	0.47	0.58
40 mm	2	−0.19	0.64	0.66	−0.03
Total	21	0.999	0.999	0.999	0.999

[a]The reproducibility coefficient is introduced in Section 4.5.1. Its calculation for the QIBA 3a data is discussed in detail in Appendix 1.

nodule sizes, although the ICC values by different algorithms within the same nodule size are comparable. The ICC tends to be very small when the estimated between-nodule variability is small. In contrast, when all nodules were combined the overall ICCs are all close to one for all four algorithms due to the relatively large between-nodule variability. Due to the ICC's dependency on between-nodule variability, as well as the fact that the ICCs are not constant across nodule sizes, we do not recommend the ICC for comparisons of algorithms in this example.

Testing equality of the within-nodule SDs is equivalent to testing the equality of the mean within-nodule variances between algorithms. Thus, pairwise comparisons of the precision among the algorithms were performed via GEEs while accounting for the correlation among multiple measurements made on the same nodule by different algorithms. This can be accomplished by SAS procedure GENMOD where the response variable is the observed within-nodule sample variance and the independent variables are group and algorithm with an identity link. The pairwise comparison p-values were obtained through the contrast statement in SAS. Table 9 presents the p-values adjusted for multiple comparisons with Sidak's FWER-controlling procedure.[20]

Using the results in Tables 8 and 9, we might make the following conclusions. For nodules with 8–10 mm diameter, large differences were observed between algorithms 1, 2, and 4 versus 3. With nine nodules we have insufficient power to detect differences, so we would conclude that a larger study, or meta-analysis, is needed. For nodules with 20 mm diameter, algorithms 1, 2, and 4 had significantly better precision than algorithm 3. For nodules with 40 mm diameter, there is again insufficient evidence to claim one algorithm is better than the other, given the small sample size.

4.4 Comparison of algorithms' agreement with true tumor volume

In Sections 4.2 and 4.3 we compared the algorithms' bias and precision separately to determine which algorithm(s) was superior for each metric. However, due to the inherent trade-off between

Table 9. Adjusted p-values of pairwise comparisons of wSDs.

	Nominal diameter			
Comparisons	8–10	20	40	Total
Algorithm 1 versus 2	1.0000	0.9460	0.7715	0.6257
Algorithm 1 versus 3	0.4002	0.0002	1.0000	0.9984
Algorithm 1 versus 4	1.0000	1.0000	1.0000	1.0000
Algorithm 2 versus 3	0.4512	0.0024	0.7601	0.8984
Algorithm 2 versus 4	1.0000	0.9192	0.8674	0.7724
Algorithm 3 versus 4	0.4286	0.0001	1.0000	1.0000

bias and precision, it is often of interest to evaluate the algorithms' performance in terms of agreement with the true tumor volume, where the measure of agreement takes into account both bias and precision.[4] Similar to the first example, in Table 10 agreement is assessed by the unscaled indices of LOAs and TDI with 80 and 95% CP, as well as the CP for acceptable differences of 200, 500, and 1000 mm^3, respectively, by nodule size and algorithm. (We are treating the repeated measurements on the same nodule as measurements from different nodules for illustration.)

Due to the observed bias in Table 6, the LOAs are not symmetric around zero and they tend to shift to the left for algorithm 2 and to the right for algorithm 3. For almost all scenarios, the absolute value of the estimated TDI is smaller than the maximum of the absolute values of the LOAs. To compare the algorithms to determine which one provides measurements closest to truth, we would be interested in the algorithm that provides a large percent, p, of measurements that are within an acceptable distance, d, to the truth (CP_d), or conversely, provides a small distance d between measurement and truth that is achieved with an acceptable probability p (TDI_p). (Note that LOA_p is also useful here, but is asymmetric, in contrast with TDI_p, and thus not as focused on measurements being within a distance from the truth.) Based on the estimated TDI_80% and TDI_95%, algorithm 1 tends to perform better than the other algorithms: algorithm 1's measurements are closer to the truth than the other algorithms for smaller nodules, similar to algorithm 4 and better than algorithms 2 and 3 for medium nodules, and similar to algorithm 3 and better than algorithms 2 and 4 for large nodules.

A plot of the observed absolute difference versus the CP (points are connected by straight lines) is referred to as the CP curve, and it provides a visual examination of the algorithm's performance over the range of possible differences between observed and true volume. The estimates of CP_200, CP_500, and CP_1000 in Table 10 provide three points on the curve. Figure 3 illustrates the estimated CP curves by algorithm and nodule size group. For nodule sizes of 8 and 20 mm diameter, algorithm 1 seems to perform better than the other algorithms, whereas all four algorithms perform similarly for nodule sizes of 40 mm diameter. Overall, algorithm 1 performed better than the other algorithms when all three groups of nodules were combined.

4.5 Comparison of algorithms' performance for measuring tumor volume change

The third goal of our analysis is to estimate and compare the algorithms' performance in measuring tumor volume *change*. In Section 5 we use the VOLCANO study to answer this question directly for the scenario of no real change. Here, we illustrate how to use the individual nodule volume

Table 10. Agreement of measured and true volumes by algorithm.

	LOA_80%	TDI_80% (95% CI)	LOA_95%	TDI_95% (95% CI)	CP_500 (95% CI)	CP_1000 (95% CI)
Diameter = 8 or 10 mm						
Algorithm 1	−126.6, 97.0	120.6 (61.7, 150.2)	−185.9, 156.0	214.9 (125.2, 252.1)	1.0 (NA)	1.0 (NA)
Algorithm 2	−230.8, 111.0	189.3 (96.4, 249.8)	−321.6, 210.8	264.8 (194.7, 359.2)	1.0 (NA)	1.0 (NA)
Algorithm 3	−550.5, 789.8	252.1 (181.6, 1700.8)	−960.5, 1145.9	1730.8 (216.6, 1920.8)	0.92 (0.59, 0.99)	0.92 (0.59, 0.99)
Algorithm 4	−76.0, 240.6	159.2 (129.3, 253.1)	−160.1, 324.7	301.3 (159.2, 303.4)	1.0 (NA)	1.0 (NA)
Diameter = 20 mm						
Algorithm 1	−716.1, 327.8	739.7 (243.4, 881.5)	−993.3, 605.0	1007.2 (401.9, 1163.8)	0.73 (0.44, 0.90)	0.93 (0.67, 0.99)
Algorithm 2	−1342.7, 307.0	1199.1 (527.1, 1587.8)	−1780.9, 745.2	1719.1 (992.7, 2109.1)	0.52 (0.27, 0.76)	0.71 (0.40, 0.89)
Algorithm 3	−555.4, 1283.6	989.7 (780.3, 1417.9)	−1043.9, 1772.1	1630.9 (1109.7, 1890.9)	0.43 (0.35, 0.52)	0.82 (0.65, 0.92)
Algorithm 4	−817.8, 915.9	913.0 (762.2, 989.1)	−1278.4, 1376.4	1008.5 (824.7, 1199.1)	0.34 (0.21, 0.50)	0.93 (0.79, 0.98)
Diameter = 40 mm						
Algorithm 1	−862.0, 1439.7	1428.2*	−1473.4, 2051.1	1454.9*	0.50*	0.70*
Algorithm 2	−8427.0, 2629.7	2389.2*	−11361.5, 5555.2	8999.2*	0.40*	0.50*
Algorithm 3	102.2, 1636.5	938.5*	−302.3, 2044.0	2271.6*	0.20*	0.80*
Algorithm 4	−666.6, 3148.6	2638.5*	−1680.0, 4162.0	3408.5*	0.10*	0.60*

NA. Not applicable to compute 95% CI due to estimated value on the boundary; *95% CI is not computed due to small sample size = 2.

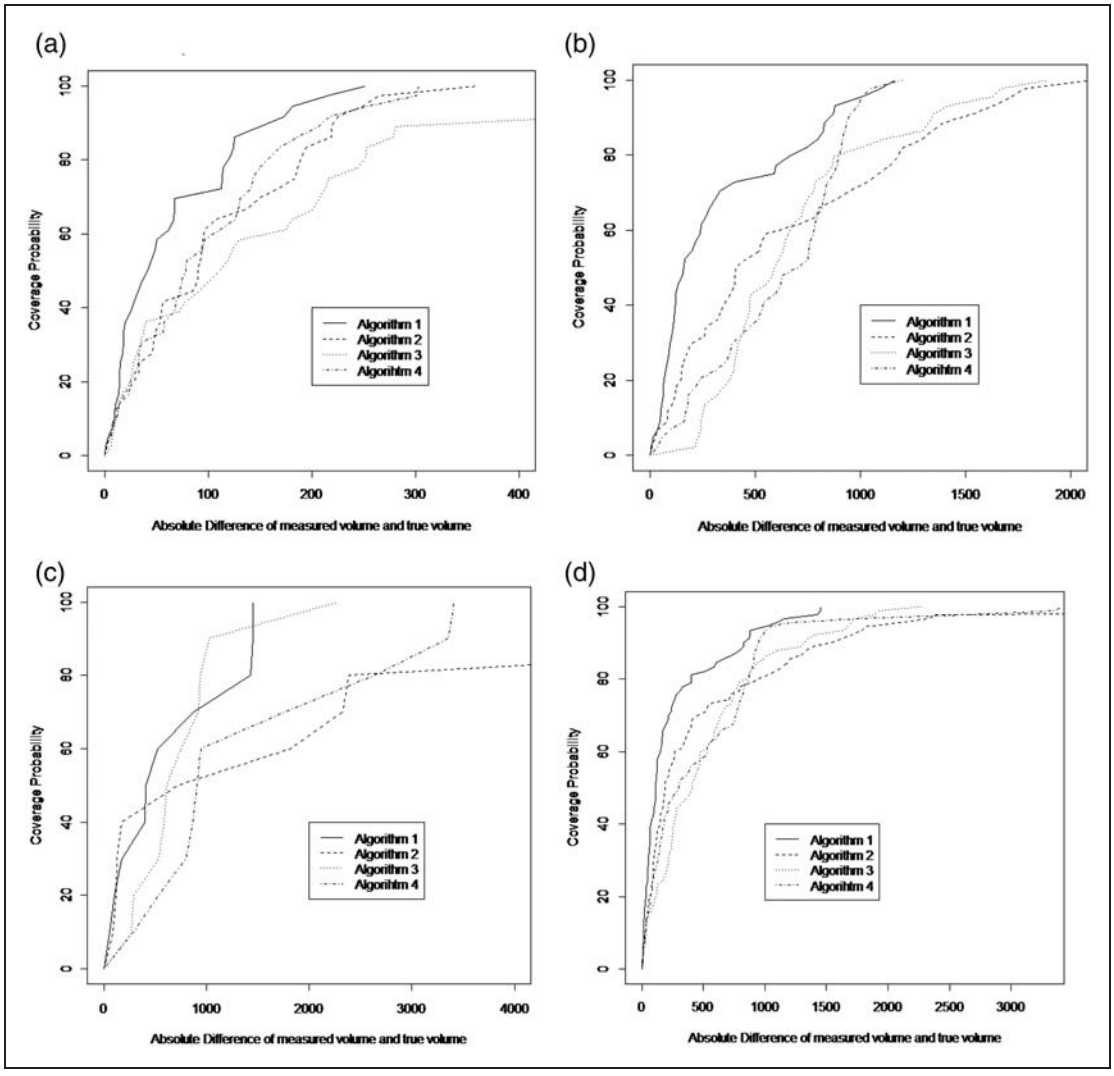

Figure 3. Coverage probability curves by group and algorithm. (a) Group 1: nominal diameter = 8 or 10 mm, (b) Group 2: nominal diameter = 20 mm, (c) Group 3: nominal diameter = 40 mm, and (d) overall.

measurements from the algorithms in the QIBA 3a study to assess the performance of the algorithms for measuring tumor volume change.

4.5.1 Nonparametric estimation

The first step is to estimate the algorithms' precision for measuring tumor volume change. There are several approaches to estimating the precision. A simple approach is to use the nonparametric estimates of within-nodule SD from Table 8. Note that the four algorithms' within-nodule SD estimates increase as the size of the tumor increases. We apply equation (2) for estimating the precision of the change in tumor volume since it does not require the assumption that the SD is constant over the range of lesion sizes. Note that we cannot use equation (3) since we do not have an

Table 11. Estimates of SD of change in tumor volume (mm^3).

Nominal diameter (mm) at two time points[a]	Upper bound on SD of the change in tumor volume			
	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4
8–10 to 20	287.18	399.57	740.08	277.36
8–10 to 40	692.53	2690.80	650.98	1242.14
20 to 40	741.36	2717.06	915.85	1266.67

[a]The mean tumor volume for tumors 8–10 mm was 535 mm^3, for 20 mm tumors was 4510 mm^3, and for 40 mm tumors was 34,085 mm^3.

estimate of the within-nodule correlation; thus, our estimates from equation (2) represent an upper bound on the SD. Table 11 summarizes the estimated SD of the change in tumor volume as a function of the lesion size at two time points.

Note that for equation (2) to be useful for detecting true change in tumor volume, the algorithms' measurements of tumor volume need to be a linear function of the true tumor volume (i.e. linearity assumption[4,6]). We can assess the validity of this assumption by inspection of the algorithms' bias, which should be linear (Table 6). From Table 6 we note the assumption of linearity is reasonable for algorithms 2 and 3, i.e. it is plausible that measurements by algorithms 2 and 3 are consistent with the relationship: $Y(t) = a + bX(t) + \varepsilon(t)$, where $Y(t)$ is the measurement by an algorithm at time t, $X(t)$ is the true size of the tumor volume at time t, and $\varepsilon(t)$ is the random error at time t. When linearity is present, we can claim that a change based on Y is equivalent to a change in true volume.

Consider several examples of assessing tumor change. Suppose at a first time point, we estimate a tumor's volume with algorithm 2 to be 507 mm^3 and then at a second time point we estimate the same tumor to have a volume of 285 m^3. We want to assess whether this change represents a true change in the tumor's volume or whether this change is due to measurement error. From Table 6, Table 8, and equation (2), the SD of the difference is estimated to be $\sqrt{2 \times 93.99^2} = 132.9$. Thus, we compute a z-value: (507–285)/132.9 = 1.67, and comparing this to a standard normal distribution, we interpret it as: there is a 10% chance of observing a difference in the tumor's volume as large as we observed given it is just due to measurement error.

As a second example, from Table 10 we estimate that we would be nearly 100% confident that a tumor change from 535 to 4510 mm^3 represents a true change in the tumor volume (i.e. SD of difference = $\sqrt{93.99^2 + 388.36^2} = 399.57$, z-value = (4510–535)/399.57 = 9.9). If it is believed that a tumor has changed in volume, clinicians often need to know the magnitude of the change. We can construct a 95% CI for the measured changed; for the second example this interval is $(4510 - 535) \pm 1.96 \times 399.57$ or [3192, 4758]. Note that this is the CI for the *algorithm's measured change* in the tumor's volume. To construct a CI for the *true change* in the tumor's volume, the linearity assumption must hold and we must know the value of the coefficient b in the relationship: $Y(t) = a + bX(t) + \varepsilon(t)$.[4] From the QIBA 3a study the point estimate of b is 0.93 from a simple least-squares linear regression; however, we recognize that a better approach to estimating b is through a meta-analysis of existing literature.[21] For illustration purposes, let $b = 0.93$. Then a 95% CI for the true change in the tumor's volume is $(4510 - 535)/b \pm 1.96 \times 399.57/b^2$ or [3369, 5180].

Note that for algorithms 1 and 4 the assumption of linearity is not consistent with the estimates of bias in Table 6. Thus, we cannot assume that an estimate of change in Y is equivalent to the same magnitude of change in true volume.

4.5.2 Parametric modeling

Instead of using the nonparametric estimates of the within-nodule SD from Table 8, we can use statistical models to estimate the reproducibility of the algorithms, thus answering the question of whether or not we are 95% confident that a true change has occurred.

For a new lesion, the *reproducibility interval* (RI) can be defined as the interval within which 95% of differences between two repeated measurements are expected to fall when the measurements are taken under *different* conditions. In contrast, *repeatability* refers to closeness of agreement between repeated measurements taken under the *same* conditions.[5,6]

A condition that changes over repeated measurements may be considered to have either random or fixed effects on the measurement. If the conditions under study are all considered to have random effects, then typically the RI takes the symmetrical form $(-RDC, RDC)$, where RDC is called the *reproducibility coefficient*, the least significant difference between two measurements taken under different conditions.[5] If some of the conditions under study are considered to have fixed effects, then the RI can be asymmetric, as we will show below for the QIBA 3a study. Note that the RI is specific to the set of conditions under study. These conditions need to be stated along with the reporting of any RI. Note that conditions other than those under study may be present in any given clinical situation and may change over repeated measurements.

For the QIBA 3a study, let Y_{itk} denote the kth volume measurement by an algorithm on nodule i with slice thickness t, where $i = 1, 2, \ldots n$, $t = 0.8$ or 5 mm, and $k = 1, 2, \ldots, K_{it}$. Consider the mixed effects model

$$Y_{itk} = \mu + \gamma_i + \tau_t + (\gamma\tau)_{it} + \varepsilon_{itk} \tag{4}$$

with overall mean μ, random effect $\gamma_i \sim N(0, \sigma^2_\gamma)$ for lesion i, fixed effect τ_t for slice thickness t, random lesion by slice-thickness interaction effect $(\gamma\tau)_{it} \sim N(0, \sigma^2_{\gamma\tau})$, and random effect $\varepsilon_{itk} \sim N(0, \sigma^2_\varepsilon)$ for measurement k within lesion i and slice thickness t. For identifiability, set $\tau_{0.8} = 0$.

From equation (4), the difference $d_{new} = Y_{i_{new}tk} - Y_{i_{new}t'k'}$ between two measurements $Y_{i_{new}tk}$ and $Y_{i_{new}t'k'}$ taken on new lesion i_{new} but with different slice thicknesses $t = 5$ mm and $t' = 0.8$ mm, respectively, has mean $\delta = \tau_5 - \tau_{0.8} = \tau_5$ and variance equal to $Vd_{new} = 2(\sigma^2_{\gamma\delta} + \sigma^2_\varepsilon)$. Let $\hat{\delta}$ be the estimate of δ with variance $V\hat{\delta}$. Considering that $d_{new} - \hat{\delta}$ has mean 0 and variance $Vd_{new} + V\hat{\delta}$ by independence of d_{new} and $\hat{\delta}$, the RI is

$$RI = \hat{\delta} \pm 1.96\sqrt{Vd_{new} + V\hat{\delta}} \tag{5}$$

Note this RI is asymmetrical, depending on the sign $\hat{\delta}$, the estimate of the fixed effect for slice thickness. For the QIBA 3a data, the slice thickness effect (Table 12) and the RI (Tables 13 to 16) of the algorithms are compared for each of three groups of nominal lesion diameters 8–10, 20, and 40 mm. To estimate $\hat{\delta}$, $V\hat{\delta}$, and Vd_{new}, we used SAS Proc Mixed to obtain restricted maximum likelihood (REML) estimates and substituted these into the formula for Vd_{new} and equation (5).

For a given group of nodule diameters, the sign and magnitude of the slice thickness effect varied by algorithm (Table 12). For algorithm 4, the slice thickness effect for two of the groups was statistically significant.

For lesions with diameters 8–10 mm, RI for algorithm 1 volume measurements was $(-248, 294)\,\text{mm}^3$, the interval within which 95% of differences are expected to lie between a measurement taken with slice thicknesses 5 mm and a measurement taken with slice thickness 0.8 mm (Table 13). A difference below $-348\,\text{mm}^3$ or above $294\,\text{mm}^3$ can be said to a significant difference in the measured volume. To the extent that the measured volume is linear in the true

Table 12. Fixed slice thickness effect REML estimates and p-values.

	Lesion size (mm)					
	8–10 mm (n = 36)		20 mm (n = 44)		40 mm (n = 10)	
Alg	$\hat{\delta}$	p-value	$\hat{\delta}$	p-value	$\hat{\delta}$	p-value
1	23.09	0.6279	−189.71	0.0896	806.36	0.3900
2	11.51	0.7886	−40.31	0.7650	−1929.51	0.5727
3	15.68	0.8996	−100.85	0.6635	78.75	0.9227
4	158.25	0.0033	279.36	0.0089	2227.69	0.0909

Table 13. Reproducibility interval *RI* (mm^3), nodules with nominal diameters 8–10 mm, average true volume 524.95 mm^3 (n = 10 nodules, 36 total volume measurements per algorithm).

	REML estimates								
			Variance components[a]						
Alg	Mean	$\hat{\delta}$	Les	Thc × Les	Error	SE $\hat{\delta}$	SD d_{new}	RI	RI as a % of average true volume
1	520.91	23.08	0	3633	4977	44.05	131.23	−248, 294	−47.3, 56.1
2	472.52	11.51	4807	440.3	10,498	40.14	147.9	−289, 312	−55.0, 59.4
3	666.02	15.68	245,832	0	71,960	106.01	379.4	−756, 788	−144.1, 150.1
4	642.1	158.25	12,886	400.39	3074.49	25.26	83.4	−13, 329	−2.4, 62.7

[a]Variance components for lesion (Les), slice thickness by lesion interaction (Thc*Les), and error.

Table 14. Reproducibility interval *RI* (mm^3), nodules with nominal diameters 20 mm, average true volume 4567.3 mm^3 (n = 10 nodules, 44 total volume measurements per algorithm).

	REML estimates								
			Variance components[a]						
Alg	Mean	$\hat{\delta}$	Les	Thc × Les	Error	SE $\hat{\delta}$	SD d_{new}	RI	RI as a % of average true volume
1	4341.9	−189.71	0	11,607	76,610	98.24	420.04	−1035, 656	−22.7, 14.4
2	3993.7	−40.31	0	18,527	139,594	130.35	562.35	−1172, 1091	−25.7, 23.9
3	4893.1	−100.85	407,034	0	510,165	223.30	1010.11	−2128, 1927	−46.6, 42.2
4	4563.1	279.36	72,993	18,385	24,126	81.28	291.59	−314, 873	−6.9, 19.1

[a]Variance components for lesion (Les), slice thickness by lesion interaction (Thc*Les), and error.

volume, a significant difference may indicate a difference in the true volume. As a percentage of the average true volume 524.95 mm^3 for these lesions, the RI interval is (−47.3, 56.1%).

Suppose linearity with unit slope can be assumed within each slice thickness, that is $E(Y_{it}) = \mu + \tau_t = \alpha_t + X_{it}$, where X_{it} is the true volume for nodule i on the occasion when slice

Table 15. Reproducibility interval *RI* (mm^3), lesions with nominal diameters 40 mm, average true volume 34,024.56 mm^3 (n = 2 nodules, 10 total volume measurements per algorithm).

	REML estimates								
			Variance components[a]						
Alg	Mean	$\hat{\delta}$	Les*	Thc × Les	Error	SE $\hat{\delta}$	SD d_{new}	RI	RI as a % of average true volume
1	34,345	806.36	0	21,969	216,611	566.68	934.43	−1336, 2948	−3.9, 8.7
2	31,657	−1929.51	4,834,456	352,263	4,998,706	2429.78	4135.45	−11,330, 7471	−33.3, 22.0
3	34,877	78.75	290,892	294,496	254,897	645.14	1048.23	−2334, 2491	−6.9, 7.3
4	35,055	2227.69	822,866	0	231,642	320.27	680.65	753, 3702	2.2, 10.9

[a]Variance components for lesion (Les), slice thickness by lesion interaction (Thc*Les), and error.

thickness t is used to take the measurement. In the QIBA 3a study, the expected difference in measured volume $d_i = Y_{it} - Y_{it'}$ for nodule i using slice thicknesses t and t' is $\delta = \alpha_t - \alpha_{t'}$ ($= \tau_t$, the slice thickness effect) because by design true volume is unchanged. However, in practice, for new lesion i_{new} volume may have changed, thus $E(d_{new}) = \alpha_t - \alpha_{t'} + X_{i_{new}t} - X_{i_{new}t'}$. Therefore $d_{new} - \hat{\delta}$ has mean $X_{i_{new}t} - X_{i_{new}t'}$, the true change in volume, and variance $Vd_{new} + V\hat{\delta}$. Thus, a 95% CI on the true change in volume is

$$d_{new} - \hat{\delta} \pm 1.96(Vd_{new} + V\hat{\delta})^{1/2}$$

The estimate $d_{new} - \hat{\delta}$ of true change adjusts for $\hat{\delta}$, which estimates $\alpha_t - \alpha_{t'}$, the difference in the intercepts in the linearity equations for the two slice thicknesses.

For example, for algorithm 2, suppose a tumor has an initial estimated volume of 508 mm^3 using 0.8 mm slice thickness. At a second occasion, suppose the estimate using 5 mm slice thickness was 226 mm^3. Adjusting for fixed slice thickness effect of 11.51 with standard error of 40.14 (Table 13), the estimated true change in volume is $(226-508) - 11.51 = -293.51$ with 95% CI $-293.51 \pm 1.96 * (147.9^2 + 40.14^2)^{1/2} = (-593.9, 6.9)$.

In the QIBA 3a study, slice thickness was controlled to be either 0.8 or 5 mm. Thus, slice thickness was modeled as having fixed effects on the algorithm measurements. For other studies, the condition(s) being varied may be regarded as having random effects, e.g. day, instrument, operator, clinical site, reader, etc. For a model of random effects for slice thickness that corresponds to the model in equation (4) and for the calculation of *RI* under this model, see Appendix 1. The *RDC* results are provided in Table 8 for comparison within its nonmodeling counterpart. In Appendix 1, *RDC* is shown to be greater than a similar nonmodeling counterpart because *RDC* includes expressly the variation due to the level of a condition being changed between two measurements.

5 Volcano '09 project

The mission of the VOLCANO[23] is to advance the state of the art in quantitative lesion *change* evaluation from image data with an initial focus on pulmonary nodules in CT scans. A major goal of the VOLCANO program is to provide a public set of reference image datasets with extensive documentation. These data sets may be used to benchmark the performance of new image

analysis methods. Further, VOLCANO will provide a standardized comparative analysis to other methods. To date, two VOLCANO studies have been conducted that establish image documentation: VOLCAMAN '09 which evaluates a set of different computer algorithms on a change-in-size measurement task and VOLCAMAN'10 which evaluates experienced radiologists on the same task. A subset of the former study was available for analysis in this paper.

The VOLCANO'09 Challenge was part of the Second International Workshop on Pulmonary Image Analysis held in conjunction with MICCAI 2009. The data for this competition was prepared for inclusion in the Public Lung Database To Address Drug Response and was provided by the Weill Medical College of Cornell University. The dataset consisted of several subgroups: (a) zero-change cases, (b) zero-change cases with different slice thickness scans, (c) cases with actual size change, and (d) a synthetic nodule case. Both the zero-change cases have images acquired within a few minutes and therefore there is no change in the actual lesion size. The (b) subset, zero-change with a slice difference, was selected to evaluate the bias introduced by a slice thickness change between scanner protocols for the two images. The real change cases were to explore changes in size other than zero although we do not have the true value for these cases. The single phantom case was included to test if the behavior on the phantom data was comparable to the performance on real lesions.

For response to therapy the appropriate measurand has been identified as the proportional change in size rather than the absolute size as, for example, the recommendation first of the World Health Organization in 1981[24] then of RECIST.[25] Similarly, for diagnosis of cancer, the appropriate measurand is the growth rate, which is derived from the proportional change in size over time.[9] Further, following the tradition of RECIST, the proportional change is made relative to the baseline or first size observation; i.e. given that a lesion starts with a volume size measurement of V1 and at a later time has a volume V2, a typical measurand is $((V2 - V1)/V1)$. This is the measurand used in the VOLCANO study.

In the challenge instructions, participants were invited to download the CT images and to complete a spreadsheet in which the change in size measurement for the nodule in each image pair was recorded. Responses from 13 teams from both academia and industry were received with size change results for a total of 17 different methods. Of the 17 computer algorithm methods evaluated only one did not require any human participation in the measurement process. For the purposes of this paper, we have nine cases (i.e. nine nodules from nine patients) in which the amount of actual change is unknown and nine cases (i.e. nine nodules from nine different patients) from a test–retest design where the subjects were re-measured after an interval of a few minutes and therefore there is no real change in the nodule size. The change measurements from six different algorithms were provided for each nodule.

Figure 4 illustrates an example of the study images. The first and middle columns illustrate image slices through the nodule, while the last column represents a three-dimensional representation of the nodule generated by one of the algorithms.

5.1 Assessment of bias in volume change measurements

In this analysis we first characterize the performance of the algorithms for measuring the change in tumor volume, then we test the null hypothesis that the agreement with the true value (i.e. no change) is equivalent across the algorithms. The assessment of agreement with no change is based on data from the nine cases with no true volume change. Figures 5 and 6 illustrate the proportional change measurement by nodule and algorithm, respectively. The estimated change is quite large for nodule #3 compared with the other nodules (see Figure 5). For this nodule the estimated change is 0.09, whereas the estimated mean change (across nodules) is 0.01. The estimated mean change ranges

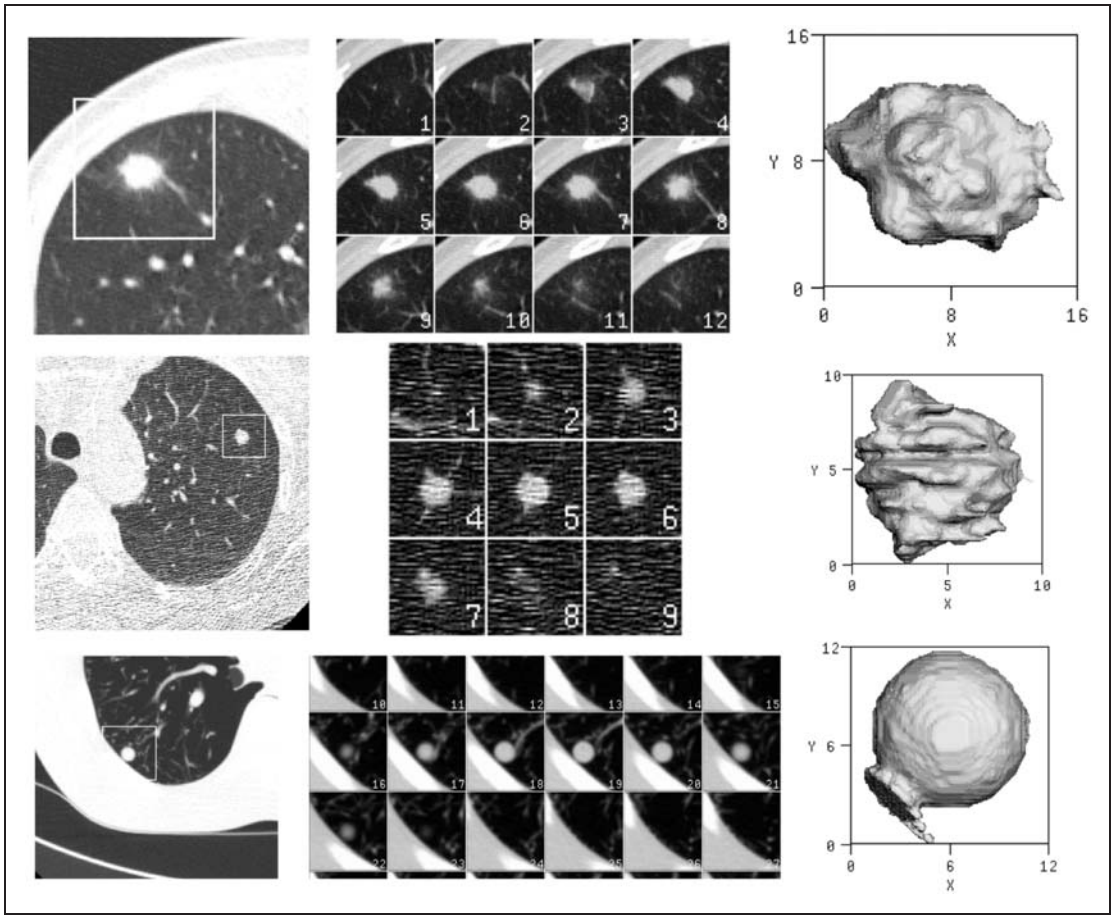

Figure 4. Images of three moderate-sized nodules from the Volcano study: left column, a central slice through the nodule; central column, a montage of all the images slices that contain the nodule; right column, a 3D visualization of an algorithm segmentation. First and second rows show real nodules, the third row shows a spherical phantom nodule within an anthropomorphic phantom.

from 0.006 for Algorithm D to 0.02 for Algorithm E (Figure 6). There does not appear to be a relationship between baseline nodule size and measured change (Figure 5). In a simple regression of measured change on baseline lesion size, neither slope nor intercept was significantly different from zero for any of the algorithms, suggesting that the measured change is not linearly related to the lesion size and that the mean change does not differ from zero, respectively.[5] This finding is expected since the change measurement is a proportional change measurand.

Figures 5 and 6 also indicate that the differences in measurements between algorithms are small relative to the between-subject variability. In a two-way ANOVA with nodule and algorithm as two fixed factors, there was no significant algorithm effect ($p = 0.789$; see Table 16). The nodule effect was significant ($p < 0.01$), however, suggesting that the noise introduced by performing the scan at two different times varied between nodules. Similarly, we ranked the absolute value of the proportional change measurements, then applied Friedman's nonparametric test with nodule being the blocking

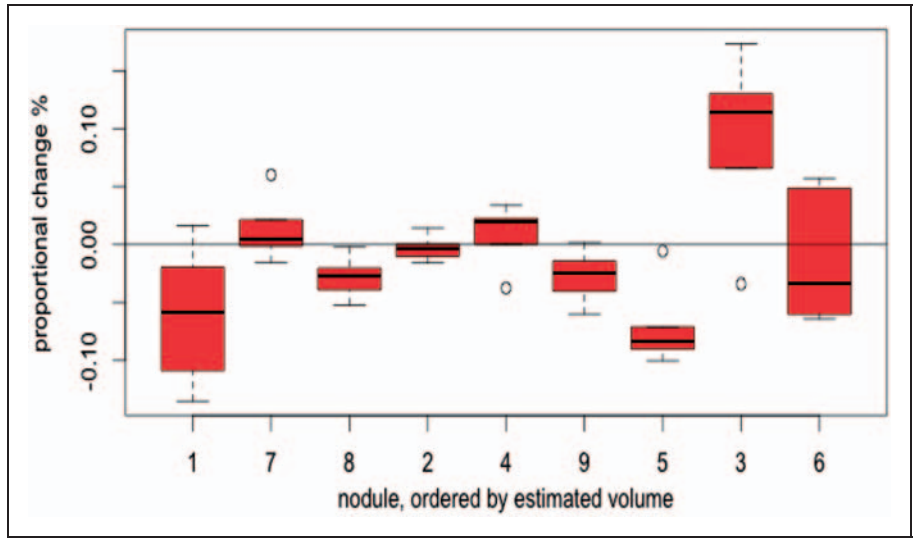

Figure 5. Box-plots of the mean proportional change of the six algorithms for each nodule, ordered by estimated volume. The true change is zero.

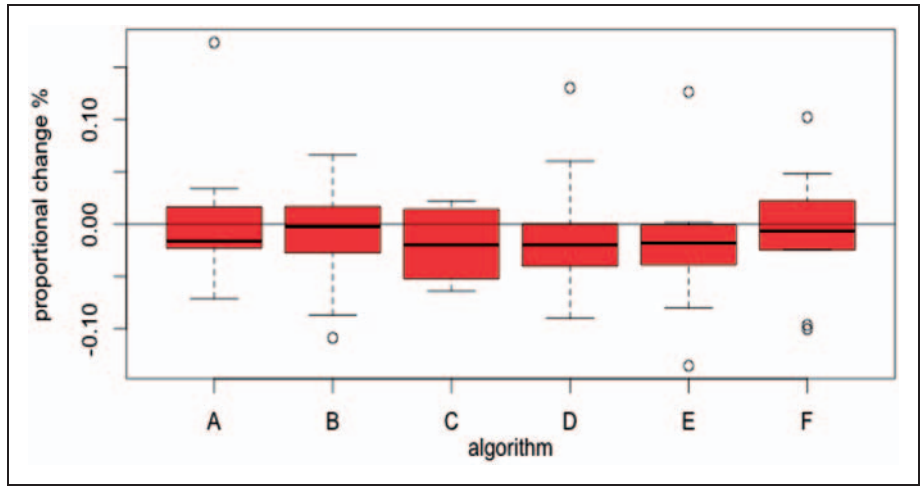

Figure 6. Box-plots of the mean proportional change of the nine nodules for each algorithm. The true change is zero.

variable. We tested the null hypothesis that the algorithms are similar with respect to their agreement with zero change versus the alternative hypothesis that one or more algorithms differ. The p-value was 0.909, indicating no evidence that the algorithms differ. The open-source R software package for two-way ANOVA and Friedman's rank sum test was used for the analyses.

The LOAs[19] of the algorithms with a hypothesis of no change in tumor volume can be calculated in order to assess the distribution of the difference between a measured change in tumor volume and

Table 16. Estimates of mean change and associated 95% CIs.

	Estimate	−CI	+CI	p-value*
Algorithm A	0.006	−0.036	0.047	0.784
Algorithm B	−0.011	−0.052	0.031	0.599
Algorithm C	−0.020	−0.061	0.022	0.340
Algorithm D	−0.006	−0.047	0.036	0.789
Algorithm E	−0.022	−0.063	0.020	0.292
Algorithm F	−0.008	−0.049	0.034	0.708

*p-value associated with a t-test of the null hypothesis that the mean change is zero.

a change of zero for a randomly chosen nodule, where the distribution is across nodules (i.e. the second kind of LOAs from Section 3.3). The LOAs are illustrated in Figure 7 for the six algorithms. Interestingly, Algorithm C, with the second highest mean change but smallest between-patient variability, has the narrowest LOA.

The second goal of the analysis is to test algorithm performance against a specified performance claim. If there is no agreed upon performance criterion, we might consider a range of values. Figure 8 illustrates the CP plot of the six algorithms using the parametric estimates of the CP.[5,26] The proportion of nodules with bias less than an agreement criterion varying from 0.01 to 0.1 is plotted for each algorithm. Algorithm C offers the best CP. On the other hand, we may decide a priori that a 30% change in tumor volume is clinically relevant, and we would like to compare the algorithms' ability to detect this change. For the nine test–retest cases where there is no change, the proportion of cases with a change $<30\%$ is an estimate of the specificity. For this small dataset, none of the six algorithms reported a change exceeding 30% (estimated specificity of 1.0 with lower 95% confidence bound of 0.67).[27]

5.2 Assessment of agreement

The last goal for this analysis is to measure the agreement among the algorithms for measuring change in tumor volume and determine which algorithms agree best. The assessment of agreement is based on the pooled results of the nine cases where the change in tumor volume is known to be zero and the nine cases where the change in tumor volume is unknown. For illustration purposes, we use the total deviance index here to describe the agreement among the algorithms. We first set the predetermined boundary for the proportion, π_0, to represent, say, 90% of the differences. Pooling over all tumors and algorithm comparisons, 90% of the absolute differences between the algorithms are less than 0.1639 (16.39%). Table 17 summarizes the estimates of TDI_90% for the pairwise comparisons between algorithms. Algorithm D tends to have the highest TDI_90% values, suggesting that the proportional change measurements from algorithm D differ the most from the other algorithms' measurements.

6 Discussion

In order to apply a quantitative imaging algorithm in practice, clinicians need to know the error in their measurements. If the task is to measure the absolute volume of a pulmonary nodule and if a phantom accurately represents these nodules in image presentation, then the phantom studies described in this paper provide direct information on the desired measurement error. In the case

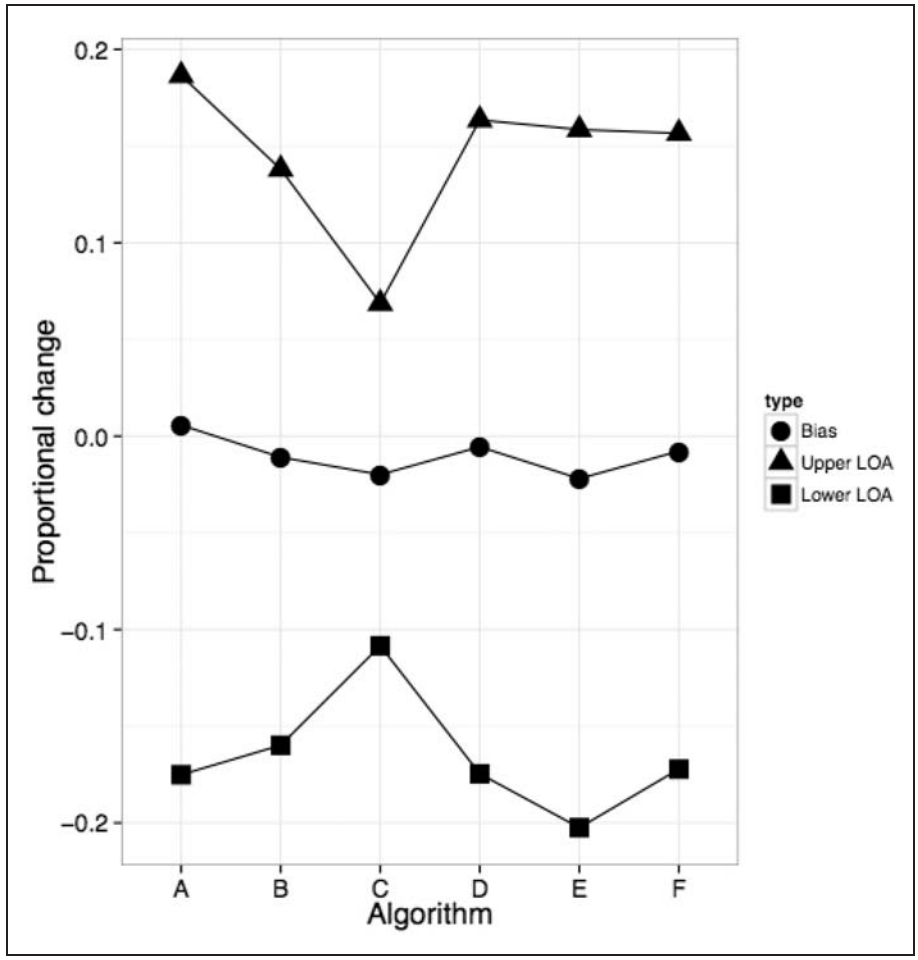

Figure 7. Limits of agreement.

of pulmonary nodule measurement, however, an important clinical task is *size-change* measurement. Actual nodules have a much wider range of presentations than can be represented by phantoms; thus, it is difficult to generalize results on size-change measurements from simulated nodules to actual nodules.

In the VOLCANO study, real nodule image data are involved and the size-change issue is directly addressed. Measurement reproducibility may be determined, as well as the agreement between different measurement methods. However, without the true value it is not directly possible to estimate the measurement error of the size-change measurements. The only time the true value is known is for zero-change cases for which precision and bias may be determined but only for the measurement value of zero. For real, nonzero change cases only the degree of agreement between different algorithms may be determined. Also, the practical small sample size limitation constrains the range of image presentations that can be evaluated.

For the volumetric measurement of pulmonary nodules the large range of nodule size is a further complication for analysis. Algorithms typically have an optimal range at intermediate sized nodules

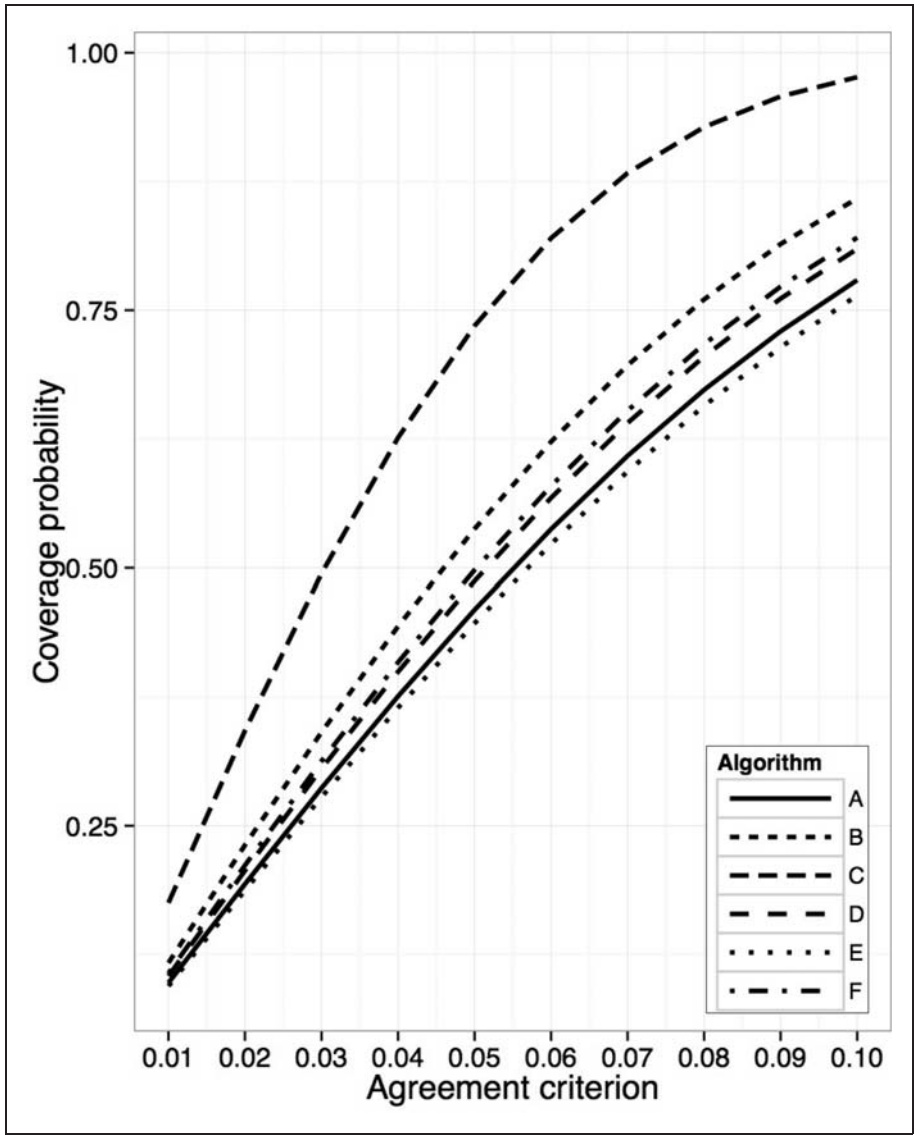

Figure 8. Coverage probability.

Table 17. TDI_90% values.

	Alg B	Alg C	Alg D	Alg E	Alg F
Alg A	0.131	0.207	0.253	0.168	0.172
Alg B		0.125	0.186	0.082	0.103
Alg C			0.164	0.190	0.170
Alg D				0.350	0.266
Alg E					0.101

where the measurement error is low. Higher error is experienced both at the lower end of the measurement range where noise and partial pixel effects are prominent and also at the higher end of the range where the nodule has a more complex shape and interacts with other lung structures. It is important that one never extrapolate results across lesion sizes.

Collaborative efforts are necessary to characterize imaging test performance with respect to specified intended clinical uses, given that the sources of variability are not all understood or no single study design suffices. Such collaboration requires a common understanding of the critical performance factors and a consistent approach to expressing them. It is important to remember that quantifying the performance of a biomarker—determining its fitness for a given purpose—is its own scientific undertaking. This paper provides examples of how the quantitative results from images can be evaluated to establish the technical characteristics of these measures when used as noninvasive biomarkers. The results are highly informative to QIBA in identifying performance characteristics and setting target levels, as well as providing insight for each participant into the specifics of how their method performs relative to their peers. Efforts such as the 3A and Volcano challenges set in place a theoretical framework for how to assess and compare algorithms.

Acknowledgement

The authors acknowledge and appreciate the Radiologic Society of North America (RSNA) and NIH/NIBIB contract # HHSN268201000050C for supporting two workshops and numerous conference calls for the authors' Working Group. The authors appreciate the expert advice of Tatiyana V. Apanasovich on earlier drafts of this paper.

References

1. Radiological Society of North America (RSNA). Quantitative Imaging Biomarkers Alliance (QIBA). Quantitative imaging and imaging biomarkers. http://www.rsna.org/research/qiba.cfm, 2010.
2. Buckler AJ. *Quantitative Imaging Biomarker Alliance on volumetric CT (presentation)*. Medical Imaging Continuum: Path Forward for Advancing the Uses of Medical Imaging in the Development of New Biopharmaceutical Products. Bethesda, MD: DIA, 2–3 October 2008.
3. Buckler AJ, Bresolin L, Dunnick NR, et al. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. *Radiology* 2011; **258**: 906–914.
4. Performance Working Group. Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment. Submitted to *SMMR*.
5. Algorithm Comparison Working Group. Quantitative imaging biomarkers: A review of Statistical methods for computer algorithm comparisons. Submitted to *SMMR*.
6. QIBA Metrology Working Group. The emerging science of quantitative imaging biomarkers: Terminology and definitions for scientific studies and for regulatory submissions. Submitted to *SMMR*.
7. Warfield SK, Zou KH and Wells WM. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004; **23**: 903–921.
8. Reeves AP, Biancardi AM, Apanasovich TV, et al. The lung image database consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements. *Acad Radiol* 2007; **14**: 1475–1485.
9. Kostis WJ, Yankelevitz DF, Reeves AP, et al. Small pulmonary nodules: Reproducibility of three-dimensional volumetric measurement and estimation of time to follow-up CT. *Radiology* 2004; **231**: 446–452.
10. Kostis WJ, Reeves AP, Yankelevitz DF, et al. Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. *IEEE Trans Med Imaging* 2003; **22**: 1259–1274.
11. Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ, et al. Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. *Radiology* 2008; **248**: 625–631.
12. Browder WA, Reeves AP, Apanasovich T, et al. Automated volumetric segmentation method for growth consistency of nonsolid pulmonary nodules in high-resolution CT. In: *SPIE international symposium on medical imaging*, 2007, pp.65140Y-1.
13. Zhang L, Yankelevitz DF, Henschke CI, et al. Zone of transition: A potential source of error in tumor volume estimation. *Radiology* 2010; **256**: 633–639.
14. Gietema HA, Wang Y, Xu D, et al. Pulmonary nodules detected at lung cancer screening: Interobserver variability of semiautomated volume measurements. *Radiology* 2006; **241**: 251–257.
15. Taylor JR. *An introduction to error analysis: The study of uncertainties in physical measurements*, 2nd ed. Sausalito, CA: University Science Books, 1997.
16. Chen B, Barnhart H, Richard S, et al. Quantitative CT: Technique dependence of volume estimation on pulmonary nodules. *Phys Med Biol* 2012; **57**: 1335–1348.
17. Norton NC, Bieler GS, Ennett ST, et al. Analysis of prevention program effectiveness with clustered data using generalized estimating equations. *J Consult Clin Psychol* 1996; **64**: 919–926.

18. Sherman M and le Cessie S. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear model. *Commun Stat Simulat* 1997; **26**: 901–925.
19. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–160.
20. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc* 1967; **62**: 626–633.
21. Huang EP, Wang X-F, Roy Choudhury K, et al. Meta-analysis of the technical performance of an imaging assay: Guidelines and statistical methodology. Submitted to *SMMR*.
22. Quan H and Shih WJ. Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics* 1996; **52**: 1195–1203.
23. Reeves AP, Jirapatnakul AC, Biancardi AM, et al. The VOLCANO'09 challenge: Preliminary results. In: *Second international workshop of pulmonary image analysis*, September 2009, pp.353–364.
24. WHO. *Handbook for reporting results of cancer treatment*. Geneva, Switzerland: World Health Organisation, 1979, p.48.
25. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 2000; **92**: 205–216.
26. Lin L, Hedayat AS, Sinha B, et al. Statistical methods in assessing agreement: Models, issues, and tools. *J Am Stat Assoc* 2002; **97**: 257–270.
27. Hanley JA. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* 1983; **249**: 1743–1745.

Appendix 1: Derivations in Section 3

For the first kind of LOAs, we are interested in knowing the limits within which 95% of such replicated differences fall for a given nodule size j and slice thickness s. Assuming such limits are the same for all nodules for a given nodule size j and slice thickness s, then the LOAs can be estimated as $\widehat{LOAs}_{95\%js}(replications) = \bar{D}_{js} \pm 1.96\widehat{wSD}_{js}$. Specifically, for a given nodule i of a given size j and given slice thickness s, let Y_{ijsk} be the kth replications, X_{ijs} be the true volume, and D_{ijs} be the expected deviation of replications from true volume. Then conditional on i, j, s, we have $Y_{ijsk} - X_{ijs} = D_{ijs} + \varepsilon_{ijsk}$ where $D_{ijs} = E(Y_{ijsk}|ijs) - X_{ijs}$. We are interested in estimating the limits of agreement, $LOAs_{95\%}(replications)$, such that 95% of the replicated differences, $Y_{ijsk} - X_{ijs}$, are within these limits for given nodule i of size j and slice thickness s. Assume that the deviation is the same across all nodules for a given size and slice thickness, i.e. $D_{ijs} = D_{js}$ and the random error $\varepsilon_{ijsk}|ijs$ has the same normal distribution with mean 0 and variance wSD_{js}^2 for all nodules i of a given size j and slice thickness s. Then the 95% LOA is $D_{js} \pm 1.96wSD_{js}$ which can be estimated as $\widehat{LOAs}_{95\%js}(replications) = \bar{D}_{js} \pm 1.96\widehat{wSD}_{js}$.

For the second kind of LOAs, we are interested in knowing the limits within which 95% of differences of measurements and true values for a random sample of nodules fall. Such 95% LOAs is $LOAs_{95\%js}(nodules) = D_{js} \pm 1.96 * SD_{js}(measured - truth)$, where $SD_{js}(measured - truth) = \sqrt{wSD_{js}^2 + \sigma_{Njs}^2 + \sigma_{Tjs}^2}$ that is contributed from variations from within (wSD_{js}^2) and between (σ_{Njs}^2) nodules as well as variation of true volumes (σ_{Tjs}^2) across nodules and the mean bias (D_{js}). The derivation is as follows. Let Y_{ijs} be ONE measurement of volume and X_{ijs} be the true volume for nodule i of size j and slice thickness s. We are interested in the 95% LOAs such that 95% of the differences from all nodules of a given size j and slice thickness s, $Y_{ijs} - X_{ijs}$, are within these limits. Assume that $Y_{ijs} - X_{ijs}|js$ has a normal distribution with mean B_{js} and standard deviation, $SD_{js}(measured - truth)$, then LOAs are $LOAs_{95\%js}(nodules) = D_{js} \pm 1.96 * SD_{js}(measured - truth)$. Since we have K replicated measurements, Y_{ijsk}, rather than only one measurement of volume for a nodule, we would need to derive $SD_{js}(measured - truth)$ for the situation with replications. Let $Y_{ijsk} - X_{ijs} = \mu_{ijs} - X_{ijs} + \varepsilon_{ijsk}$ for given j and s with $E(Y_{ijsk}|js) = E(\mu_{ijs}|js) = \mu_{js}$, $E(X_{ijs}|js) = X_{js}$, $D_{js} = \mu_{js} - X_{js}$, where ε_{ijsk} is assumed to be normal with mean 0 and variance wSD_{js}^2, and μ_{ijs}, X_{ijs}, and ε_{ijsk} are assumed to be conditionally independent given j and s. Then $Var(Y_{ijsk} - X_{ijs}|js) = Var(\mu_{ijs} - X_{ijs} + \varepsilon_{ijsk}|js)$

$$= Var(\mu_{ijs} - \mu_{js}|js) + Var(X_{ijs} - X_{js}|js) + Var(\varepsilon_{ijsk}|js)$$
$$= \sigma^2_{Njs} + \sigma^2_{Tjs} + wSD^2_{js}$$

where σ^2_{Njs} is the between-nodule variability and σ^2_{Tjs} is the availability of true volume across nodules. Thus, we have $SD_{js}(measured - truth) = \sqrt{wSD^2_{js} + \sigma^2_{Njs} + \sigma^2_{Tjs}}$.

Following the same approach as in Bland and Altman[19] on estimating the SD^2_{js} by using data with replications, the unbiased estimator is $\widehat{SD}^2_{js} = s^2_{\bar{d}_{js}} + (1 - \frac{1}{K})\widehat{wSD}^2_{js}$, where $s^2_{\bar{d}_{js}}$ is the sample variance of mean differences $\bar{d}_{ijs} = \bar{Y}_{ijs} - X_{ijs}$ with $\bar{Y}_{ijs}$ as the mean of K replicates. This is because

$$E\left(\widehat{SD}^2_{js}|js\right) = Var\left(\bar{Y}_{ijs.} - X_{ijs}|js\right) + \left(1 - \frac{1}{K}\right)wSD^2_{js}$$
$$= Var\left(\bar{Y}_{ijs.} - \mu_{ijs} + \mu_{ijs} - X_{ijs}|js\right) + \left(1 - \frac{1}{K}\right)wSD^2_{js}$$
$$= Var\left(\bar{\varepsilon}_{ijs.}|js\right) + Var\left(\mu_{ijs}|js\right) + Var(X_{ijs}|js) + \left(1 - \frac{1}{K}\right)wSD^2_{js}$$
$$= \frac{wSD^2_{js}}{K} + \sigma^2_{Njs} + \sigma^2_{Tjs} + \left(1 - \frac{1}{K}\right)wSD^2_{js}$$
$$= \sigma^2_{Njs} + \sigma^2_{Tjs} + wSD^2_{js}$$

Note that if there are unequal replicates K_{ijs}, then $\left(1 - \frac{1}{K}\right)$ is replaced by $\left(1 - \frac{1}{n_{js}}\sum_{i=1}^{n_{js}}\frac{1}{K_{ijs}}\right)$ following the same argument as in Bland and Altman.[19]

Derivations in Section 4

In the QIBA 3a study, volume measurements on a nodule were taken using two different slice thicknesses 0.8 and 5 mm. Suppose for QIBA 3a, slice thickness is regarded as having not fixed effects as in equation (4), but random effects. For the kth measurement of volume by an algorithm on nodule i with slice thickness t, $i = 1, 2, \ldots, n$, $t = 0.8\,\text{mm}, 5\,\text{mm}$, $k = 1, 2, \ldots, K_{it}$, consider the model

$$Y_{itk} = \mu + \gamma_i + \tau_t + (\gamma\tau)_{it} + \varepsilon_{itk} \tag{6}$$

with random effect $\gamma_i \sim N(0, \sigma^2_\gamma)$ for nodule i, random effect $\tau_t \sim N(0, \sigma^2_\tau)$ for slice thickness t, random nodule-by-slice thickness interaction effect $(\gamma\tau)_{it} \sim N(0, \sigma^2_{\gamma\tau})$, and random effect $\varepsilon_{itk} \sim N(0, \sigma^2_\varepsilon)$ for repeated measurement k within nodule i and slice thickness t.

From equation (6), the difference $d_{new} = Y_{i_{new}tk} - Y_{i_{new}t'k'}$ between two measurements $Y_{i_{new}tk}$ and $Y_{i_{new}tk'}$ taken on new lesion i_{new} has mean 0 and variance $2(\sigma^2_\tau + \sigma^2_{\gamma\tau} + \sigma^2_\varepsilon)$. The reproducibility coefficient (RDC) may be defined as the least significant difference between two measurements when conditions (e.g. slice thickness) changed between the measurements. For model (6), RDC is

$$RDC = 2.77\sqrt{\sigma_\tau^2 + \sigma_{\gamma\tau}^2 + \sigma_\varepsilon^2} \tag{7}$$

RDC specifically includes the variance components σ_τ^2 and $\sigma_{\gamma\tau}^2$ due to using different slice thicknesses when making two measurements. When slice thickness does not change between the two measurements, these variance components vanish and *RDC* becomes

$$RDC_S = 2.77\sigma_\varepsilon$$

Because nodules are re-positioned in QIBA 3a, conditions are changing and thus we still consider $2.77\sigma_\varepsilon$ as being a reproducibility coefficient when some conditions change but others are held constant. In general depending on the design of an experiment, when the random factors being modeled are held constant, 2.77 times the square root of the residual error variance could be regarded are the *repeatability coefficient*, the least significant difference between two measurements taken under the same conditions.[5]

For each algorithm, we estimated *RDC* (mm^3) and RDC_S for each of the three groups of nodules with diameters 8–10, 20, and 40 mm (Tables 18 to 20). We used SAS Proc Mixed to obtain REML estimates of the variance components. We also obtained these *RDC*s as a percent of the true volume.

For algorithms 2 and 3, the variance component estimates for slice thickness and slice thickness by nodule interaction were, in general, small relative to the error variance estimate, resulting in an *RDC* similar to RDC_S for all three nodule groups. In contrast, for algorithms 1 and 4, variance components involving slice thickness were large, resulting in an *RDC* larger than RDC_S. For example, for lesions 40 mm in diameter, algorithm 4 has a large variance component for slice thickness, resulting in an *RDC* as a percent of true volume of 12.9% compared with 3.8% for RDC_S.

As an alternative to model (6), a model may be used that does not specifically separate from residual error the variance components due to changing slice thickness. Consider the one-way random effects model

$$Y_{ik} = \mu + \gamma_i + \eta_{ik}, \quad i = 1, \ldots, n, \; k = 1, \ldots, K_i \tag{8}$$

Table 18. *RDC* (mm^3), Lesion sizes 8–10mm (n = 10, 36 total measurements).

		REML variance component estimate								
Alg	Mean	Les	Thc	Thc × Les	Error	RDC_S	RDC_S as % of average measured volume	Within Les total SD[a]	RDC	RDC as % of average measured volume
1	517.57	6246.2	0	3182.66	4961.39	195.2	37.7%	90.2	250.1	48.3%
2	472.52	4933.5	0	0	10398.00	282.6	59.8%	102.0	282.6	59.8%
3	663.02	244360	0	0	69471.00	730.6	110.1%	263.6	730.6	110.1%
4	641.4	12845	12201	396.76	3078.98	153.8	24.0%	125.2	347.0	54.1%

[a]Within lesion Total SD is square root of sum of variance components for Slice Thickness (Thc), the slice thickness by lesion interaction (Thc*Les), and error.

where $\gamma_i \sim N(0, \sigma_\gamma^2)$ is the random effect for lesion i and $\eta_{itk} \sim N(0, \sigma_\eta^2)$ is the residual error for K_i repeated measurements on nodule i. The variation due to measurement with different slice thicknesses is subsumed into this residual error term. Two estimates of σ_η are

$$\overline{wSD} = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{\sum_{k=1}^{K_i}(Y_{ik} - Y_{i.})^2}{K_i - 1}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\frac{1}{nK_i - n}(Y_{ik} - Y_{i.})^2$$

$$MSE = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K_i}(Y_{ik} - Y_{i.})^2}{\sum_{i=1}^{n}K_i - n} = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\frac{1}{\sum_{i=1}^{n}K_i - n}(Y_{ik} - Y_{i.})^2$$

$\overline{wSD}$ is the nonparametric straight average of the wSDs for each nodule described in Section 4.3. MSE is the REML estimate pooling the wSDs according to the number of measurements per nodule. Note that $\frac{1}{nK_i - n} = \frac{1}{\sum_{i=1}^{n}K_i - n}$ for all i if $K_i = K$, but in general $wSD \neq MSE$.

Table 19. *RDC* (mm^3), Lesion Size 20 mm (n = 10, 44 total measurements).

		REML variance component estimate[a]								
Alg	Mean	Les	Thc	Thc × Les	Error	RDC_S	RDC_S as % of average measured volume	Within Les total SD[a]	RDC	RDC as % of average measured volume
1	4342.2	0	12851	12892	75673	762.5	17.6%	318.5	882.7	20.3%
2	3993.2	2907.4	0	12165	139222	1034.2	25.9%	389.1	1078.5	27.0%
3	4894.6	403976	0	0	499658	1959.3	40.0%	706.9	1959.3	40.0%
4	4563.5	72350	35696	18279	24209	431.3	9.45%	279.6	775.0	17.0%

[a]Within lesion Total SD is square root of sum of variance components for Slice Thickness (Thc), slice thickness by lesion interaction (Thc*Les), and error.

Table 20. *RDC* (mm^3), Lesion Size 40 mm (n = 2, 10 total measurements).

		REML variance component estimate[a]								
Alg	Mean	Les	Thc	Thc × Les	Error	RDC_S	RDC_S as % of average measured volume	Within Les Total SD*	RDC	RDC as % of average measured volume
1	34,364	0	155,747	235,084	213,733	1281.4	3.73%	777.5	2155.2	6.27%
2	31,232	9,968,781	0	0	5,601,553	6560.2	21.0%	2366.8	6560.2	21.0%
3	34,915	473,030	0	71,932	254,323	1397.8	4.00%	571.2	1583.2	4.53%
4	35,061	806,713	2,428,659	0	232,265	1335.8	3.81%	1631.2	4521.5	12.9%

[a]Within lesion Total SD is square root of sum of variance components for Slice Thickness (Thc), slice thickness by lesion interaction (Thc*Les), and error.

We now show that except in trivial cases $RDC > 2.77 \times E(MSE)$ in the balanced case of the same number of repeated measurements for all cases i and slice thicknesses s, where RDC is computed under model (6) and MSE is computed under model (8). That is, by not modeling explicitly the variance components due to slice thickness and defining the reproducibility coefficient explicitly as the least significant difference when slice thickness changes between measurements, $2.77 \times MSE$ will tend to underestimate RDC.

To index for slice thickness, rewrite model (8) as

$$Y_{itk} = \theta + \gamma_i + \eta_{itk} \tag{9}$$

where in the balanced case $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, T$, and $k = 1, 2, \ldots, K$. From model (9), the REML estimate of σ^2_η is $MSE = SSE/dof$, where sum of squares $SSE = \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{k=1}^{K} (y_{itk} - y_{i\bullet\bullet})^2$ has degrees of freedom $dof = n(SK - 1)$. This REML estimate is unbiased, i.e. the expectation $E(MSE) = \sigma^2_\eta$.

In the balanced case, the analysis of variance for model (6) is as follows:

Factor[a]	Dof	SS	E(MS)
Les	$n - 1$	$SK \sum_{i=1}^{n} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$	$SK\sigma^2_\gamma + K\sigma^2_{\gamma\tau} + \sigma^2_\varepsilon$
Thc	$S - 1$	$nK \sum_{s=1}^{S} (\bar{y}_{\bullet s\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$	$nK\sigma^2_\tau + K\sigma^2_{\gamma\tau} + \sigma^2_\varepsilon$
*Les*Thc*	$(n - 1)(S - 1)$	$K \sum_{i=1}^{n} \sum_{s=1}^{S} (\bar{y}_{is\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet s\bullet} + \bar{y}_{\bullet\bullet\bullet})^2$	$K\sigma^2_{\gamma\tau} + \sigma^2_\varepsilon$
Error	$nS(K - 1)$	$\sum_{i=1}^{n} \sum_{s=1}^{S} \sum_{k=1}^{K} (y_{isk} - y_{is\bullet})^2$	σ^2_ε

[a]Factors are lesion (*Les*), slice thickness (*Thc*), slice thickness by lesion interaction (*Les*Thc*), and *Error*.

From this analysis of variance table, SSE under model (9) is the sum of the sum of squares $SSThc$, $SSLes * Thc$, and $SSError$ because they are mutually orthogonal. By multiplying the respective expected mean squares by their degrees of freedom, summing, and dividing by dof, we see that MSE under model (8) has under model (6) the expectation

$$E(MSE) = \frac{nK(T-1)}{n(TK-1)} \left(\sigma^2_\tau + \sigma^2_{\gamma\tau} \right) + \sigma^2_\varepsilon$$

Note $E(MSE)$ equals reproducibility variance $\sigma^2_\tau + \sigma^2_{\gamma\tau} + \sigma^2_\varepsilon$ only in the trivial cases $K = 1$ or $\sigma^2_\tau = \sigma^2_{\gamma\tau} = 0$, but otherwise $E(MSE) < \sigma^2_\tau + \sigma^2_{\gamma\tau} + \sigma^2_\varepsilon$. Thus, from equation (7), $2.77 \times E(MSE) < RDC$ except in the trivial cases.