The VOLCANO'09 Challenge: Preliminary Results

Anthony P. Reeves[1], Artit C. Jirapatnakul[1], Alberto M. Biancardi[1], Tatiyana V. Apanasovich[2], Chris Schaefer[3], Jeffrey J. Bowden[3], Markus Kietzmann[4], Rene Korn[4], Markus Dillmann[4], Qiang Li[5], Jiahui Wang[5], Jan H. Moltz[6], Jan-Martin Kuhnigk[6], Tatsuro Hayashi[7], Xiangrong Zhou[7], Hiroshi Fujita[7], Thomas Duindam[8], Bram van Ginneken[8], Rick Avila[9], Jane P. Ko[10], Kira Melamud[10], Henry Rusinek[10], Rafael Wiemker[11], Grzegorz Soza[12], Christian Tietjen[12], Matthias Thorn[12], Michael F. McNitt-Gray[13], Yanisley Valenciaga[13], Maryam Khatonabadi[13], Yoshiki Kawata[14], Noboru Niki[14]

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca (NY), USA
[2]Jefferson Medical College, Thomas Jefferson University, Philadelphia (PA), USA
[3] Imaging Services, Biomedical Systems, USA
[4] Definiens AG, Germany
[5] Department of Radiology, Duke University Medical Center, USA
[6] Institute for Medical Image Computing, Fraunhofer MEVIS, Germany
[7] Gifu University, Japan
[8] Image Sciences Institute, University Medical Center Utrecht, the Netherlands
[9] Kitware, Inc.
[10] Radiology Department, New York University, USA
[11] Philips Research Hamburg
[12] Computed Tomography, Healthcare Sector, Siemens AG, Germany
[13] David Geffen School of Medicine at UCLA, USA
[14] University of Tokushima

reeves@ece.cornell.edu, tatiyana.apanasovich@jefferson.edu, cschaefer@biomedsys.com, mkietzmann@definiens.com, jiahui.wang@duke.edu, jan.moltz@mevis.fraunhofer.de, hiroshi.fujita.gifu@gmail.com, bram@isi.uu.nl, rick.avila@kitware.com, hr18@nyu.edu, rafael.wiemker@philips.com, grzegorz.soza@siemens.com, mmcnittgray@mednet.ucla.edu, niki@opt.tokushima-u.ac.jp

Abstract. The VOLCANO'09 Challenge invited participants to evaluate the change in size of pulmonary nodules in CT images; the challenge data set consisted of 50 pairs of CT scans each scan containing a single nodule. This is the first challenge for CAD methods on pulmonary nodules in which size change rather than volume estimation is the primary endpoint. Responses from 13 teams were received with size change results for a total of 17 different methods. In this paper the challenge data set is described and statistical results computed from the submissions are presented. The dataset consisted of several subgroups: (a) zero-change cases, cases with different slice thickness scans, cases with actual size change and a synthetic nodule case. No statistical difference was found between the methods; a slice thickness change was significant and there was an interesting bias observed for some zero-change nodules.

Keywords: Size Change Estimation, Variability Analysis, Pulmonary Nodules.

1 Introduction

The target of the challenge is three-dimensional change analysis of pulmonary nodules in CT images. The focus of the challenge is not directly on image segmentation itself (which tells us little of the underlying disease) but rather the change in size of the nodule recorded on two time-separated images. This size change is a critical measurement for (a) diagnosing cancer and (b) evaluating response to therapy. One of the most important indicators of malignancy is the relative change in size of a nodule over a period of time. The critical issue for the challenge, the precision of size change measurement, is needed to establish the minimum time delay between sequential scans and the associated magnitude of the measurement required to determine malignancy or response to therapy. There has been one previous pilot study in this area, Biochange 2008 [1].

Most evaluation methods for CAD systems, including challenges, involve a ground truth established be experts. However, for the task of nodule size estimation it is well known that there is a large amount of variation or disagreement in expert size estimations [2]. Further, it has not been established that expert's manual estimations are superior to automated measurements. In this challenge, while the change in size of nodules will be reviewed by experts, we explore the issue of ground truth through the submitted responses to the challenge.

1.1 Motivation for the study

Current approaches to quantification of nodule volume change measurement exhibit two main problems that complicate their direct comparison. First, these methods require a large unified database of both stable and growing nodules. Second, there is no single commonly used evaluation technique that would assess the measurement quality of a particular method. Therefore we invited interested parties to take part in this unique study that addresses both of these issues by providing a single evaluation image dataset and a common methodology for assessing the quality of the measurement algorithm.

1.2 The VOLCANO Challenge

The challenge involved measuring the change in nodule size for 50 scan pairs. Four additional scan pairs were made available for training. Teams reported the fractional change in nodule size for each of the 50 scan pairs. Thirteen different teams submitted their measurement change results from on a total of 17 different methods. In 12 of these methods, the actual volumes recorded for each nodule were also reported.

The participants were only informed that there were 50 nodule pairs; however, the data may be divided into four subgroups:

A. (14) zero-change in which the scans were taken minutes apart and therefore there is no real change in the nodule size.
B. (13) zero-change cases as in A above except one scan had a slice thickness of 1.25 mm and the second scan had a larger slice thickness (2.5 or 5.0 mm)

C. (19) nodules with a significant time interval between scans and therefore some real change and (3) nodules with a large change in size of greater than 150%, one of which was known to be malignant. Of these nodules, 19 were considered to be stable or benign by biopsy and 3 were diagnosed as malignant.
D. (1) synthetic phantom nodule with a known size recorded multiple times with different slice thicknesses

In general, the main interest is to learn the smallest size change which can be reliably detected and the precision in that size change measurement. A number of studies on repeat scans have been reported in the literature ([3] refers to several of these). In these studies the limits of agreement for repeat scans of the same nodule are in the order of 20-25% by volume. For theses reasons most of the cases in data set C were selected to have a size change within the range of ±50%. For completeness, three cases with a very large size change (150% or more) were included to characterize the measurement methods for such situations.

1.3 Overview

In this paper we report the initial statistical findings for the data submitted by the VOLCANO teams that provides collective information about the nodule size change measurement process; we do not provide a performance rating for the different teams. This paper addresses the following issues:

1. What is the precision of change measurement to be expected from computer assisted measurement methods? This question is addressed by considering the size change measurements from group A.
2. What is the impact on the computer methods of changing the slice thickness of the CT scan? This is addressed by the analysis of any bias for the size change measurements of group B and a comparison of the variation of groups A and B. Also, we can observe if there is a bias in group D.
3. What is the variation to be expected between different computer methods when there is an actual size change? This is addressed by the variation in measurements for group C.
4. For the teams that provided volume information, what is variation in volume estimates made by the different methods? For comparative volume measurements we consider groups A-C and for absolute volume measurement we consider group D.

2 Materials and Methods

The image data used in the study was acquired for the Public Lung Database to address drug response [4] and was provided by the Weill Cornell Medical College with the exception of one case of a synthetic “phantom” nodule provided by the FDA [5]. Cases were selected that contained at least one nodule of solid consistency which was present in at least two scans with a whole-lung field of view; only nodules visible on at least three slices on both scans were included.

Table 1. Summary of scan parameters

Group	Current (mA)	kVp (kVp)	Scanner Models
A	40-250	120	GE LightSpeed (Ultra, QX/i, Pro 16)
B	20-80	120	GE LightSpeed (Ultra, QX/i)
C	40-80	120	GE LightSpeed (Ultra, Pro 16, VCT)
D	40	120	Philips MX8000 IDT 16

The VOLCANO'09 dataset consists of 50 pairs of CT scans of pulmonary nodules: 49 scan pairs of real pulmonary nodules and one phantom nodule. The data consisted of four subgroups A-D as outlined in the introduction. The size distribution for the nodules used in this study is shown in Figure 1.

The scan pairs in group A had the same slice thickness for both scans; in 13 cases, both scans had a slice thickness of 1.25 mm while one case had scans with a slice thickness of 2.5 mm. For group B, 11 cases had scans with 1.25 mm and 2.5 mm slice thickness, while two cases had scans of 1.25 and 5.0 mm slice thickness. All the scans in group C had a slice thickness of 1.25 mm. Scans in these three groups were acquired without overlap. Other relevant scan parameters are listed in Table 1.

For group C, the status of the nodule was determined by a radiologist; stable nodules were either biopsied (3) or did not have any clinical change in 2 years (16), while the three malignant nodules were biopsied.

Group D was comprised of the synthetic phantom nodule, a 10 mm (523.60 mm^3) sphere with two different slice thickness reconstructions, 1.5 mm and 3.0 mm with 50% overlap. All scans have a whole-lung field of view. The phantom was placed in a chest phantom with simulated vascular structures [5].

In five zero-change cases, the patient was oriented in a different position between the scans; one case was in group A while four cases were in group B.

2.1 VOLCANO Data Preparation

Prior to making the images available for the challenge, all identifying patient information was removed. The original dates of the scans were replaced with dates corresponding to a time interval of 100 days between scans, with the order of the scans was randomized. The scans were then clipped in the axial direction, with the five slices below and above the region containing the nodule included if possible. This was done (a) because some of the scans did not cover the whole-lung in the axial dimension and (b) to reduce the amount of data to be downloaded for this study.

Along with the scans, teams were provided with a spreadsheet identifying the approximate center of the nodule established by a human observer.

2.2 The data measurement methods

Submissions to the VOLCANO challenge were received from 13 teams. Several teams submitted multiple size change measurement methods for a total of 17 submitted methods. Although not required, 12 teams provided volume measurements for each nodule. The teams and their methods are summarized in Table 2.

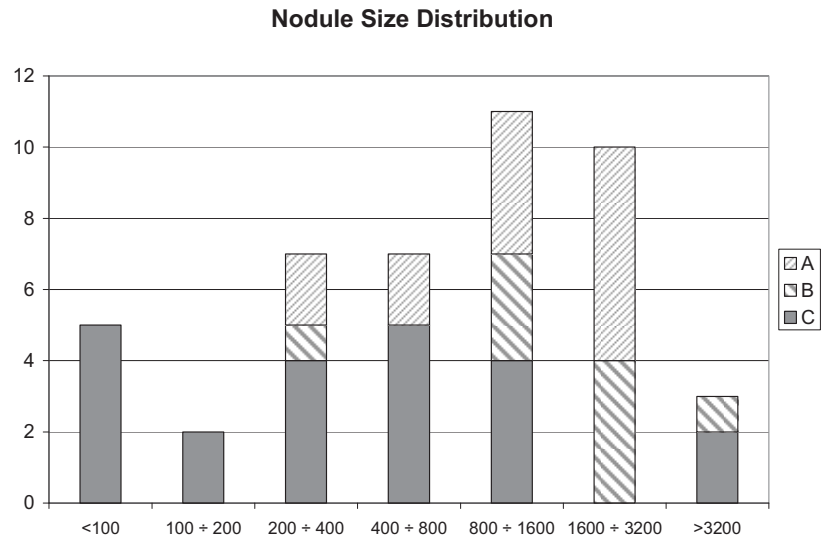


Fig. 1. The distributions of the approximate nodule sizes, expressed as volumes in mm^3, according to the different groups they belong to. The listed volumes (for which an equivalent diameter is defined as the diameter of the sphere having that volume) correspond to the following equivalent diameters, respectively: 5.8, 7.3, 9.1, 11.5, 14.5, and 18.3 mm.

Table 2. The methods submitted to VOLCANO'09. (PVC = partial voxel compensation)

Method	Team/Method	Automation	Method
1	Tokushima	3	Image filtering
2	ISI, SPHERE	4	Sphere fitting
3	ISI, SEG	4	Image filtering
4	ISI, REG	4	Elastic registration
5	NYU, HYB	7	Image filtering and PVC
6	NYU, HYBA	7	Image filtering and PVC
7	UCLA	6	Image filtering
8	VIA, GAD	1	Density change
9	VIA, GAS	4	Image filtering
10	Kitware	4	Fast marching and shape detection level set
11	Duke	1	Spiral scanning, dynamic programming
12	Gifu	1	Image filtering
13	Biomedsys	2	Image filtering
14	MeVIS	3	Image filtering and PVC
15	Siemens	3	Image filtering and PVC
16	Philips	1	Active contour
17	Definiens	5	Image filtering

The methods had varying levels of operator interaction which can be divided broadly into three groups: completely automated after specification of a seed point, manual parameter control, and modification of the resulting boundary or indicating additional control points. These categories were further subdivided according to the fraction of cases which required manual intervention. Teams were asked to rank the level of automation required by their algorithms using the following scale:

1. Totally automatic using seed points
2. Limited parameter adjustment (on less than 15% of the cases)
3. Moderate parameter adjustment (on less than 50% of the cases)
4. Extensive parameter adjustment (more than 50% of the cases)
5. Limited image/boundary modification (on less than 15% of the cases)
6. Moderate image/boundary modification (on less than 50% of the cases)
7. Extensive image/boundary modification (more than 50% of the cases)

The level of automation required for each method is detailed in Table 2.

Although each method was unique, there were several common approaches to the task of volume change measurement. One approach taken by seven methods (Tokushima, ISI-Seg, UCLA, VIA-GAS (Vision and Image Analysis Group, Cornell), Gifu, Biomedsys, Definiens) is based on simple image filtering operations [6]. Generally, methods using this approach extracted a volume of interest (VOI) around the seed point for each nodule. Definiens additionally used an automatically generated anatomical model to further restrict the VOI. This volume of interest was resampled and a threshold applied to identify voxels belonging to high-intensity structures. Next, either region growing or connected component analysis is applied to the volume of interest to eliminate non-nodule structures, followed by the removal of attached structures such as vessels or the chest wall using morphological filtering or other more advanced techniques. Four methods (MeVIS, NYU-HYB, NYU-HYBA, Siemens) extended this approach to better address partial voxels along the perimeter of the nodule [7, 8]. These methods determine a region around the border of the nodule where the voxels have intensity between that of solid tissue and the lung parenchyma; based on histogram analysis, these voxels are weighted when computing the nodule volume.

The remaining six methods used different approaches. Most methods resampled the CT scans into isotropic space. ISI-Sphere estimated the best fitting spherical volume of interest at the seed point of the nodule from a thresholded, resampled volume of interest. The volume of the nodule was estimated from the number of voxels above a threshold. The ISI-Registration method applied non-rigid registration to deform the first scan to the second; this transformation was then applied to a segmentation obtained for the first scan to obtain the volume of the nodule on the second scan. Kitware required only a manually specified seed point and a bounding box. Their method computed several features for each voxel, including lung wall, vesselness, gradient, and intensity features which were aggregated and used to guide a fast marching algorithm to generate an initial guess of the nodule boundary; this guess was refined using a shape detection level set. The volume was computed from the surface of the resulting level set. Duke used a spiral-scanning technique to convert the 3D volume of interest around the nodule to a 2D generalized polar coordinate system. Dynamic programming techniques were used to obtain the nodule boundary on the 2D image which was then transformed back into 3D space [9]. This boundary was

applied to the original 3D image to estimate the nodule volume. Philips used an active contour-based approach utilizing a radial basis function energy minimization algorithm [10]. In contrast to the methods described thus far, VIA-GAD (Cornell) did not explicitly segment the nodule; instead, the change in nodule size was estimated from the change in density of a Gaussian-weighted region around the nodule [11].

2.3 Data analysis

Teams were requested to provide a size change metric for each nodule. For methods using volumes, the size change metric would be:

$$RVC = \frac{V_2 - V_1}{V_1} \quad \textbf{(1)}$$

The data metric that was requested for results (relative change with respect to time 1) is asymmetric with respect to the order of presentation: no change in size has a value of 0, a 10 times increase in size results in a value of 9 while a 10 times decrease in size results in a value of - 0.9. The negative size change is bounded by -1 while the positive size change is unbounded. The order of the scan pairs was randomized. It is possible to reverse the order: that is computing the relative size change from scan two to scan one by the following transformation:

$$RVC' = \frac{1}{RVC + 1} - 1 = \frac{V_1}{V_2} - 1 \quad \textbf{(2)}$$

where RVC represents the reported relative size change. However, if any of the methods have an order bias, flipping the results to match the correct time sequence would mask such an effect. The one data set where ordering is important is data set B; we flipped the results so that the thin slice scan was always the first in the pair to determine the effect caused by a change in slice interval. For the other data sets flipping to restore the correct temporal order of the scans was not performed.

To quantify the variation and bias of each group, the median of the median of absolute deviation (MMAD) and median of the absolute median (MAM) were computed for each group. The median of the absolute deviation (MAD) is computed by taking the median size change metric for each nodule, and, for each method, computing the absolute deviation from the median. This results in a MAD value for each nodule; the median of these is then reported as the MMAD. The MAM is computed by taking the median of the absolute median for each nodule. Both these measures can be computed for each method instead of for each nodule.

The statistical agreement between methods was established for the size change measurements using the Friedman test. Groups were compared by applying the Wilcox rank-sum test to the MAD values computed for each nodule.

3 Results

Box and whisker plots for the different groups are shown in Figures 2 to 6. In the plots, the median of each nodule is indicated by a thick line. The 25% to 75% interquartile range (IQR) is represented by the box. Methods with values inside the box are generally in good agreement; 50% of the methods lie inside the IQR. The lines above and below the box ("whiskers") represent the largest and smallest values that are within 1.5 times the IQR; any more extreme values are considered outliers and indicated by an open circle. The number shown x-axis is a random case identifier that does not correspond to the case identifiers given to the teams; the results are ordered according to median change. For group C, the three cases with the most change are plotted separately for visibility.

For each of the groups, the median of the absolute medians (MAM) and the median of the median of the absolute deviations (MMAD) were computed across methods. These are measures of the bias and variation, respectively, and reported in Table 3.

Table 3. Summary of the median absolute median (MAM) and median of median absolute deviation (MMAD) of the relative size change for each group. Note that the results of group D are on a single nodule.

Group:	A	B	C	D
MAM	0.0572	0.1020	0.1740	0.0620
MMAD	0.0388	0.0899	0.0592	0.0761

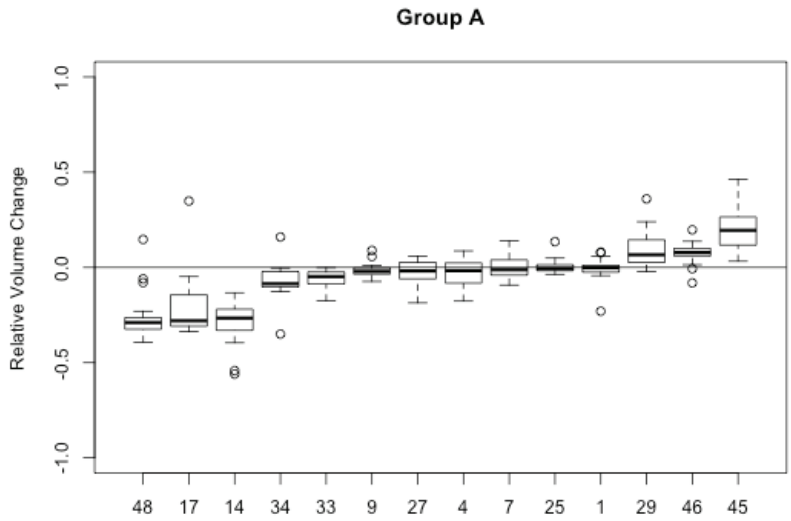


Fig. 2. Box plot for group A, the zero-change dataset

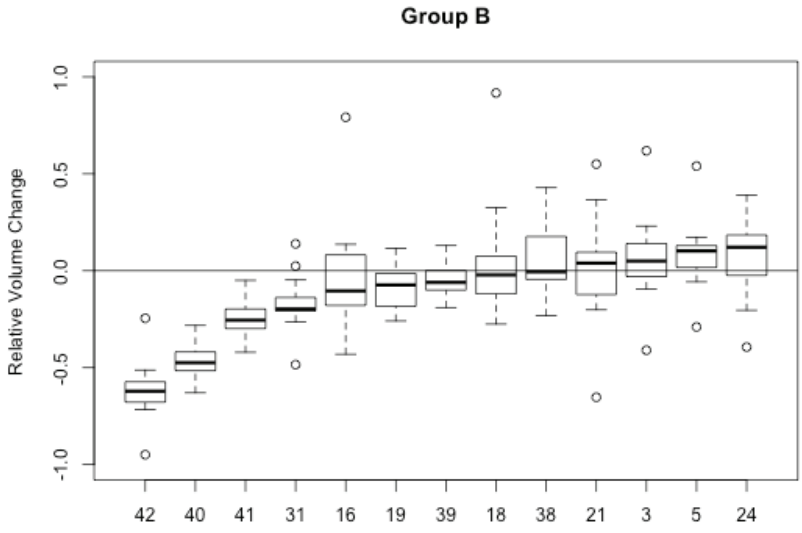


Fig. 3. Box plot for group B, zero-change with different slice interval, with the volume change inverted when necessary to order by small to large slice interval. All nodules were imaged on scans of 1.25 and 2.5 mm slice thickness except for 3 and 21 which were 1.25 and 5.0 mm.

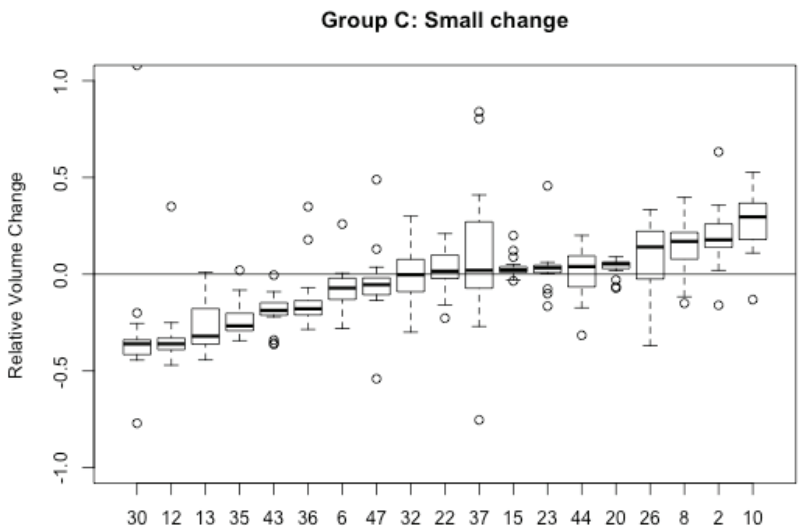
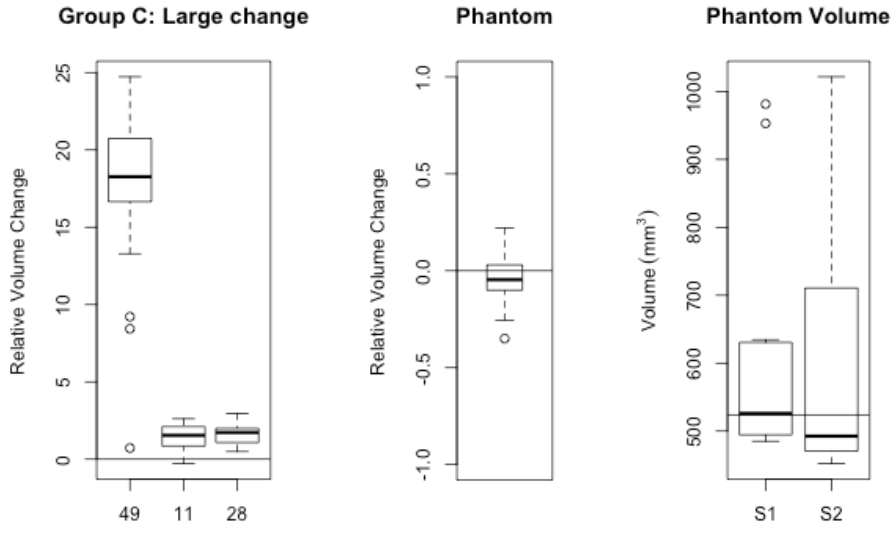


Fig. 4. Box plot for group C, the actual small change nodules. One outlier is not shown.



(a) (b) (c)

Fig. 5. (a) Group C nodules with large change, with the volume change for the first nodule inverted for visibility and (b) size change and (c) volume variation results for the synthetic nodule for scan 1 (S1) and scan 2 (S2). The volume results for the phantom are presented for 12 of 17 methods that provided volumes.

4 Discussion

4.1 The repeat measurement behavior of the computer assisted methods

In group A, the scans are a repeat measure on nodules that have not changed in size. The median MAD was 0.0387, which indicates the level of variation between the methods. The bias from the true zero value of size change is represented by the median absolute median of 0.0572.

To examine the variation due to the nodule and/or scanner, the median MAD was computed across nodules. The value of 0.0700 was larger than the median MAD computed across methods; thus, the variation due to the nodule/scanner is larger than the variation due to the methods. In addition, the median of the absolute medians was lower, 0.0217, suggesting that any bias that exists is due to the nodule/scanner and not the method. The 85% confidence interval of the absolute medians across methods (i.e., omitting the two largest cases) is 0.266 or a 26% volume change. It is interesting to note that for 8 of the 14 cases the interquartile range (IQR) does not include zero. Therefore there is evidence of a systematic bias introduced by the scanner/nodule combination for these cases.

4.2 The impact of change in the CT scan slice thickness

The size change measurements in group A for all the methods are in agreement ($p=0.92$ using the Friedman test). In contrast, the size change measurements for group B, where there is a change in slice thickness, are significantly different between the methods ($p < 0.01$) according to the Friedman test. To determine if the variation is greater in group B, the median absolute deviation (MAD) was computed for each nodule in group A and group B and the Wilcoxon rank-sum test was performed. The variation was higher in group B than group A ($p < 0.01$); with $p=0.05$, the variation is 40% higher in group B than group A.

We observe from Fig. 5c that there is a greater dispersion in the IQR of the volume measurement for the thicker slice scan of the phantom nodule (131.01 mm^3 vs. 207.79 mm^3) which is consistent with the above.

4.3. The variation of the methods when there is a real size change

The variation with respect to no-change and real change data sets is achieved by comparing the median of the median of absolute deviation (MAD) values between group A and the 19 small change cases in group C. The median MAD for group A and the small change cases in group C were 0.0387 and 0.0590 respectively. The small increase might suggest that the results obtained from a zero-change dataset capture similar behavior of a dataset with a small amount of change as represented by group C. Furthermore, comparison with the 3 nodules with larger size changes in group C showed that there was an increased variation (median MAD of 0.4446) for a larger size change.

4.4 A comparison of volume estimation and size change measurement

For the 11 teams that provided volume information, the variation in the size change measurement was compared to the volume measurements for group A and the small change subset of group C. The Friedman test was performed to determine if the methods were in agreement.

In group A, there was no significant difference between the size change measurements of the methods ($p = 0.92$), but there was significant disagreement between the volume measurements of the methods ($p < 0.01$). For the small change subset of group C, again there was no significant difference between the size change measurements of the methods ($p = 0.97$), but significant disagreement in the volume measurements ($p < 0.01$). This may be due to a bias in volume measurement between methods that might be neutralized when computing size change.

5 Conclusion

Change in size measurements made on 50 nodule image pairs were reported for 17 different methods. The analysis of the results showed (a) that there was no statistical difference between the methods on scans of the same slice thickness, (b) that there was a statistical difference in the methods when the scan slice thickness is changed,

and (c) that the behavior of the methods for nodules with a small real change in size was similar to that for the zero-change data. The last point has implications for the validity of using zero-size change datasets for evaluating nodule measurement performance. For 11 of the methods volume measurements were provided in addition to the size change measurements. The volume measurements did show a statistical difference between methods; therefore, caution is needed when extrapolating from studies that focus only on volume estimation when size change is the intended task.

Acknowledgments

We are grateful to David Yankelevitz of the Weill Medical College of Cornell University for reviewing the cases used in this study. We gratefully acknowledge Lisa Kinnard of the FDA for providing us with the synthetic nodule CT images used in this study.

References

1. Fenimore, C. Biochange 2008 Pilot, http://www.itl.nist.gov/iad/894.05/biochange2008 /Biochange2008-webpage.htm
2. Reeves, A. P., Biancardi, A. M., Apanasovich, T. V. et al.: The Lung Image Database Consortium (LIDC): A Comparison of Different Size Metrics for Pulmonary Nodule Measurements. Academic Radiology 14, 1475--1485 (2007).
3. Gavrielides M.A., Kinnard L.M., Myers K.J., Petrick N.: Noncalcified lung nodules: volumetric assessment with thoracic CT. Radiology 251, 26--37 (2009).
4. Reeves, A. P., Biancardi, A. M., Yankelevitz, D. F., Fotin, S., Keller, B. M., Jirapatnakul, A. C., Lee, J.: A Public Image Database to Support Research in Computer Aided Diagnosis, In: 31st Annual International IEEE EMBS Conference, in press, IEEE Press, New York (2009).
5. Gavrielides, M.A., Zeng, R., Kinnard, L.M., Myers, K.J., Petrick, N.: A model-based approach for the analysis of lung nodules in a volumetric CT phantom study. In: SPIE Medical Imaging Conference, vol. 7260, pp. 726009, SPIE (2009).
6. Kostis, W. J., Reeves, A. P., Yankelevitz, D. F., Henschke, C. I.: Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. IEEE Trans. on Medical Imaging 22, 1259--1274 (2003).
7. Ko, J.P., Rusinek, H., Jacobs, E.L., Babb, J.S., Betke, M., McGuinness, G., Naidich, D. P.: Small Pulmonary Nodules: Volume Measurement at Chest CT: Phantom Study. Radiology 228, 864--870 (2003).
8. Kuhnigk, J., Dicken, V., Bornemann, L., Bakai, A., Wormanns, D., Krass, S., Peitgen, H.: Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. IEEE Trans. on Medical Imaging 25, 417--434 (2006).
9. Wang, J., Engelmann, R., Li, Q.: Segmentation of pulmonary nodules in three-dimensional CT images by use of a spiral-scanning technique. Medical Physics 34, 4678--4689 (2007).
10. Opfer R., Wiemker R.: A new general tumor segmentation framework based on radial basis function energy minimization with a validation study on LIDC lung nodules. In: SPIE Medical Imaging Conference, vol. 6512, pp. 651217, SPIE (2007).
11. Jirapatnakul, A. C., Reeves, A. P., Biancardi, A. M., Yankelevitz, D. F., Henschke, C. I.: Semi-automated measurement of pulmonary nodule growth without explicit segmentation. In: IEEE International Symposium on Biomedical Imaging. pp. 855-858. IEEE Press, New York (2009).