NISTIR 7879

Statistical Analysis of Reader Measurement Variability in Nodule Sizing with CT Phantom Imaging Data

Zhan-Qian John Lu
Charles Fenimore
Nicholas Petrick
Rongping Zeng
Marios A. Gavrielides
David Clunie
Kristin Borradaile
Robert Ford
Hyun J. Grace Kim
Michael F. McNitt-Gray
Binsheng Zhao
Andrew J Buckler

http://dx.doi.org/10.6028/NIST.IR.7879

National Institute of Standards and Technology
U.S. Department of Commerce

AUTHENTICATED U.S. GOVERNMENT INFORMATION GPO

NISTIR 7879

Statistical Analysis of Reader Measurement Variability in Nodule Sizing with CT Phantom Imaging Data

Zhan-Qian John Lu
Statistical Engineering Division, Information Technology Laboratory

Charles Fenimore
Information Access Division Information Technology Laboratory

Nicholas Petrick
Rongping Zeng
Marios A. Gavrielides
US Food & Drug Administration, Center for Devices & Radiological Health

David Clunie
Kristin Borradaile
Core Lab Partners, Princeton, NJ

Robert Ford
Princeton Radiology, Princeton, NJ

Hyun J. Grace Kim
Michael F. McNitt-Gray
David Geffen School of Medicine at UCLA, Los Angeles, CA

Binsheng Zhao
Columbia University Medical Center, New York, NY

Andrew J Buckler
Buckler Biomedical, Wenham, MA

http://dx.doi.org/10.6028/NIST.IR.7879

November 2012



U.S. Department of Commerce
Rebecca Blank, Acting Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary of Commerce for Standards and Technology and Director

Abstract

A study was performed under the auspices of Radiological Society of North America (RSNA) as a component of the Quantitative Imaging Biomarker Alliance (QIBA) to assess reader measurement variability of both spherical and complex (non-spherical) nodules sizing measures based on CT imaging scans. This paper reports the statistical data analysis of intra-reader and inter-reader variability of the three sizing measurements (1D, 2D, and 3D) performed as part of this collaborative effort. The data analysis strategy is based on careful graphical displays of measurement data and appropriate transformations of 1D, 2D and 3D data so that the performance of the three sizing measures can be compared on the same footing. An analysis of variance strategy is presented to analyze a well-designed experiment with complex factor settings and reader variability is defined. The general conclusion is that the 3D volume-derived diameter measure based on thin slice multi-detector CT has reduced bias and comparable variability over previous 1D and 2D sizing measures.

Keywords: Image biomarkers, shape and size measures, quantitative CT imaging, inter-reader and intra-reader variability, analysis of complex designed experiment.

1. Introduction

Computed Tomography (CT) has gone through rapid evolutions in the past 40 years and CT imaging has become a routine diagnostic tool in image-based medical practice [1]. Indeed, image biomarkers have been recognized as an important part of the biomarkers which have been considered by the industry and regulatory agencies as suitable parameters to evaluate clinical trials, with the goal of considerably speeding up the development of safe and effective medical therapies and procedures [2]. The de facto current standard, the Response Evaluation Criterion in Solid Tumors (RECIST) [3, 4] is based on one-dimensional sizing (1D diameter) alone (replacing the 2D area measure). Currently, the availability of high-resolution thin slice CT scans has made volume metric or volume-based change highly desirable surrogate endpoints in assessing therapy response [5, 6]. Recognizing these opportunities, the Radiological Society of North America (RSNA) at its annual meeting in 2007 formed the Quantitative Imaging Biomarker Alliance (QIBA) program to investigate quantitative imaging biomarkers in disease detection and responses to treatment and in which the measurement science plays an important role [7, 8]. It has been well documented that inter-reader variability is an important component of variability in nodule sizing measurements [9, 10, 11].

This paper presents the statistical data analysis of a large reader study of CT image data performed under QIBA to assess reader measurement variability of both spherical and complex non-spherical nodule sizing measures from CT images collected under various experimental settings that has not been reported in [12], and the contribution of this paper is the following. First, we emphasize data analysis based on careful graphical displays and innovative analysis of variance procedures for such complex multi-factorial experimental design data involving nodules of different size and shape. Secondly, we recommend appropriate transformations of the data so that the 1D, 2D and 3D data can be compared on the same footing across nodules of different size and shapes. Thirdly, we focus on quantifying the intra-reader and inter-reader variation in the context of taking into account of contribution of various other experimental factors to the measurement variability for sizing measurements in 1D, 2D and 3D sizing measures of simple and complex nodule shapes.

2. Data and Graphical Displays

Under the auspices of the QIBA program, a study on CT imaging measurement study was performed in which 6 radiologists measured the size of 10 synthetic nodules embedded within an anthropomorphic thorax phantom from CT scan data with multiple experimental factors. The synthetic nodules consist of five shape/size types: spherical (10, 20 mm) and non-spherical (20 mm elliptical, 10 mm lobulated, and 10 mm spiculated) and are made of two densities of -10 (Hounsfield Unit) HU and +100 HU (Table 1). Some examples of the non-spherical phantom nodules are shown in Figure 1 and we refer the full details on the FDA phantom to [13]. The experimental factors for

CT imaging collection include two repeat CT scans, two slice thickness of 0.8 mm and 5.0 mm. Six readers are recruited to use semi-automated computer assisted diagnostic tools to provide three sizing measurements, 1D longest in-slice dimension; 2D area from longest in-slice dimension and corresponding longest perpendicular dimension; 3D semi-automated volume, and there were two reading sessions for each reader. The resulting data are summarized in an Excel style sheet, with columns of nodule, density, scan, slice thickness, case number, reader, reading session, sizing technique, measured sizes, nominal (derived) size. The nominal size values were provided by ex vivo measurement method at FDA and are used here as reference values (Table 1). In order to compare data across 1D, 2D and 3D sizing measures and, in the literature it is often recommended to consider the relative size percent errors defined by:

$$y=100\times(\text{measured size} - \text{nominal size})/\text{nominal size}, \quad (1)$$

where measured size is the measured 1D, 2D, or 3D size value, while the nominal size is the nominal (reference) 1D, 2D or 3D size value on each of the 5 nodules given in Table 1.

Note that we may approximately define the (relative) bias[1] as the systematic departures of (1) from 0s and the variance as the variability around a common value (mean) at a given experimental condition. Figure 2 shows the 3D volume data, while Figure 3 and Figure 4 show the 1D and 2D data. Several important observations can be made based on the graphical analysis of data.

1. The 1D and 2D measurements suffer from under-estimation for non-spherical shape nodules, most noticeably nodule 3 (elliptical shape at one of the orientation) and nodule 5 (spiculated). Further illustration of the orientation effect on non-spherical nodules (nodules 3 and 5) is shown in Figure 10 of [12]. The volume measure is least affected by the non-spherical shape and is almost unbiased.
2. Other than vulnerability against bias for non-spherical nodules, 1D and 2D measurements have less variability than the 3D measurements, if we separate out the effect of orientation on the non-spherical nodules on the 1D and 2D data. Furthermore, it appears that the 3D measurement method is best suited for thin slice data, as there is more variability with 5.00 mm data than the 0.8 mm data.
3. The effect of nodule size and shape on the statistical characteristics of relative size measurements as defined in (1) is also apparent. For non-spherical nodules, in addition to potential bias due to positioning orientation, there is clear difference in variability across nodules, with smaller nodule (nodule 1) and complex shapes (nodule 4 and 5) having more measurement variability.

So the transformation (1) does not completely remove the effect of scale. Since in change analysis, it is the relative change in a chosen sizing measure that is of interest, one may argue that one can apply any monotone transformation on the 2D and 3D data so that they can be better compared with the 1D sizing measure. The natural transformation is to convert the 2D area S and 3D volume V into equivalent corresponding diameters R_2, R_3

[1] Strictly speaking, we cannot define the bias because we do not know the true size measure of a nodule. The nominal size values are reference values and are subject to their uncertainties.

of spheres with the same area or volume, as defined in the following *linearization transformations* [14]:

$$R_2 = (\frac{S}{4\pi})^{1/2}, R_3 = (\frac{3V}{4\pi})^{1/3}. \qquad (2)$$

After applying the transformations (2) to the 2D and 3D measurement data and nominal values, one can use (1) to plot the relative errors on all nodules.

Figure 5 through 7 show the data in relative errors after applying transformation (2), and note that Figure 6 duplicates Figure 3 except it is re-plotted with the same range as Figure 5 and 7 for easy comparison.

It appears that the transformations (2) make the 2D and 3D data more comparable to 1D data in the sense that the range of relative errors are *similar* in the linear scale (all withine ±20% range), if we single out the effect of orientation on the non-spherical nodules (nodules 3 and 5) on 1D and 2D measurements.

In the following sections, we see how the variances due to various experimental factors can be defined both graphically and quantitatively, and the different variance components that experimental factors including readers may contribute to the overall variability can be characterized as well. We emphasize that one should examine the individual nodule variances and to treat the nodule size and shape as *apriori* covariate factors before pooling them in some way to produce the overall variances.

3. Intra-reader Variability

Analogous to [15], an obvious way to define the intra-reader variability is the difference in observations by a single reader based on the same image over two different occasions or based on different images in which the nodule being observed is known to be unchanged in the chosen size measure. Note that by experimental design, each of the six readers has made two observations on two separate occasions on each image, and we use the difference between the two observations to assess the intra-reader variability. We consider the relative (percent) difference defined by

y=100×(measured size at time 1 – measured size at time 2)/nominal size value. (3)

To better compare data across different sizing scale, we also consider transformations of 2D and 3D data of (3) which convert the 2D (area) and 3D (volume) data into the linear (diameter) scale. Figures 8, 9, 11 show the intra-reader relative change for 1D, 2D and 3D data, while Figures 10, and 12 show corresponding plot using the transformed 2D and 2D data in equivalent spherical diameter scale.

There are several interesting observations:

1. Scanner slice thickness (5.0 mm) seems to have little effect on 1D and 2D intra-reader variability. But this is not the case for 3D volume data, where the thick slice (5.0 mm) data have more variability than the thin slice thickness (0.8 mm).

2. The scale of the data does have an important effect on the reader measurement variability, as the range of relative change in 2D data (Figure 9) is more than twice as large than that of 1D data (Figure 8), while the range of change in 3D data (Figure 11) is much larger than 2D and 1D, three times or more than that of 1D data, even in the case of thin slice data.
3. The linearization transformations (2) have the desired effect of making the 3D and 2D data more comparable to 1D data in terms of intra-reader variability, as shown in Figure 10 and Figure 12. Moreover, the effect of nodule shape and size is less pronounced in the transformed data scale, and we observe that the relative variability due to the intra-reader difference is around 10 % for 1D and 2D data, and is also around 10 % for the 3D volume data in the transformed diameter scale in the case of thin slice (0.8 mm) data.

4. Analysis of Variance

It should be made clear that the measurand is the size of a single nodule and standard analysis of variance method such as [16] is designed to study the effects of measurement and experimental factors on a single measurand. Since there are 10 nodules (5 identical nodules made of two densities and measured with two positioning orientations) under consideration, it is imperative that bias and variance associated with a measurement process should be associated with per nodules, and performance evaluation should be based on a per nodule basis before pooling and generalizing an inference on the population of potentially many nodules of different sizes and shapes can be made.

Ideally, the bias and variance of a measurement process should be characterized based on many repeated measurements. Our experimental data consist of: There are 480 observations in total, 48 observations for each of the 10 nodules, of 5 identical shape and size combinations made of two densities, with repeated scans of CT images at two CT slice thicknesses, and are measured by 6 readers repeated in two separate sessions. The design factors are clearly heavily nested [17]. One can define the various technical variations due to various experimental factors such as we have done for intra-reader variability in Section 3, in this section we focus on the more interesting situations: under experimental conditions that there is no underlying change in nodule shape or size (no change hypothesis), what is the expected measurement variability due to various experimental factors, including reader effects. We have seen that the two repeated scans clearly serve this purpose and so does the density factor for the 3D method and the slice thickness factor for 1D and 2D. (The effect of density on the 1D and 2D measurements is due to the different orientation, which is a critical factor for measuring non-spherical objects.) In a sense, we're looking for experimental conditions similar to the coffee-break experiments [18] in order to assess the measurement effects of interest.

Let y denote the relative errors defined as (1) for each observation and we focus on the 3D data first. The first step is to compute the relative bias, and this is given by computing the means (averages) of y for every combination of nodule, slice thickness and reader,

resulting in a 5x6x2 array data (the dimensions representing 5 nodules, 6 readers and 2 slice thickness). Note that, since the computation is performed on a per nodule basis, computing the sample means of *y* is equivalent to computing the bias from the volume measurements directly. That is, we can approximate the relative bias by:
Rel. Bias=100×(the observed averages of measured size – nominal size)/nominal size. (4)

Similarly, we can compute the measured intra-reader standard deviation for every combination of nodule, slice thickness and reader, resulting in another 5x6x2 array. These two arrays are the stage two data summaries, which form the basis for our final statistical performance metrics to be defined later on. Figure 13 and Figure 14 show the reader relative bias plot and intra-reader standard deviation (as a percentage of the nominal size value), respectively. Note that in the plots, the x-axis representing the data sequence in the order of data for nodules 1,2,3,4,5 and then repeat at another image setting. It is interesting to observe that it appears that readers 4,5, 6 tend to underestimate while readers 1,2,3 tend to overestimate. Also there are more bias and intra-reader variability with the 5.0 mm slice thickness data, and clearly the larger nodules (nodule 2 and nodule 3) have less bias and intra-reader variability.

At the final stage of analysis, performance metrics need to be defined which, hopefully, do not have to depend on individual nodules, and which should be as simple as possible, providing a single or a few numbers which can summarize the overall reader measurement performance. First, because the bias is the systematic behavior of a measurement process, we recommend that, in order to combine bias across readers, one can take the absolute values and then take averages or even maximum across readers. One can also compute the standard deviation of individual reader biases, resulting in inter-reader variability[2]. These are given in Table 2. It should be pointed out that all these computations are done on a per nodule basis, but one can easily combine these metrics across different nodules if they are deemed similar, or comparable. Indeed, it clearly shows that nodule 2 and nodule 3 are similar and could be grouped together. If one does not care about the specific shape or size effect and is only interested in the overall performance, one can either compute the average across the nodules or even the maximum. These final summary statistics for the relative bias are given in the last row of Table 2. The intra-reader variability can be easily combined across readers by taking the means of individual variances and then take the square root. The results are given in the last two columns of Table 2.

In conclusion, Table 2 summarizes the right level of data reduction and summary performance metrics that a decision can be based upon. It shows that for the case of 0.8 mm data, reader bias and standard deviation are all within 15% of the nominal volume, and for the optimistic cases of nodules (with simple shape and large size), the uncertainty is reduced to less than 8 %.

The same analysis strategy can be applied directly to 1D and 2D data, except that the significant factor density should be replaced by slice thickness. For both 1D and 2D data, the relative bias can be up to -60 % for some non-spherical nodules, while the relative

[2] Strictly speaking, one should adjust for a small factor due to the intra-reader variability, see e.g. [19].

standard deviation ranges from 5 % to 12 % for 1D, and 10 % to 15 % for 2D for nodules from simple to complex, and [12] provides more detailed comparison results. In summary, we conclude that the 3D method do not suffer from the underestimation in the 1D and 2D method for complex (non-spherical) nodules while still exhibiting comparable variability in the case of thin slice data.

5. Discussion

We demonstrated that with the availability of thin slice (0.8mm or less) CT scan image data, the 3D (volume) measure is very promising as the most reliable sizing method for complex nodule shapes in comparison to traditional 1D and 2D measures. For thick slice data, 1D and 2D measures still have some advantages as having the less variability and less dependence on slice thickness. For this synthetic nodules considered, we demonstrate that there is very good repeatability and reproducibility between and within the readers for the volume metric measurement, and the measurements of 3D (volume) measurements are within 20% of the nominal values for even complex (non-spherical) nodules. For future work, we think more extensive study in the factorial settings of nodule size and nodule shape is desirable, and the connection between nodule size analysis and clinically significant tumor change analysis [20] should be further studied.

Acknowledgement

We thank CoreLab Partners Inc[3] for support and participating in the reader study performed in this work. We thank the QIBA Volumetric CT Technical Committee, especially the members of the QIBA Volumetric CT Part 1A subcommittee for contributions in this work. We thank Adam Pintar and Jeeseong Hwang of NIST for careful reviewing of this paper.

REFERENCES

[1] Jiang Hsieh. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances.* Second Edition, Wiley-Interscience and SPIE Press, 2009.

[2] J. J. Smith, A. G. Sorensen, J. H. Thrall, "Biomarkers in Imaging: Realizing Radiology's Future." *Radiology,* June 2003, 227, pp.633-638.

[3] P. Therasse, P., S.G. Arbuck, E.A. Eisenhauer, J. Wanders, R.S. Kaplan, L. Rubinstein, J. Verweij, M.V. Glabbeke, A.T. van Oosterom, M.C. Christian, S.G. Gwyther, "New Guidelines to Evaluate the Response to Treatment in Solid Tumors,"

[3] Commercial service was identified in this paper to foster understanding. Such identification does not imply recommendation nor endorsement by the National Institute of Standards and Technology, nor does it imply that the service is necessarily the best available for the purpose.

Journal of the National Cancer Institute, Vol.92 , No.3, February 2, 2000, pp.205-216. doi: 10.1093/jnci/92.3.205.

[4] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, J. Verweij, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)." *European Journal of Cancer*, Vol. 45, 2009, pp.228-247.

[5] P.D. Mozley, L.H. Schwartz, C. Bendtsen, B. Zhao, N. Petrick, A.J. Buckler "Change in lung tumor volume as a biomarker of treatment response: a critical review of the evidence. " Ann Oncol. 2010;21(9):1751-5.

[6] C. R Meyer, S. G Armato, III, C. P Fenimore, G. McLennan, L. M Bidaut, D. P Barboriak, M. A Gavrielides, E. F Jackson, M. F McNitt-Gray, P. E Kinahan,N. Petrick,and B. Zhao, "Quantitative Imaging to Assess Tumor Response to Therapy: Common Themes of Measurement, Truth Data, and Error Sources." *Translational Oncology*, Vol. 2, No.4, December 1, 2009, pp.198–210.

[7] Radiological Society of North America (RSNA). Quantitative Imaging Biomarkers Alliance (QIBA). Quantitative imaging and imaging biomarkers. http://www.rsna.org/research/qiba.cfm. 2010.

[8] A.J. Buckler. 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 DIA Bethesda, MD Oct 2-3. 2008.

[9] J. J. Erasmus, G.W. Gladish, L. Broemeling, B.S. Sabloff, M.T. Truong, R.S. Herbst, and R. F. Munden." Interobserver and Intraobserver Variability in Measurement of Non-Small-Cell Carcinoma Lung Lesions: Implications for Assessment of Tumor Response." *Journal of Clinical Oncology*, Vol.21, No.13, July 1, 2003, pp. 2574-2582.

[10] L.E. Dodd, R.F. Wagner, S.G. Armato III, M.F. McNitt-Gray, S. Beiden, H-P CHAN, D. Gur, G. McleNnan; C.E. Metz, N. Petrick; B. Sahiner, J. Sayre, "Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the Lung Image Database Consortium." *Academic Radiology, Vol.* 11, No. 4, 2004, pp. 462-475

[11] C.R. Meyer, T.D. Johnson, G. McLennan, D.R. Aberle, E. A. Kazerooni, H. MacMahon, B. F. Mullan, D.F. Yankelevitz, E. J.R. van Beek, S.G. Armato III, M.F. McNitt-Gray, A.P. Reeves, D. Gur, C.I. Henschke, E.A. Hoffman, R. H. Bland, G. Laderach, R. Pais, D. Qing, C. Piker, J. Guo, A. Starkey, D. Max, B.Y. Croft, L.P. Clarke, " Evaluation of Lung MDCT Nodule Annotation Across Radiologists and Methods." *Academic Radiology*, Vol.13, No.10, October, 2006, pp.1254-1265.

[12] N. Petrick, H.J. G. Kim, D. Clunie, K. Borradaile, R. Ford, R. Zeng, M.A. Gavrielides, M.F. McNitt-Gray, Z.Q.J. Lu, C. Fenimore, B. Zhao, A.J. Buckler. "Evaluation of 1D, 2D and 3D nodule size estimation by radiologists for spherical and complex nodules through CT thoracic phantom imaging." *Manuscript submitted for publication*.

[13] M.A. Gavrielides, L.M. Kinnard, K.J. Myers, J. Peregoy, W. F. Pritchard, R. Zeng, J. Esparza, J. Karanian, and N. Petrick. "A Resource for the Assessment of Lung Nodule Size Estimation Methods: Database of Thoracic CT Scans of an Anthropomorphic Phantom." *Optics Express*, Vol.18, No.14, July 5, 2010, pp. 15244-15255.

[14] G. Pólya, G. Szegö. *Isoperimetric Inequalities in Mathematical Physics.* Princeton University Press, Princeton, 1951. (P. 3)

[15] D.G. Altman and J.M. Bland, Measurement in Medicine: the Analysis of Method Comparison Studies. *The Statistician* 32 (1983), pp. 307-317.

[16] H. Scheffe, *The Analysis of Variance.* John Wiley & Sons, New York, 1950.

[17] D.R. Cox, *Planning of Experiments*. John Wiley & Sons, New York, 1958.

[18] B. Zhao, L.P. James, C.S. Moskowitz, P. Guo, M.S. Ginsberg, R.A. Lefkowitz, Y. Qin, G.J. Riely, M.G. Kris, L.H. Schwartz, "Evaluating Variability in Tumor Measurements from Same-day Repeat Scans of Patients with Non-Small Cell Lung Cancer," *Radiology*, Vol.252, No.1, July 2009, pp.263-272.

[19] P.S.R.S. Rao, *Variance Components Estimation: Mixed Models, Methodologies and Applications.* Chapman & Hall, London, 1997, P.12.

[20] Z.Q. J. Lu, C. Fenimore, R.H. Gottlieb, C.C. Jaffe, "An Empirical Bayes Approach to Robust Variance Estimation: a Statistical Proposal for Quantitative Medical Imaging Testing." *Open Journal of Statistics*, Vol.2, No.3, July 2012, 260-268.
DOI: 10.4236/ojs.2012.23031

Table 1. Five Nodule Types: nominal size values* provided for references

Nodule	Volume (mm^3)	Area (mm^2)	1D (diameter, mm)
1: 10mm, sphere	522.41	100.11	10.01
2: 20mm, sphere	4259.41	406.42	20.16
3: 20mm, elliptical	4261.84	506.82	31.74
4: 10mm, lobulated	527.03	148.49	12.89
5: 10mm, spiculated	526.67	340.86	22.58

*Provided based on derived ex vivo measurements at FDA.



Figure 1. Figure showing FDA non-spherical phantom nodules (courtesy of Marios Gavrielides, Center for Devices & Radiological Health, US Food & Drug Administration).

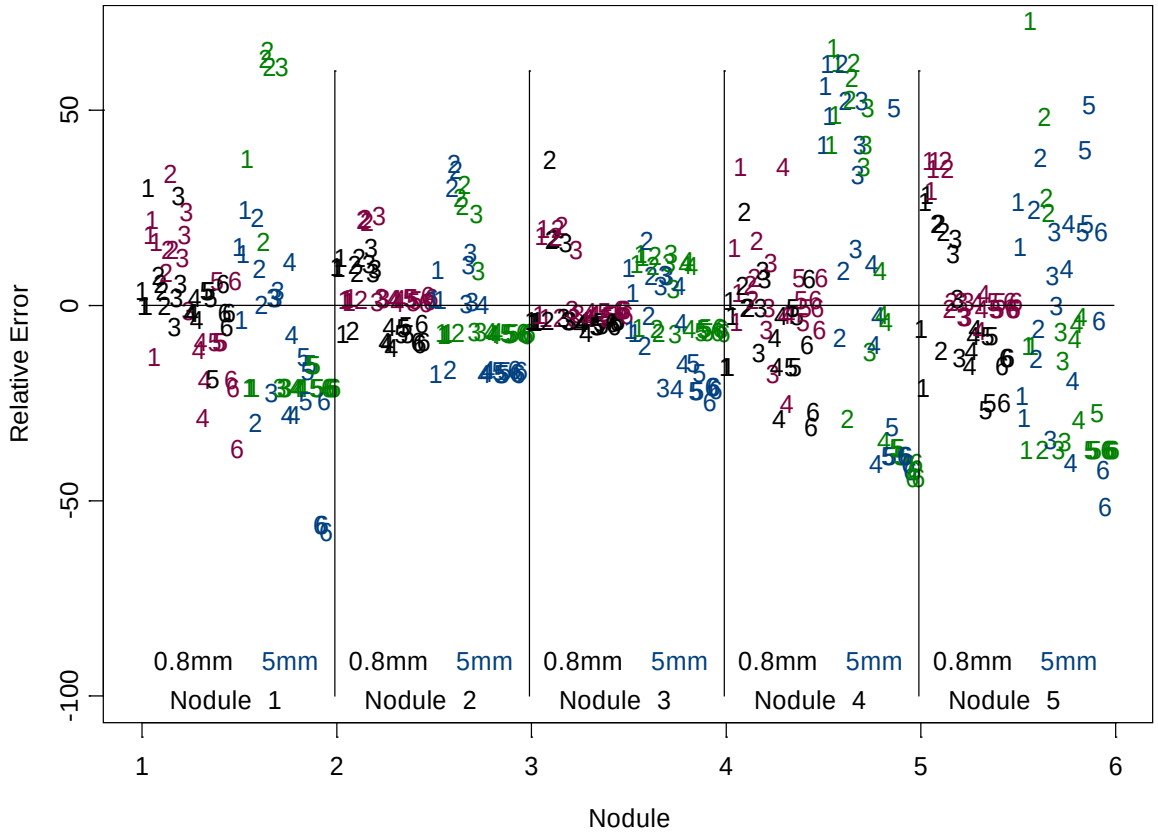


Figure 2. Plot of relative percent errors in volume measurements for all five nodules. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors indicate phantom density (-10, black or blue, 100, pink or green) at two slice thickness (0.8 mm, black or pink, 5.0 mm, blue or green). The data sequence is in the order of nodules 1, 2, …, and 5.

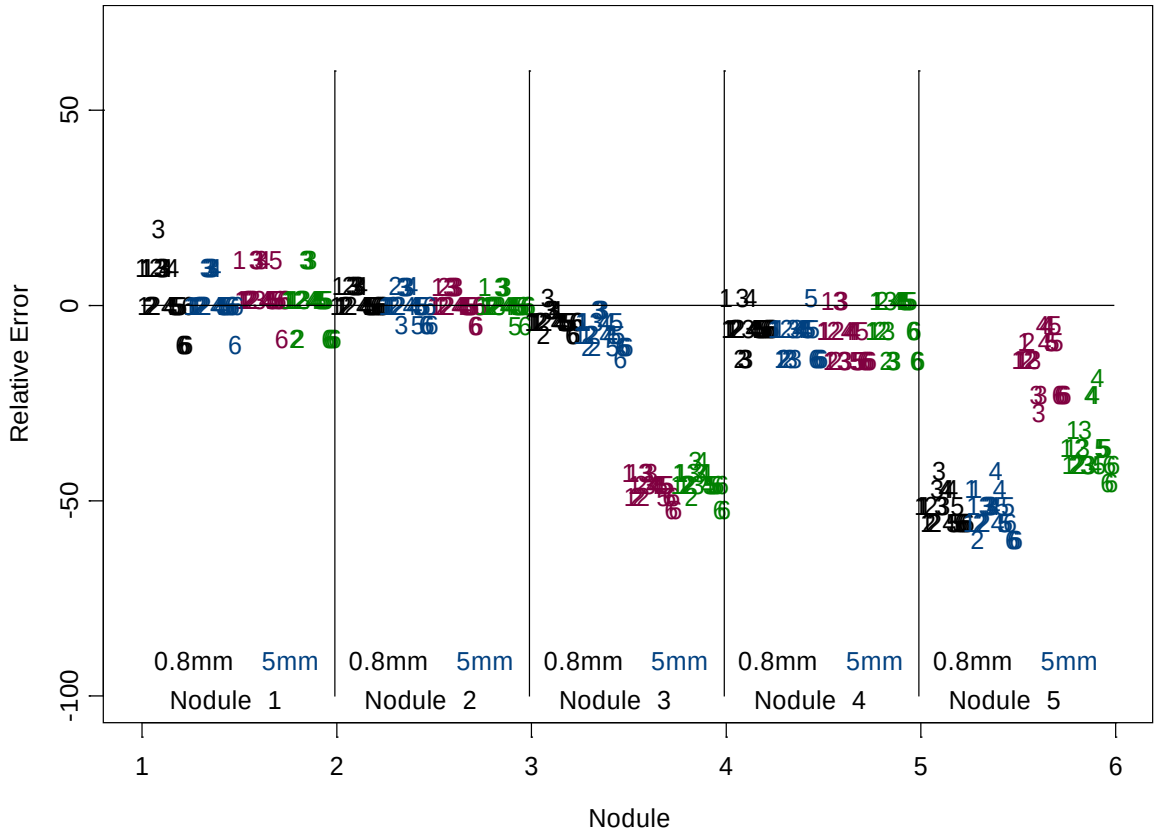


Figure 3. Plot of relative percent errors in 1D (diameter) measurements for all five nodules. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors indicate phantom density (-10, black or blue, 100, pink or green) at two slice thickness (0.8 mm, black or pink, 5.0 mm, blue or green). The data sequence is in the order of nodules 1, 2, …, and 5.

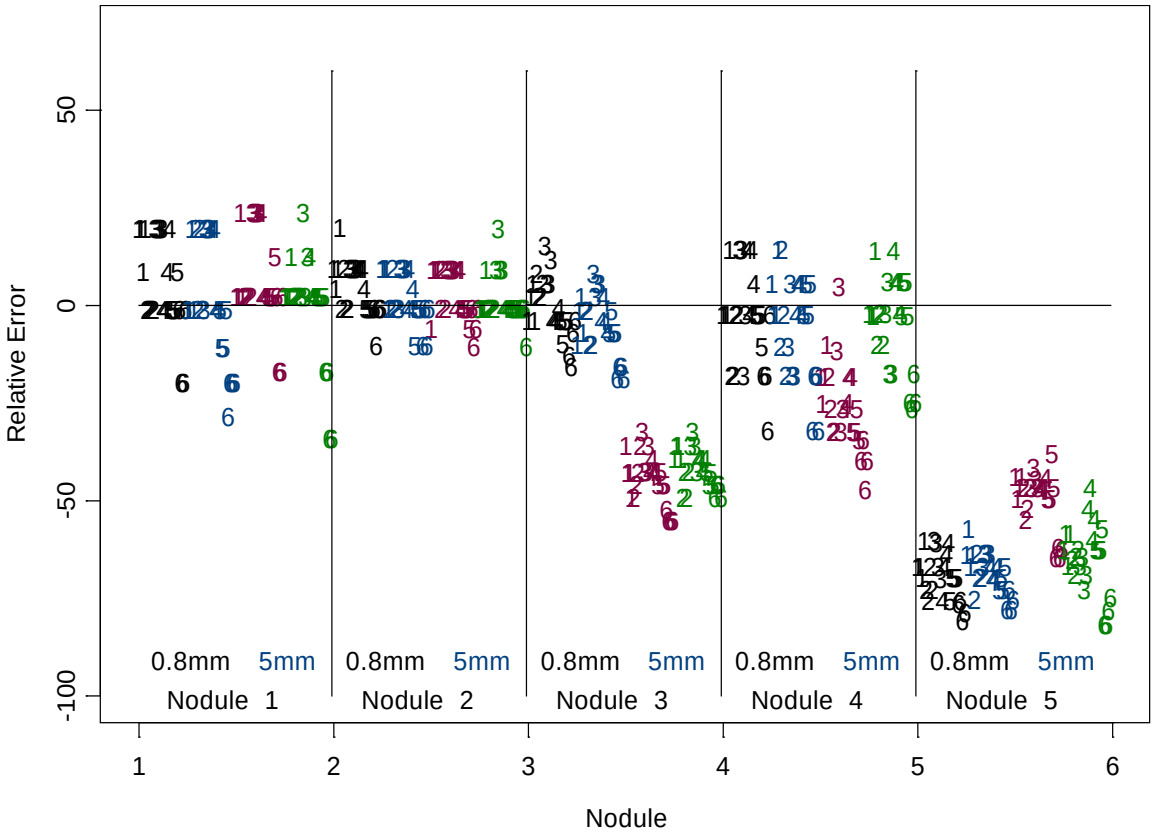


Figure 4. Plot of relative percent errors in 2D (area) measurements for all five nodules. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors indicate phantom density (-10, black or blue, 100, pink or green) at two slice thickness (0.8 mm, black or pink, 5.0 mm, blue or green). The data sequence is in the order of nodules 1, 2, …, and 5.

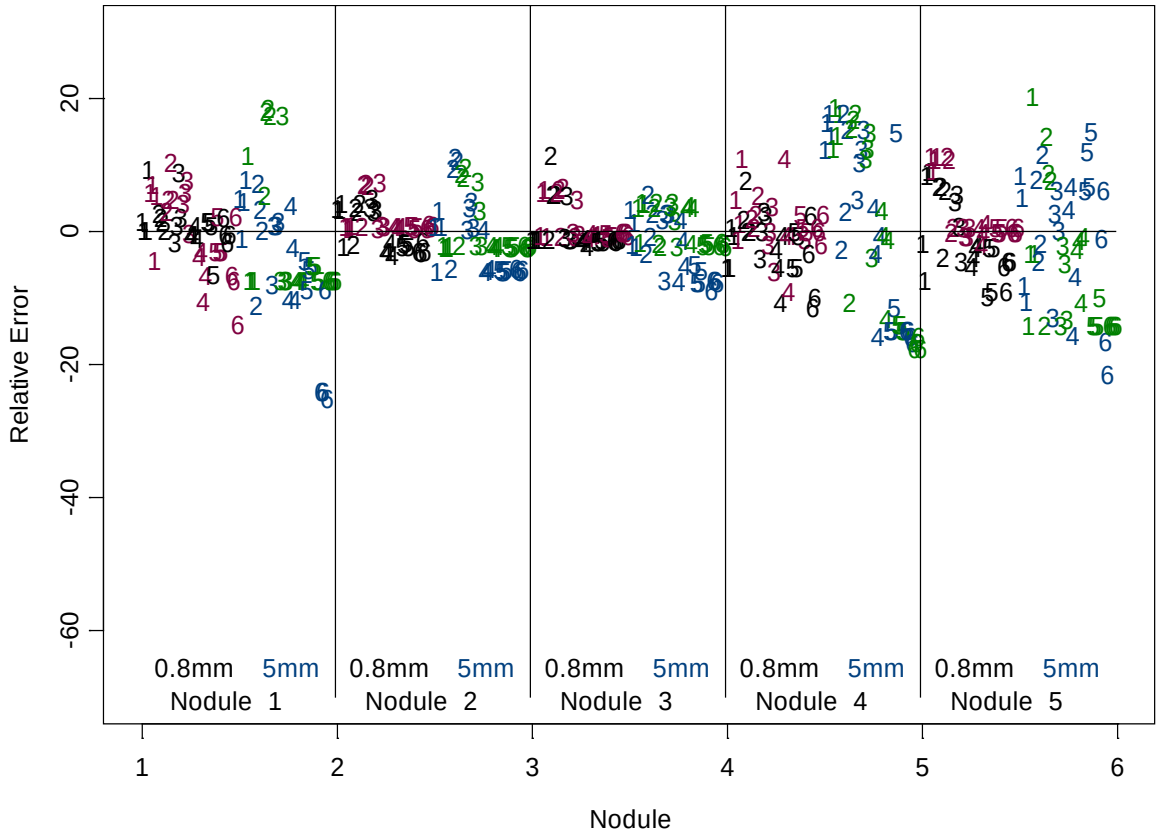


Figure 5. Plot of 3D data in relative percent errors after both the observed volume and nominal volume are converted in the linear scale as defined in (2). The colors and symbols are the same as defined in Figure 2.

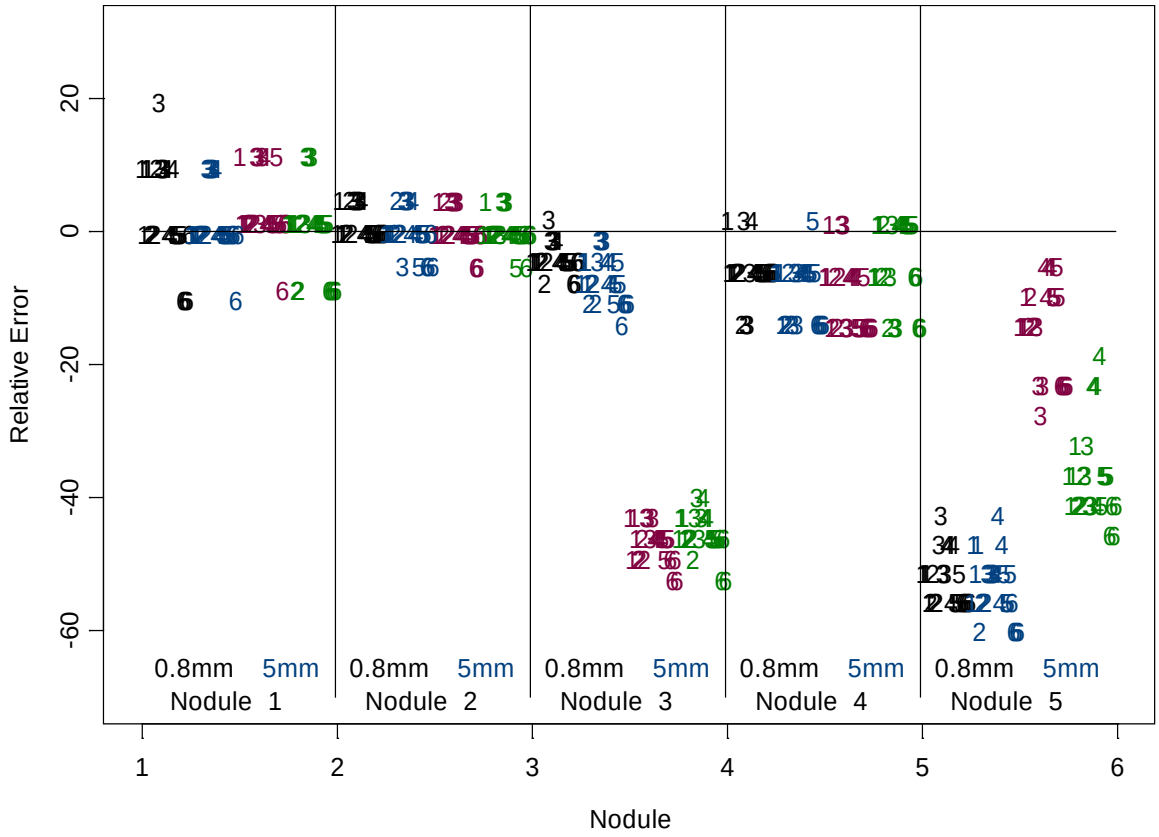


Figure 6. Plot of the 1D data in relative percent errors. This is almost the same figure as Figure 3 except the slight change in data range in the y-axis.

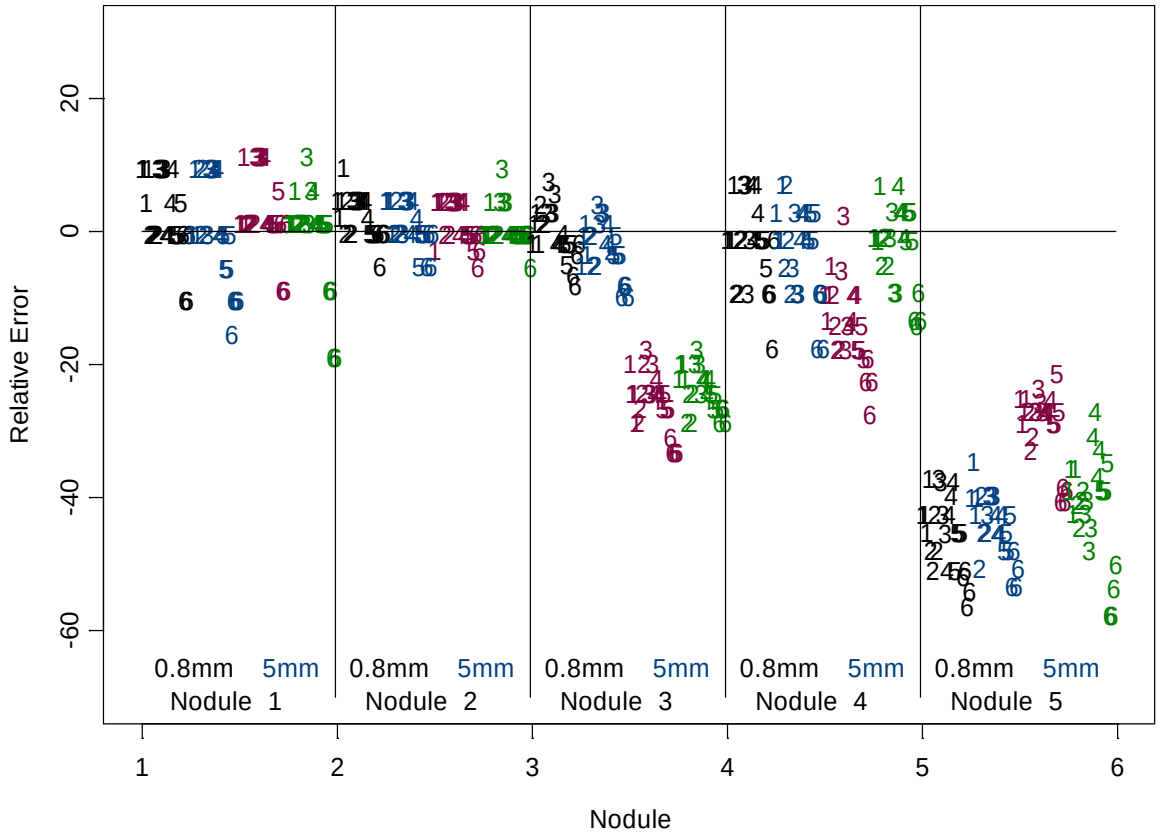


Figure 7. Plot of 2D data in relative percent errors after both the observed 2D and nominal 2D data are converted in the linear scale as defined in (2). Symbols and colors same as in Figure 4.

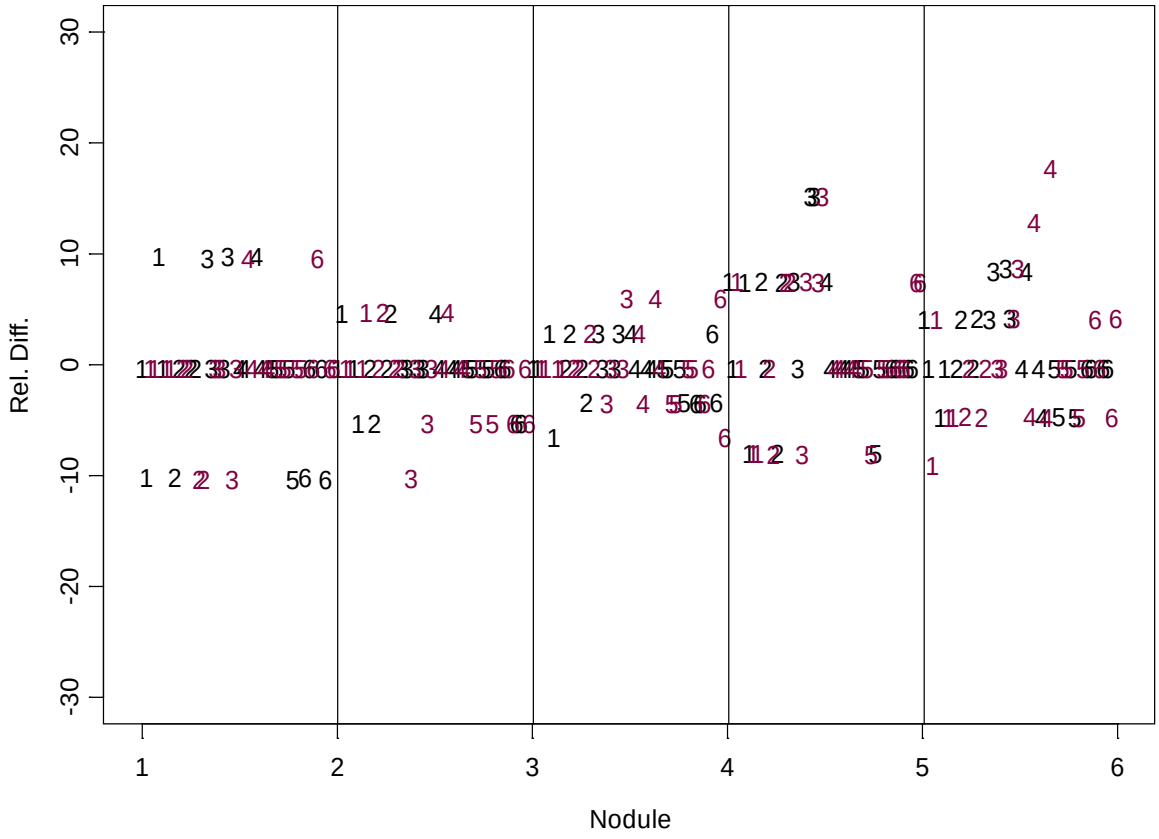


Figure 8. Plot of relative changes in 1D (diameter) measurements (by the same reader on two occasions) for all five nodules. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors at two slice thickness (0.8 mm, black, 5.0 mm, pink). The data sequence is in the order of nodules 1, 2, …, and 5.

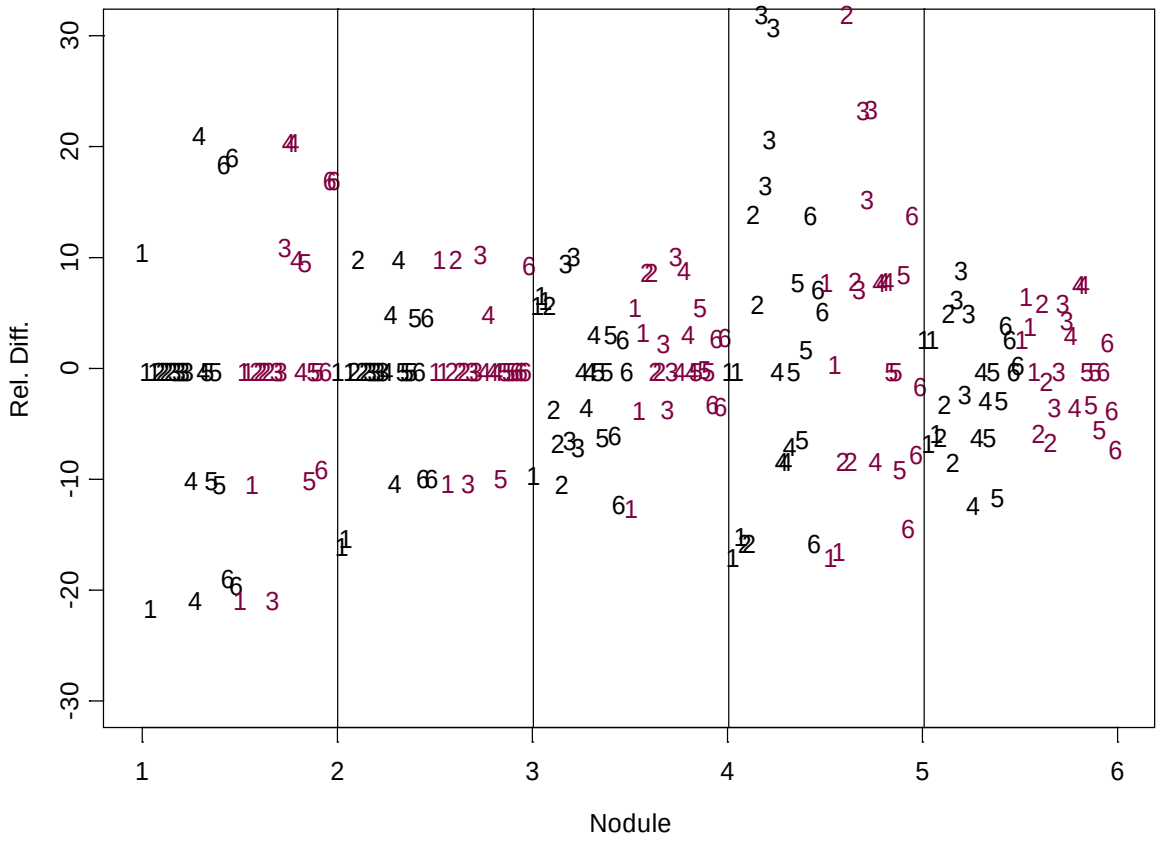


Figure 9. Plot of relative changes in 2D (area) measurements (by the same reader on two occasions) for all five nodules. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors at two slice thickness (0.8 mm, black, 5.0 mm, pink). The data sequence is in the order of nodules 1, 2, …, and 5.

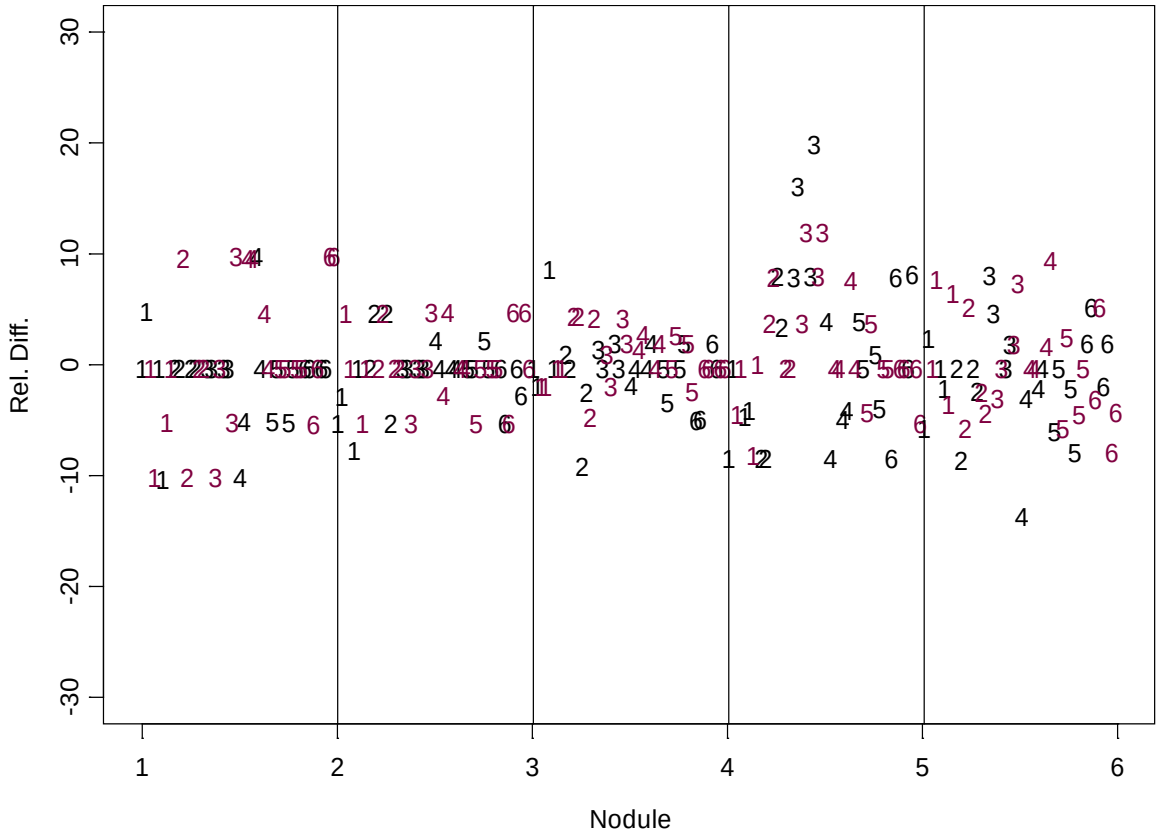


Figure 10. Plot of relative changes in 2D intra-reader measurements in the linear scale after the linearization transformation (2) (by the same reader on two occasions) for all five nodules. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors at two slice thickness (0.8 mm, black, 5.0 mm, pink). The data sequence is in the order of nodules 1, 2, …, and 5.

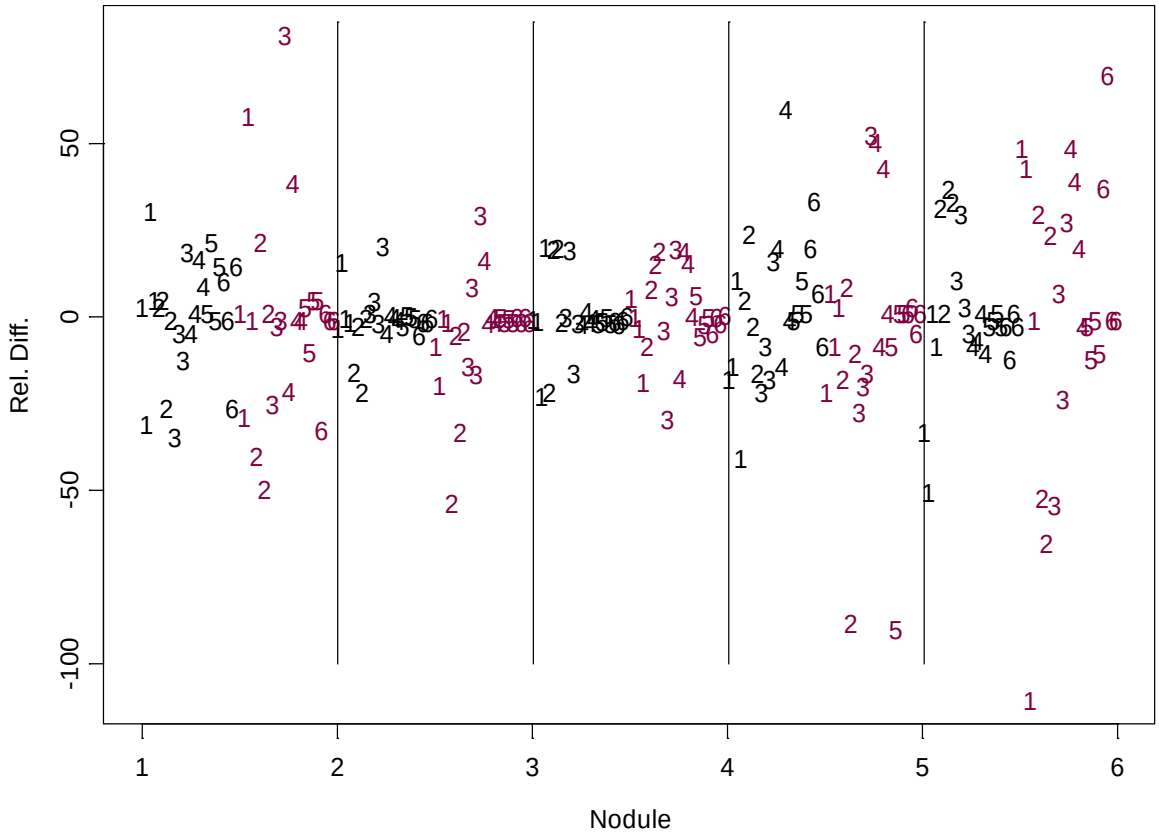


Figure 11. Plot of relative changes in 3D (volume) measurements (by the same reader on two occasions) for all five nodules. (Note this figure has a much larger scale in the y-axis than comparable Figures 8-10 and 12). Readers are coded by 1, 2, 3, 4, 5, 6, and the colors at two slice thickness (0.8 mm, black, 5.0 mm, pink). The data sequence is in the order of nodules 1, 2, …, and 5.

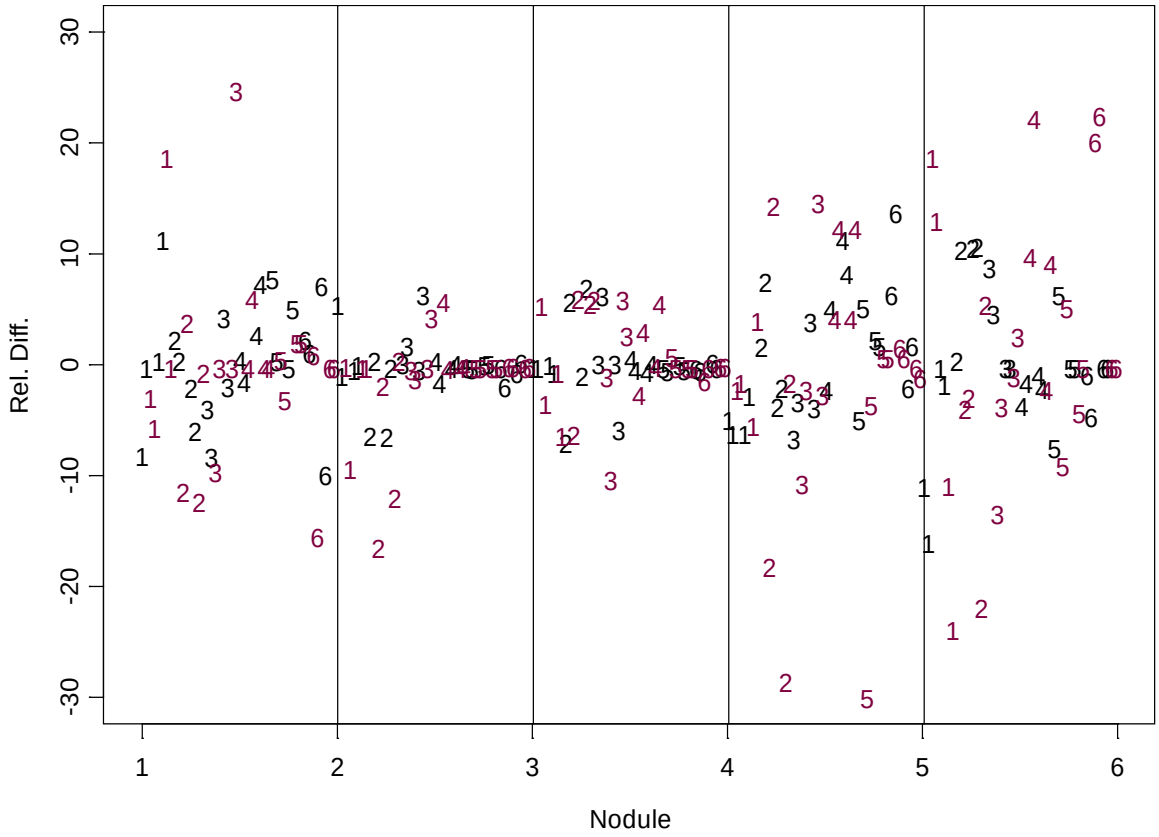


Figure 12. Plot of relative changes in 3D (volume) measurements in the linear scale after the linearization transformation (2) (by the same reader on two occasions) for all five nodules. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors at two slice thickness (0.8 mm, black, 5.0 mm, pink). The data sequence is in the order of nodules 1, 2, …, and 5.

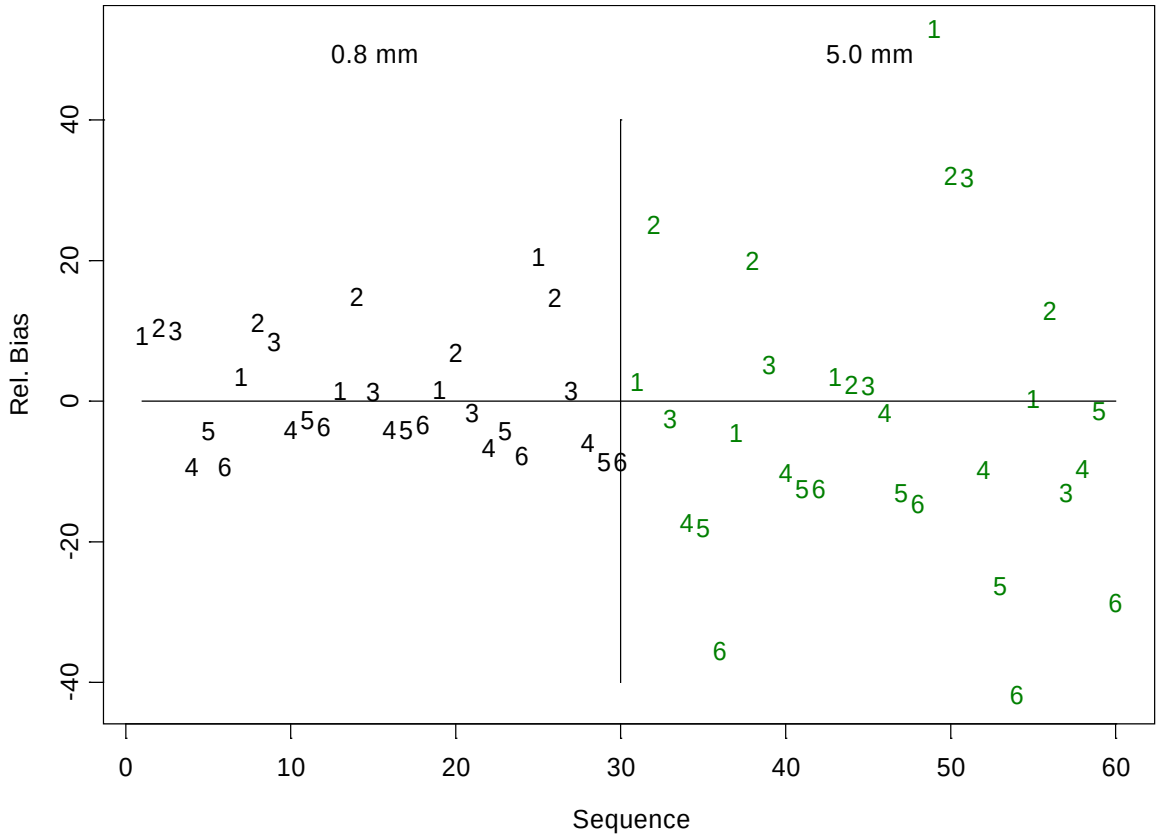


Figure 13. Plot of measured volume relative bias: 100*(measured means-nominal volume)/nominal volume. The data sequence is in the order of nodules 1, 2, …, 5, then repeat at another image setting. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors indicate the two slice thickness (0.8 mm is in black, 5.0 mm in green).

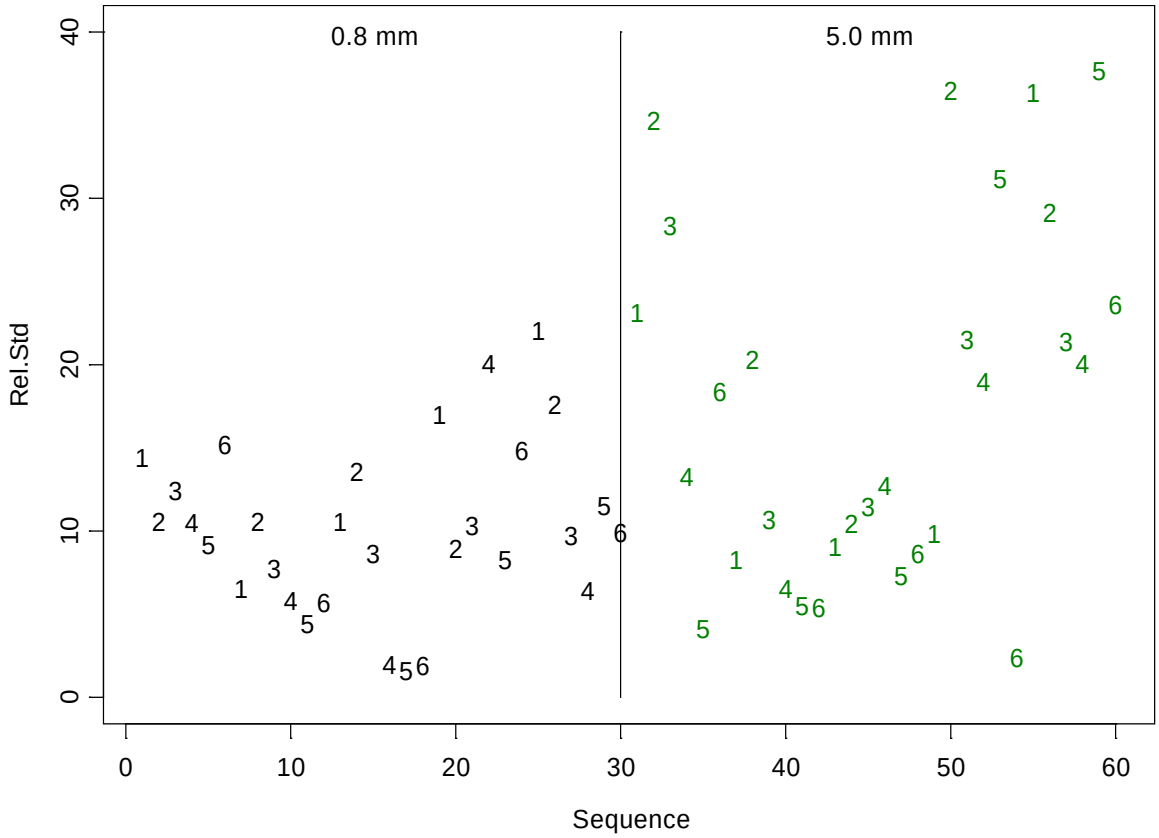


Figure 14. Plot of measured volume intra-reader standard deviation: 100×measured stdev/nominal volume. The data sequence is in the order of nodule 1, 2, …, and 5, then repeat at another image setting. Readers are coded by 1, 2, 3, 4, 5, 6, and the colors indicate the two slice thickness (0.8 mm is in black, 5.0 mm in green).

Table 2: Summary of reader relative bias and standard deviation for 3D (volume) data: the second and third columns are the average absolute reader bias, third and fourth are the maximum absolute reader bias, and the sixth and seventh columns are the inter-reader standard deviation defined as the sample standard deviation across individual reader bias. The eighth and ninth columns are the intra-reader standard deviation.

Nodules	Individual Mean Bias		Individual Maximum Bias		Inter-reader variability (stdev)		Intra-reader variability (stdev)	
	0.8 mm	5.0 mm	0.8 mm	5.0 mm	0.8 mm	5.0 mm	0.8 mm	5.0 mm
1	8.8	17.0	10.1	35.9	9.8	20.9	12.1	22.5
2	5.6	10.9	10.7	19.5	6.6	12.8	6.9	10.6
3	4.9	6.2	14.5	15.0	7.3	8.1	7.8	9.9
4	5.0	32.4	8.2	52.5	5.5	37.7	13.8	23.0
5	10.0	11.1	20.1	29.0	12.6	14.1	13.8	28.7
Overall	**6.8**	**15.5**	**20.1**	**52.5**	**8.7**	**21.3**	**11.2**	**20.4**