Department of Radiology, Sir Run Run Hospital Affiliated with Nanjing Medical University, [1]Department of Radiology, BenQ Medical Center, Nanjing Medical University, [2]Department of Cell Biology, Collaborative Innovation Center for Cancer Personalized Medicine, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Cancer Center, Key Laboratory of Human Functional Genomics of Jiangsu Province, Nanjing Medical University, Nanjing, China

Systematic analysis of measurement variability in lung cancer with multidetector computed tomography

Binghu Jiang, Dan Zhou[1], Yujie Sun[2], Jichen Wang[1]

Address for correspondence:
Prof. Yujie Sun, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, No.101, Longmian Avenue, Jiangning District, 211166, Nanjing, China.
E-mail: yujiesun@njmu.edu.cn
Dr. Jichen Wang, Department of Radiology, BenQ Medical Center, Nanjing Medical University, No.71, Hexi Street, Jianye District, 210019, Nanjing, China.
E-mail: fskwjc@126.com

Access this article online

Quick Response Code:

Abstract:

OBJECTIVE: To systematically analyze the nature of measurement variability in lung cancer with multidetector computed tomography (CT) scans.

METHODS: Multidetector CT scans of 67 lung cancer patients were analyzed. Unidimensional (Response Evaluation Criteria in Solid Tumor criteria), bidimensional (World Health Organization criteria), and volumetric measurements were performed independently by ten radiologists and were repeated after at least 5 months. Repeatability and reproducibility measurement variations were estimated by analyzing reliability, agreement, variation coefficient, and misclassification statistically. The relationship of measurement variability with various sources was also analyzed.

RESULTS: Analyses of 69 lung tumors with an average size of 1.1–12.1 cm (mean 4.3 cm) indicated that volumetric technique had the minimum measurement variability compared to the unidimensional or bidimensional technique. Tumor characteristics (object effect) could be the primary factor to influence measurement variability while the effect of raters (subjective effect) was faint. Segmentation and size in tumor characteristics were associated with measurement variability, and some mathematical function was established between the volumetric variability and tumor size.

CONCLUSION: Volumetric technique has the minimum variability in measuring lung cancer, and measurement variability is associated with tumor size by nonlinear mathematical function.

Key words:
Computed tomography, lung cancer, measurement variability

Tumor imaging plays a fundamental role in clinical care and trials of lung cancer where computed tomography (CT)-based tumor measurement is the preferred technique.[1] Compared with the World Health Organization (WHO) criteria, the Response Evaluation Criteria in Solid Tumors (RECIST) shows a better result in determining response to therapy.[2,3] Recently, volume technique obtained with automated segmentation tool improves accuracy of assessment.[4,5]

Owing to measurement variability, however, measurements of lung tumor size on CT scans are often inconsistent and can lead to an incorrect interpretation of tumor growth or response. Although a number of significant factors leading to measurement variability have been documented,[4-13] those of primary importance and the quantitative relationship between those potential factors and variability have yet to be determined. The purpose of this study was to systematically analyze the measurement variability in CT interpretation in cases of nonsmall cell lung cancer.

Methods

This retrospective study was approved by our institutional ethics committee, and the requirement for informed consent was waived because of the retrospective nature.

Patients' characteristics

Patients were identified in the Picture Archiving and Communication System from January 2014 to December 2015. All the identified patients had solid pulmonary nodules or masses diagnosed as nonsmall-cell lung cancer by biopsy or surgical specimen, and all tumors were imaged by CT with 2.0 mm or thinner collimation.

We identified 67 patients with 69 lung tumors, including 20 women and 47 men with mean age

of 67.1 years ±12.2 standard deviation (SD). Sixty-five patients had one focus each, and two patients had two foci each.

Computed tomography data acquisition
Patients underwent imaging using a 64-detector CT scanner (LightSpeed VCT, GE Healthcare, USA, Chicago, IL) with 64 mm × 0.625 mm collimation and a 16-detector CT scanner (Sensation 16, Siemens Medical Systems, Forchheim, Germany) with 16 mm × 0.75 mm collimation. Scans were obtained with the patients at full inspiration. Exposure settings were 50–80 mAs at 120 kVp. Axial images of 1.25 mm or 1.5 mm thickness were reconstructed with 512 × 512 matrix. Air calibration was conducted every morning before CT scanning.

Tumor measurement
Pulmonary tumors were analyzed independently by 10 raters (with 2–10 years of experience in radiology, respectively) on a workstation (Leonardo; Siemens Medical Systems) using a lung window (width, 1500 HU; center, −500 HU), and if necessary, the window settings were allowed to be changed. After instruction to measure tumors on preselected images, the raters performed measurements on transverse slices using a digital caliper according to the RECIST and WHO criteria and obtained the volume of each tumor using the computer-aided semi-automated evaluation software (LungCare; Siemens Medical Solutions). Four measurements were generated: longest diameter on native axial slice (RECIST criteria), longest perpendicular diameter in the same image, product of these two diameters (WHO criteria), and volumetric quantification of the tumor. The raters were not aware of each other's selected slices. At least 5 months later, a duplication of measurement procedure was performed by each rater for all tumors.

In addition, two experienced raters (D. Z. and B. J., with 20 and 10 years of experience in radiology, respectively) visually assessed tumor morphological characteristics by consensus. Moreover, four subgroups were generated: regular group (well-defined boundary) versus irregular group (undefined boundary) and isolated group (nearly no interface between tumor and adjacent structures) versus nonisolated group (interface ≥45°).

Statistical analysis
Statistical analysis was performed with the SPSS software (PASW Statistics 18; SPSS Inc., Chicago, IL, USA) and two-tailed $P < 0.05$ was considered statistically significant. The required sample size to detect a significant association at $\alpha = 0.05$ and with a power of 90% was estimated to be 60. Continuous variable is expressed as mean ± SD.

We estimated the intraobserver reliability with formula of (between_subject SD^2 + between_observer SD^2)/(between_subject SD^2 + between_observer SD^2 + measurement_error SD^2) and interobserver reliability with formula of (between_subject SD^2)/(between_subject SD^2 + between_observer SD^2 + measurement_error SD^2), which are the mathematical derivation of equation of (SD of subject's true values)^2/([SD of subject's true values]^2 + [SD of measurement error]^2) by Bartlett and Frost,[14] and the agreement by Bland–Altman plots. The variation coefficient (VC), defined as the ratio of the SD to the mean, was also calculated. The variation sources of the tumor measurements were modeled with the analysis of variance.[7] We also explored the relationship between measurement variability and potential factors by curve estimation.

Results

Tumor size ranged from 1.1 cm to 12.1 cm (mean, 4.3 cm) by unidimensional measurements, 1.1 to 104.9 cm^2 (mean, 19.3 cm^2) by bidimensional measurements, and 0.6 to 553.4 cm^3 (mean, 66.2 cm^3) by volumetric measurements [Table 1].

Misclassification rates
Because of unavailable criteria for volumetric technique at present, we used RECIST criteria as the reference for volumetric measurement. Misclassification rates demonstrated the potential impact of measurement variability. For each rater and each tumor, the difference between the smallest and largest measurement was computed. All measurement differences were assessed relative to the smaller measurement using RECIST and WHO criteria for progressive disease (RECIST >20% and WHO >25%) and relative to the larger measurement using criteria for response (RECIST >30% and WHO >50%). A misclassification was recorded in each group if the relative change exceeded these criteria. For inter-rater misclassification, only the first replication was used for this estimate. Volumetric technique showed the lowest misclassification rates [Table 2].

Agreement and reliability
For the repeatability (intra-rater) study, the 95% limits of agreement varied from −12.1 mm (−26.9%) to 12.9 mm (28.9%) for unidimensional, −984.0 mm^2 (−45.1%) to 960.3 mm^2 (47.6%) for bidimensional, and −6666.4 mm^3 (−11.2%) to 7221.8 mm^3 (11.6%) for volumetric measurement [Table 1]. The significant difference was found among RECIST versus WHO ($P < 0.001$), RECIST versus volume ($P < 0.001$), and WHO versus volume ($P < 0.001$), respectively. For the reproducibility (inter-rater) study, the 95% limits of agreement varied from −13.7 mm (−31.2%) to 13.9 mm (31.2%) for unidimensional, −1095.0 mm^2 (−52.4%) to 1153.4 mm^2 (53.6%) for bidimensional, and −19593.2 mm^3 (−23.9%) to 22622.5 mm^3 (25.8%) for volumetric measurement. The significant difference was found among RECIST versus WHO ($P < 0.001$), RECIST versus volume ($P < 0.001$), and WHO

Table 1: Results from tumor measurements

	RECIST (mm)	WHO (mm²)	Volume (mm³)
Minimum	11.2	111.4	641.6
Maximum	121.1	10,494.4	553,448.8
Mean	43.0	1925.7	66,216.8
Repeatability			
Agreement (95% CI)	−26.9%, 28.9%	−45.1%, 47.6%	−11.2%, 11.6%
Reliability	0.998	0.998	1.000
VC (%)	6.7	11.4	2.9
Reproducibility			
Agreement (95% CI)	−31.2%, 31.2%	−52.4%, 53.6%	−23.9%, 25.8%
Reliability	0.971	0.982	0.997
VC (%)	7.8	13.6	6.7

CI = Confidence interval, VC = Variation coefficient, RECIST = Response Evaluation Criteria in Solid Tumors, WHO = World Health Organization

Table 2: Measurement variability and the corresponding misclassification

	Unidimensional	Bidimensional	Volumetric*
Variability of measurements (%)			
Intra-rater (10 pairs)			
Minimum (variability)	0	0	0
Maximum (variability)	77.1	139.7	35.4
Inter-rater (45 pairs)			
Minimum (variability)	0	0	0
Maximum (variability)	92.3	161.1	52.4
Misclassifications/rater (%)			
Progressive disease (RECIST >20%, WHO >25%)			
Intra-rater (10 pairs)	8.4 (12.2)	14.8 (21.4)	0.5 (0.7)
Inter-rater (45 pairs)	10.7 (15.5)	18.2 (26.4)	8.0 (11.5)
Response (RECIST >30%, WHO >50%)			
Intra-rater (10 pairs)	3.8 (5.5)	3.2 (4.6)	0.2 (0.3)
Inter-rater (45 pairs)	5.2 (7.5)	4.6 (6.7)	2.2 (3.2)

*Aligned with RECIST criterion. RECIST = Response Evaluation Criteria in Solid Tumors, WHO = World Health Organization

versus volume ($P < 0.001$). In the long run, we expect the difference between two volumetric measurements on a subject to differ by no more than −11.2%, 11.6% for repeatability study and −23.9%, 25.8% for reproducibility on 95% of occasions [Figure 1]. This means that increases and decreases less than the threshold can be a result of the inherent variability and may be indistinguishable from changes caused by variability alone and are unproven as a marker of efficacy in clinical trials.

The intra-rater and inter-rater reliability were 0.998 and 0.971 for unidimensional measurements, 0.998 and 0.982 for bidimensional measurements, and 1.000 and 0.997 for volumetric measurements. In addition, the volumetric technique had the smallest VC [Table 1].

Sources of variation
For the analysis of variance, the dependent variable was the tumor size measured and the independent variables were tumor, rater, and replication. The results indicated that tumor effect (measurement variability resulted from tumor characteristics alone) and rater effect (measurement variability resulted from rater characteristics alone) were significant in producing measurement variability, and the vast majority of variability was contributed by tumor effect [Table 3].

Influence of tumor characteristics
Compared with unidimensional and bidimensional techniques, volumetric technique had the lowest misclassification rate and VC and the highest agreement and reliability. Therefore, volumetric technique was optimal for therapeutic response assessment of lung cancer [Table 4].

For repeatability (intra-rater) study, tumor size ($P < 0.001$) and interface ($P = 0.001$) influenced the volumetric measurement: the lower variability was found in isolated tumors with interface of <45°, and the lowest variability could be obtained at tumor size of 57 mm by the fitted function of $Y = 0.001X^2 − 0.114X + 7.524$ [Figure 2]. For reproducibility (inter-rater) study, variability was only associated with tumor size ($P < 0.001$) and the lowest variability appeared at 40 mm by the fitted function of $Y = 0.004X^2 − 0.317X + 16.079$ [Figure 2].
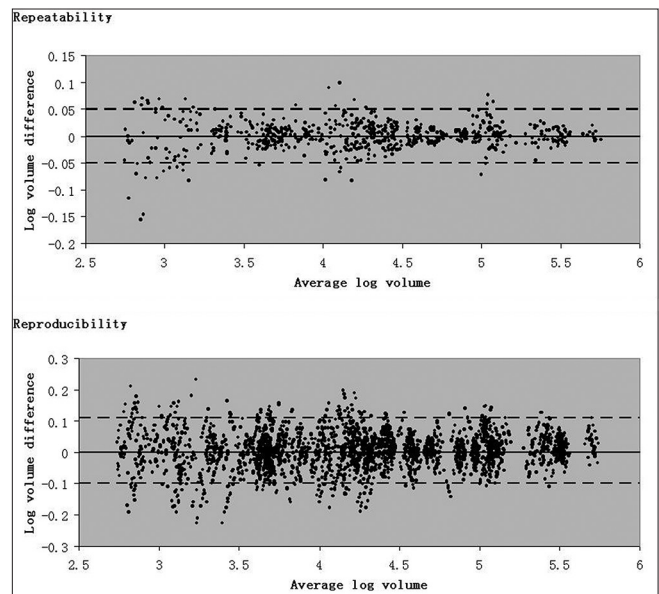

Figure 1: Bland–Altman plots demonstrating the agreement between intra-rater (repeatability) and inter-rater (reproducibility) measurements of volume, which is logarithmically transformed. As presented in the Bland–Altman plots, the level of agreement is significantly higher for intra-rater measurements than that for inter-rater measurements

Discussion

Compared to unidimensional and bidimensional techniques, our study showed that volumetric technique had the minimum variability in measuring lung cancer with CT scans, and the vast majority of variability was produced by tumor effect. Furthermore, variability was associated with tumor size by nonlinear mathematical equation. To clarify the significance of our results, we will elucidate the following key points:
1. Why should reliability be introduced into analysis of measurement variability in lung cancer?
2. Is conventional inter-observer variability really a result of observer (rater) heterogeneity or subjective effect?
3. Is there linear or nonlinear relationship between measurement variability and tumor size?

Table 3: Analysis of variance of tumor measurements

Source	Type III sum of squares	df	Mean square	P	Weight (%)
Unidimensional model (R^2=0.963)					
Corrected model	753,121.895	78	9655.409	0	
Intercept	2,555,130.006	1	2,555,130.006	0	
Tumor	750,267.319	68	11,033.343	0	97.3
Rater	2802.296	9	311.366	0	2.7
Replication	52.280	1	52.280	0.124	
Error	28,738.619	1301	22.090		
Total	3,336,990.520	1380			
Bidimensional model (R^2=0.967)					
Corrected model	5.974E9	78	7.659E7	0	
Intercept	5.117E9	1	5.117E9	0	
Tumor	5.960E9	68	8.765E7	0	98.3
Rater	1.335E7	9	1,483,846.652	0	1.7
Replication	48,358.231	1	48,358.231	0.578	
Error	2.037E8	1301	156,547.304		
Total	1.129E10	1380			
Volumetric model (R^2=0.995)					
Corrected model	1.298E13	78	1.665E11	0	
Intercept	4.991E12	1	4.991E12	0	
Tumor	1.298E13	68	1.909E11	0	99.8
Rater	3.880E9	9	4.311E8	0	0.2
Replication	2.660E7	1	2.660E7	0.486	
Error	7.127E10	1301	5.478E7		
Total	1.805E13	1380			

df=Degrees of freedom

Table 4: Influence of tumor characteristics on volumetric variability

	Presence (%)	Absence (%)	P
Intra-rater (n=690)			
Shape (regular)	3.8±3.5 (n=380)	4.3±4.5 (n=310)	0.084
Interface (<90°)	3.6±3.2 (n=350)	4.6±4.8 (n=340)	0.001
Size	$Y=0.001X^2-0.114X+7.524$ (R^2=0.316)		<0.001
Inter-rater (n=3105)			
Shape (regular)	9.4±8.4 (n=1710)	9.6±8.4 (n=1395)	0.378
Interface (<90°)	9.5±7.8 (n=1575)	9.6±9.0 (n=1530)	0.715
Size	$Y=0.004X^2-0.317X+16.079$ (R^2=0.191)		<0.001

Y=Variation of measurements (%), X=Size of tumor (mm)

Repeatability (intra-rater) refers to the variation in repeat measurements made on the same subject under identical conditions. This means that measurements are made by the same instrument or method, the same observer (or rater), and that the measurements are made over a short period, over which the underlying value can be considered to be constant. Reproducibility (inter-rater) refers to the variation in measurements made on a subject under changing conditions. The changing conditions may be due to different measurement methods or instruments being used, measurements being made by different observers or raters, or measurements being made over a period, within which the "error-free" level of the variable could undergo non-negligible change.[14]

Reliability and agreement

Repeatability and reproducibility are characterized by the concepts of agreement and reliability. Agreement quantifies how close two measurements made on the same subject are and is measured on the same scale as the measurements themselves. Reliability relates the magnitude of the measurement error in observed measurements to the inherent variability in the "error-free," "true," or underlying level of the quantity between patients.

In previous studies,[4-10,12,13] agreement has been emphasized and most of these studies used the Bland–Altman plots to demonstrate the agreement. Compared to agreement, however, reliability is rarely referred to. Reliability is critical for evaluation of therapeutic response because it represents the validity of measurement.[14,15] The agreement tells how close the first and the second measurements observed are, while reliability tells how close the measurements observed and the true size are. To a tumor with true size of 5.0 cm, intuitively, if the first measurement observed was 3.0 cm and the second measurement observed was 2.9 cm, agreement of measurements observed would be considered good because the difference of two measurements observed was so little (0.1 cm), but reliability would be poor because the measurements observed (3.0 cm or 2.9 cm) was so far from the true size of 5.0 cm.

We compared the agreement and reliability of unidimensional, bidimensional, and volumetric techniques, and the results
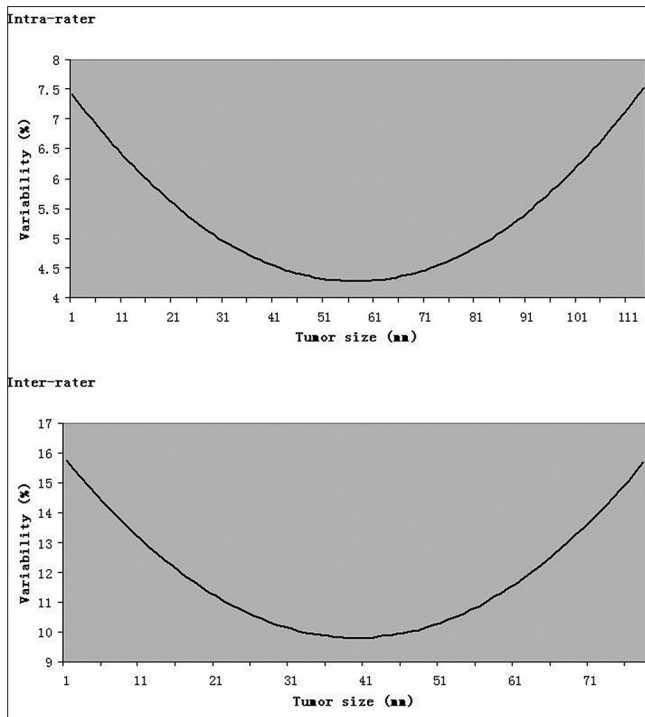
Figure 2: Fitted curves of variability (%) by tumor size (mm). For repeatability study, the lowest variability appeared at 5.7 cm of tumor size, but at 4.0 cm for reproducibility study

revealed that volumetric technique had the best agreement and reliability, indicating that volumetric measurements were optimal in consistency between raters (agreement) and between measurements observed and true measurements (reliability). In addition, given this increased interest in quantitative tumor measurements, it becomes important to understand what measurement changes are meaningful rather than a result of variability of measurement.[12] Our results showed that the 95% confidence interval (CI) of agreement of volumetric technique was from −11.2% to 11.6% for repeatability study and −23.9% to 25.8% for reproducibility study, indicating that a meaningful or true change can be determined as differences between measurements observed are beyond these 95% CIs, because measurement variability will be within these 95% CIs.

Is conventional inter-observer variability really a result of subjective effect?

Our current results indicated that both object effect (measurement variability resulted from tumor characteristics alone) and subjective effect (measurement variability resulted from rater characteristics alone) could influence inter-rater variability. However, the vast majority of variability was a result of object effect rather than subjective effect. What does that mean? It means that the inter-observer variability is primarily not a result of subjective effect. If the inter-observer variability is intrinsic to observers, it would be closely changed as observer changed, otherwise the association would be extrinsic. For example, there is a regular tumor and an irregular tumor; different observers have different measurements observed both in regular and irregular tumors. As we know, however, the differences of measurements observed would be smaller in regular tumor than that in irregular tumor to all observers.

Before the era of advanced volume technique, Erasmus *et al.*[7] concluded that measurements of lung tumor size on CT scans were often inconsistent and consistency can be improved if the same reader performs serial measurements for any one patient. With the development of computer-aided methods or automation techniques, measurement variability resulted from observers would be minimized or eliminated. Therefore, we think that the future efforts should be focused on the consistence of determining tumor borderline, which is more convenient and accurate in clinical practice.

Mathematical functions between variability and tumor size

Although the effect of pulmonary nodule characteristics on measurement has been reported in a number of studies, including nodule morphology, location, size, inspiration level, and segmentation,[8-11,13] there are limited data on object characterization in pulmonary masses. Our study showed that tumor segmentation, i.e., how to delineate the boundary of a tumor, was related with volumetric measurement variability, which is accordance with the previous study reporting that segmentation represents the most important factor contributing to measurement variability.[8] With the development of computer-aided methods or automation techniques, segmentation technique, i.e., how to delineate the boundary of a tumor, would become one of the most important points in tumor measurements.

It should be noted that nonlinear relationship was of significance between tumor size and volume variability in our study. Oxnard *et al.*[12] reported that larger tumors tend to have larger magnitude measurement changes in millimeters, but an opposite relationship occurred in relative change (percent increase or percent decrease). However, our results showed that nonlinear relationship had better goodness of fit than that of linear relation. The nonlinear relationship (point conic, or quadratic function, or U-shaped curve) reveals that medium-sized tumors tended to have the smallest variability. This is an interesting finding and the fact that medium-sized lesions are more reliably measured and very small and very large lesions are difficult to measure accurately.

Although volumetric quantification produced a promising result, accurate determination of response may require functional and molecular techniques.[16,17] In addition, we did not determine the threshold of evaluating therapeutic response for volumetric technique.

Conclusion

Volumetric technique has the minimum variability in measuring lung cancer with CT, and the vast majority of variability is a result of object effect (tumor characteristics). Moreover, medium-sized lesions are more reliably measured according to the established U-shaped curves between variability and tumor size.

References

1. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. Cancer 1981;47:207-14.

2. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, *et al.* New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 2000;92:205-16.

3. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer 2009;45:228-47.

4. Revel MP, Lefort C, Bissery A, Bienvenu M, Aycard L, Chatellier G, *et al.* Pulmonary nodules: Preliminary experience with three-dimensional evaluation. Radiology 2004;231:459-66.

5. Prasad SR, Jhaveri KS, Saini S, Hahn PF, Halpern EF, Sumner JE. CT tumor measurement for therapeutic response assessment: Comparison of unidimensional, bidimensional, and volumetric techniques initial observations. Radiology 2002;225:416-9.

6. Dinkel J, Khalilzadeh O, Hintze C, Fabel M, Puderbach M, Eichinger M, *et al.* Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer. Lung Cancer 2013;82:76-82.

7. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, *et al.* Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: Implications for assessment of tumor response. J Clin Oncol 2003;21:2574-82.

8. Gietema HA, Schaefer-Prokop CM, Mali WP, Groenewegen G, Prokop M. Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection CT – Influence of inspiration level, nodule size, and segmentation performance. Radiology 2007;245:888-94.

9. Gietema HA, Wang Y, Xu D, van Klaveren RJ, de Koning H, Scholten E, *et al.* Pulmonary nodules detected at lung cancer screening: Interobserver variability of semiautomated volume measurements. Radiology 2006;241:251-7.

10. Goodman LR, Gulsun M, Washington L, Nagy PG, Piacsek KL. Inherent variability of CT lung nodule measurements *in vivo* using semiautomated volumetric measurements. AJR Am J Roentgenol 2006;186:989-94.

11. Iwano S, Okada T, Koike W, Matsuo K, Toya R, Yamazaki M, *et al.* Semi-automatic volumetric measurement of lung cancer using multi-detector CT effects of nodule characteristics. Acad Radiol 2009;16:1179-86.

12. Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, *et al.* Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. J Clin Oncol 2011;29:3114-9.

13. Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ, de Bock GH, Gietema HA, Xu DM, *et al.* Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. Radiology 2008;248:625-31.

14. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008;31:466-75.

15. Farrell T, Cairns M, Leslie J. Reliability and validity of two methods of three-dimensional cervical volume measurement. Ultrasound Obstet Gynecol 2003;22:49-52.

16. Zhao B, Oxnard GR, Moskowitz CS, Kris MG, Pao W, Guo P, *et al.* A pilot study of volume measurement as a method of tumor response evaluation to aid biomarker development. Clin Cancer Res 2010;16:4647-53.

17. Nishino M, Dahlberg SE, Cardarella S, Jackman DM, Rabin MS, Ramaiya NH, *et al.* Volumetric tumor growth in advanced non-small cell lung cancer patients with EGFR mutations during EGFR-tyrosine kinase inhibitor therapy: Developing criteria to continue therapy beyond RECIST progression. Cancer 2013;119:3761-8.