Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer[1]

Binsheng Zhao, DSc
Leonard P. James, MD
Chaya S. Moskowitz, PhD
Pingzhen Guo, MD
Michelle S. Ginsberg, MD
Robert A. Lefkowitz, MD
Yilin Qin, MS
Gregory J. Riely, MD
Mark G. Kris, MD
Lawrence H. Schwartz, MD

Purpose: To evaluate the variability of tumor unidimensional, bidimensional, and volumetric measurements on same-day repeat computed tomographic (CT) scans in patients with non–small cell lung cancer.

Materials and Methods: This HIPAA– compliant study was approved by the institutional review board, with informed patient consent. Thirty-two patients with non–small cell lung cancer, each of whom underwent two CT scans of the chest within 15 minutes by using the same imaging protocol, were included in this study. Three radiologists independently measured the two greatest diameters of each lesion on both scans and, during another session, measured the same tumors on the first scan. In a separate analysis, computer software was applied to assist in the calculation of the two greatest diameters and the volume of each lesion on both scans. Concordance correlation coefficients (CCCs) and Bland-Altman plots were used to assess the agreements between the measurements of the two repeat scans (reproducibility) and between the two repeat readings of the same scan (repeatability).

Results: The reproducibility and repeatability of the three radiologists' measurements were high (all CCCs, ≥0.96). The reproducibility of the computer-aided measurements was even higher (all CCCs, 1.00). The 95% limits of agreements for the computer-aided unidimensional, bidimensional, and volumetric measurements on two repeat scans were (−7.3%, 6.2%), (−17.6%, 19.8%), and (−12.1%, 13.4%), respectively.

Conclusion: Chest CT scans are well reproducible. Changes in unidimensional lesion size of 8% or greater exceed the measurement variability of the computer method and can be considered significant when estimating the outcome of therapy in a patient.

[1] From the Departments of Radiology (B.Z., P.G., M.S.G., R.A.L., Y.Q., L.H.S.), Medical Physics (B.Z.), Medicine (L.P.J., C.S.M., G.J.R., M.G.K.), and Epidemiology and Biostatistics (C.S.M.), Memorial Sloan-Kettering Cancer Center, 1275 York Ave, New York, NY 10021. Received September 5, 2008; revision requested October 29; revision received January 13, 2009; accepted February 13; final version accepted March 3. Supported in part by the Experimental Therapeutics Center of Memorial Sloan-Kettering Cancer Center. **Address correspondence to** B.Z. (e-mail: *zhaob@mskcc.org*).

The two most widely accepted guidelines assessing objective response to therapy in patients with solid tumors are the World Health Organization criteria (1,2) and the Response Evaluation Criteria in Solid Tumors (3). The former determines response on the basis of an approximation of cross-sectional area (bidimensional measurement), whereas the latter uses only the tumor's greatest diameter (unidimensional measurement) measured on a transverse image, principally a computed tomographic (CT) scan. Both guidelines suggest reporting treatment results by using four categories: complete response, partial response, stable disease, and progression of disease.

Nearly 90% of patients with lung cancer have non–small cell lung cancer. Accurate and early assessment of response to a given therapy is critical for patient management and for further development of new therapies. Ideally, response to therapy should be determined with high accuracy and as quickly as possible to permit a prompt change in treatment, if necessary, and reduce the potential toxicity of an ineffective therapy.

A clinical study of non–small cell lung cancer that used multidetector CT and a three-dimensional computer segmentation software (4) showed that changes in tumor volume obtained from thin-section images could be determined as early as 3 weeks after chemotherapy was initiated, whereas changes detected by using the unidimensional and bidimensional techniques were less apparent during this same period. With the potential to measure size and/or change in size more accurately and assess response earlier and with different image postprocessing techniques, questions of CT scan reproducibility as well as measurement repeatability on CT scans need to be answered.

Variations in tumor measurements on serial CT scans can be introduced at the time of data acquisition (eg, nonuniform imaging technique/protocol, repeat CT scans) (5–8) and during the measurement procedure (eg, different measurement tools and human interpretation) (9–12). Despite the widespread use of CT scanning as a method of response assessment, little is known about the measurement reproducibility of in vivo tumors on serial CT scans. Previous studies on the variability of the repeat scans are limited because they looked at masses that were less than 2 cm or were of unknown type (primary tumor, pulmonary metastasis, benign pulmonary mass) (7,8). We designed and carried out a same-day repeat CT study to estimate the measurement variations in lung tumors seen in patients with non–small cell lung cancer. The purpose of our study was to evaluate the variability of tumor unidimensional, bidimensional, and volumetric measurements on same-day repeat CT scans in patients with non–small cell lung cancer.

Advances in Knowledge

- CT acquisition is generally accurate; the agreement between two scans obtained 15 minutes apart (reproducibility) is comparable with the agreement when the same scan is read twice (repeatability).
- By using thin-section multidetector CT and computer-aided segmentation software, reproducibility of unidimensional, bidimensional, and volumetric measurements in non–small cell lung cancer is high, with the volumetric being the most reproducible and the bidimensional the least reproducible measurement.

Materials and Methods

Patient Recruitment

This study was institutional review board approved and Health Insurance Portability and Accountability Act compliant. Patients were recruited through the oncologist's clinical practice. When patients would come in for their clinical visits, the oncologists would determine whether unenhanced chest CT was indicated in the near future and whether the patients were eligible for this trial. If both answers were "yes," the patients were offered participation in this study. If agreed, informed consent was obtained from each patient. Patient inclusion was determined by the following criteria: all patients must *(a)* be age 18 years or older, *(b)* have pathologically confirmed non–small cell lung cancer, *(c)* have measurable primary pulmonary tumors of 1 cm or larger, and *(d)* have scheduled a clinically indicated unenhanced CT scan of the chest. Exclusion criteria were *(a)* pregnant or lactating women and *(b)* those patients who were unable to consent to a repeat CT scan or for whom a repeat CT scan would be medically unsafe.

From January 2007 through September 2007, 32 consecutive patients (mean age, 62.1 years; range, 29–82 years) with non–small cell lung cancer were recruited. Of these patients, 16 were men (mean age, 61.8 years; range, 29–79 years) and 16 were women (mean age, 62.4 years; range, 45–82 years).

Implication for Patient Care

- The findings are valuable in revealing tumor variations on modern CT scanners by using advanced measurement tools, allowing more accurate evaluation of therapy response.

Repeat CT Scans

On completion of the clinically indicated CT scan, each patient was asked to

Published online
10.1148/radiol.2522081593

Radiology 2009; 252:263–272

Abbreviation:
CCC = concordance correlation coefficient

Author contributions:
Guarantors of integrity of entire study, B.Z., L.H.S., M.G.K., C.S.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, B.Z., L.P.J., G.J.R.; clinical studies, L.P.J., P.G., M.S.G., G.J.R., M.G.K., L.H.S., R.A.L., Y.Q.; statistical analysis C.S.M.; and manuscript editing, B.Z., L.P.J., C.S.M., G.J.R., M.G.K., L.H.S.

Funding:
This research was supported in part by the National Cancer Institute (grants R01 CA125143 and SP30 CA008748-40).

Authors stated no financial relationship to disclose.

leave the scanner table, walk around the CT scanner site, and return to the table for a second unenhanced scan. Both scans were obtained with the same CT scanner within 15 minutes of each other by using the same imaging protocol.

CT scans were obtained with a 16–detector row (LightSpeed 16; GE Healthcare, Milwaukee, Wis) or 64–detector row (VCT; GE Healthcare) scanner, both of which are routinely used at our center. Parameters for the 16–detector row scanner were as follows: tube voltage, 120 kVp; tube current, 299–441 mA; detector configuration, 16 detectors × 1.25-mm section gap; and pitch, 1.375:1. Parameters of the 64–detector row scanner were as follows: tube voltage, 120 kVp; tube current, 298–351 mA; detector configuration, 64 detectors × 0.63-mm section gap; and pitch, 0.984:1. The thoracic images were obtained without intravenous contrast material during a breath hold. Since the second scan was considered as a separate scan, its field of view was set given the patient's second scout image. Adjustment was allowed owing to the patient's position in the scanner. Thin-section (1.25 mm) images were reconstructed with no overlap by using the lung convolution kernel and transferred to our research picture archiving and communication system server where Digital Imaging and Communications in Medicine images are stored. These thin-section images were then used for both manual measurement and semiautomated computation of tumor sizes.

The standard clinical unenhanced chest CT exposes the patient to 6 mSv of radiation. An additional CT scan performed as part of this protocol would expose the patient to approximately 5 mSv of radiation. This exposure is equivalent to approximately 19 months of natural background radiation.

Manual Measurement of Tumor Size

The greatest diameter (ie, unidimensional measurement) and greatest perpendicular diameter of each tumor measured on a transverse image plane were manually measured by using an image viewing system developed with computer software (Interface Description Language; IDL Solutions, Germantown, Wis) in our laboratory. Bidimensional measurements are obtained by using the greatest perpendicular diameters of the tumor. Three radiologists (R.L., P.G., and M.G., with 10, 25, and 12 years experience with chest CT, respectively) independently measured 32 target lesions (one per patient). Among them, 29 were primary lung cancers and three were metastatic lesions (used because the primary tumors were nonmeasureable, as defined by the Response Evaluation Criteria in Solid Tumors criteria). The lung window settings (width, −50 HU; level, 1500 HU) were used for the manual measurements. In the same session, each radiologist first measured the 32 lesions on the images from the initial scans, then measured the 32 lesions on the images from the repeat scans, which were sequentially displayed in a different order than were scans from the first set. The radiologists were also blinded to the first measurement and the fact that these were repeat CT scans. In a separate session (2 days later), two of the three radiologists remeasured the same lesions on images from the initial scans. One radiologist performed all the measurements in one session.

Computer-aided Measurement of Tumor Size

In this study, our own semiautomated three-dimensional technique was used to separate the target lesions from surrounding anatomic structures (4,13,14). This algorithm was originally developed for small pulmonary nodules seen on high-attenuation CT images for noninvasive diagnosis (13,14) and later modified to assist in the segmentation of lung lesion masses for volumetric response assessment (4). Briefly, a region of interest that tightly encloses the tumor needs to be manually selected on one image by using the mouse. Starting with a higher density threshold level calculated on the basis of density values of the pixels inside the region of interest, the threshold level decreases in a stepwise manner. At each threshold level, the largest three-dimensional object (ie, geometrically connected voxels) can be determined and its surface gradient calculated. The threshold level that provides the maximum surface gradient is

Table 1

Descriptive Statistics of the Measurements Obtained by Three Readers

Measurement Type and Reader	Scan 1		Scan 2
	First Reading	Repeat Reading	
Unidimensional (mm)			
1	37.2 ± 18.4 (10.7–81.5)	36.9 ± 19.2 (9.2–85.5)	36.8 ± 19.2 (10.7–86.0)
2	34.0 ± 19.0 (9.3–84.7)	33.9 ± 18.6 (9.0–83.4)	33.9 ± 19.2 (8.1–83.2)
3	34.7 ± 16.9 (9.7–73.4)	35.6 ± 18.2 (8.9–83.9)	35.1 ± 18.9 (7.4–8937)
Bidimensional (mm^2)			
1	1211.3 ± 1213.2 (91.6–4906.4)	1216.1 ± 1308.7 (62.7–5561.3)	1223.7 ± 1278.6 (91.6–5255.1)
2	1035.0 ± 1219.6 (74.8–5135-9)	1002.6 ± 1100.7 (58.0–4315.5)	1024.6 ± 1193.0 (65.7–5053.1)
3	1094.5 ± 1138.8 (88.2–5048.5)	1139.0 ± 1209.2 (73.1–5710.5)	1092.4 ± 1164.0 (43.3–4851.2)

Note.—Data are the mean ± standard deviation; numbers in parentheses are the ranges.

considered as the optimal threshold level to separate the tumor from its surrounding structures. In this study, we adopted the strategies proposed by Zhao et al (4) to separate blood vessels and the chest wall from the tumor when needed.

To minimize possible variations in the segmentation results caused by the manual initiation of the computer algorithm, regions of interest of each tumor on the two repeat scans were selected side by side by an operator (B.Z.). Correctness and consistency of the computer results were visually inspected by participating radiologists (L.H.S. and P.G., with >15 and >20 years experience with chest CT, respectively). If any segmentation results were considered suboptimal, tumor contours that were superimposed on the original images were edited by a radiologist (P.G.) with our image viewing system. Once the segmentation and manual correction (if needed) were completed, the unidimensional, bidimensional, and volumetric measurements were calculated by using the computer algorithm. Tumor volume is defined as the sum of all tumor voxels, including voxels on and inside the tumor boundary, multiplied by the image resolutions in the x-, y- (in-plane), and z-directions.

Statistical Analysis

To estimate the reproducibility and the repeatability of the tumor size measurements by using repeat CT data, the concordance correlation coefficient (CCC) was initially used to quantify repeatability and reproducibility (15). If Y_1 and Y_2 are the measurements of a pulmonary lesion (i), we assumed that the (Y_{i1}, Y_{i2}) pairs were independent and followed a bivariate distribution, with means μ_1 and μ_2 and a covariance matrix $([\sigma_1^2, \sigma_{12}]^T, [\sigma_{12}, \sigma_2^2]^T)$. The formula CCC = $(2\sigma_{12})/(\sigma_1^2 + \sigma_2^2 + [\mu_1-\mu_2]^2)$ evaluates the degree to which pairs are located on the 45° line through the origin in a plot of the first and second measurements. CCCs can be broken down to a measure of precision (how far each pair of measurements deviates from the best-fit line through the data) and a measure of accuracy (the distance between the best-fit line and the 45° line through the origin). CCC values range from 1, perfect agreement between the repeated measurements, to -1, perfectly reversed agreement between measurements.

To further assess the reproducibility and the repeatability of the measurements, Bland-Altman plots were generated (16). For each of the three types of measurements, the percentage of relative difference between the repeated tumor measurements (defined as $100 \cdot [Y_{i2}-Y_{i1}]/Y_{i1}$) was plotted by using the average of the two lesion measurements. The limits of agreement were calculated by taking the mean of the percentage of relative differences between the two measurements and two standard deviations of these differences. Observations within the limits of agreement may be thought of as resulting from measurement error rather than a true change in tumor size. In contrast, observations outside of this range may reasonably be attributed to an actual change in tumor size. Note that the use of the percentage of relative difference in constructing these plots differs from the usual formula, which simply uses the difference of $Y_{i2}-Y_{i1}$. We chose to present the Bland-Altman plots and limits of agreement in this manner because the percentage of relative difference is the primary quantity of interest and more easily interpreted in this context than the difference.

Figure 1

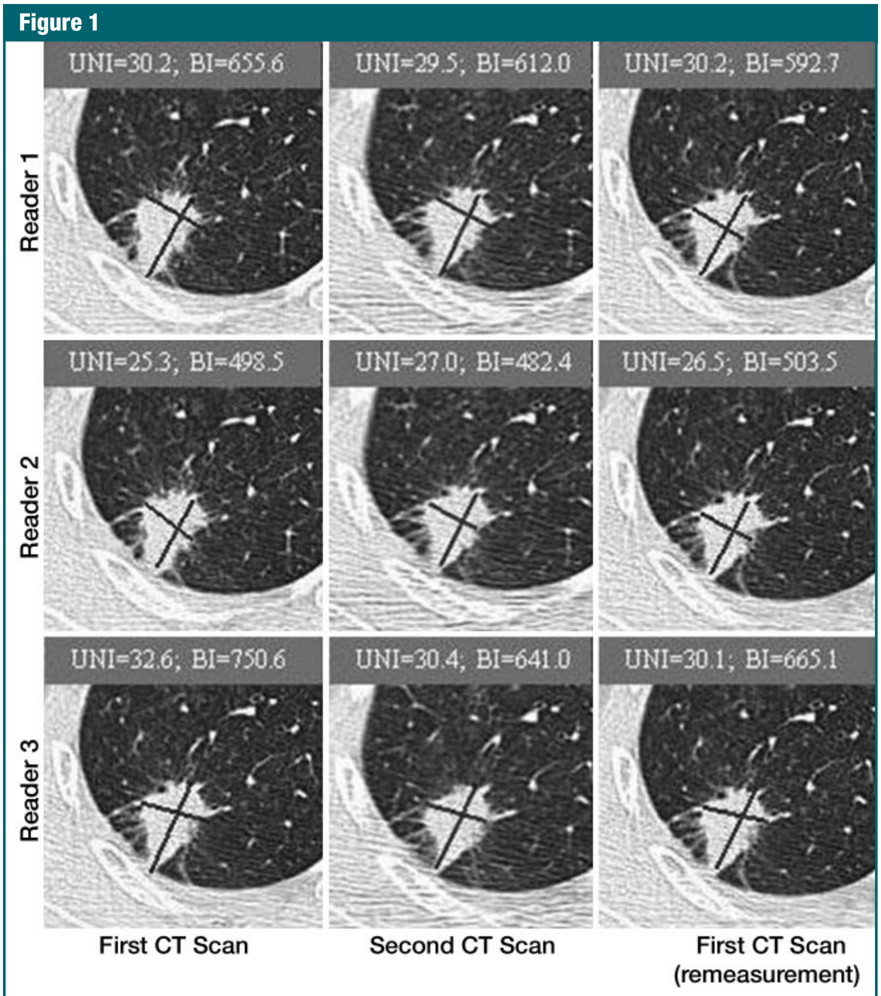

Figure 1: Three radiologists' manual unidimensional (*UNI*, in millimeters) and bidimensional (*BI*, in square millimeters) measurements of tumors on CT scans. Two greatest diameters (lines) were drawn by radiologists. Measurement values are at top of each scan.

Table 2

Radiologists' Measures of Agreement

Measurement Type, Comparison, and Reader	Concordance Correlation Coefficient*	Mean Relative Difference (%)	95% Limits of Agreement (%)
Unidimensional			
Scan 1 vs scan 2			
1	0.99 (0.98, 1.00)	−1.4	−18.3, 15.5
2	0.98 (0.97, 1.00)	−0.4	−22.1, 21.4
3	0.96 (0.94, 0.99)	0.1	−22.8, 23.0
Scan 1 repeat			
1	0.99 (0.99, 1.00)	−1.6	−11.8, 8.6
2	0.99 (0.98, 1.00)	0.1	−19.1, 19.3
3	0.98 (0.97, 0.99)	2.8	−23.1, 28.6
Bidimensional			
Scan 1 vs scan 2			
1	0.99 (0.99, 1.00)	−1.7	−25.6, 22.3
2	0.99 (0.99, 1.00)	−2.0	−33.1, 29.1
3	0.96 (0.93, 0.99)	0.1	−38.9, 39.1
Scan 1 repeat			
1	0.99 (0.99, 1.00)	−3.7	−22.1, 14.8
2	0.99 (0.98, 0.99)	−0.5	−30.2, 29.2
3	0.99 (0.98, 1.00)	6.2	−29.4, 41.8

* Data are the CCC; numbers in parentheses are the 95% confidence intervals.

To explore values for tumor measurements one might reasonably expect to see on repeated measurements of the same tumor, the conditional standard deviations for the second scan were determined on the basis of the first scan and assuming that the natural logarithms of the two scans were jointly and normally distributed by using a covariance matrix $([\tau_1^2, \tau_{12}]^T, [\tau_{12}, \tau_2^2]^T)$. For this analysis, the measurements were transformed to the logarithmic scale to have the distribution of data reasonably approximate the bivariate normal distribution. The standard deviation of the second measurement, which was conditional on the results of the first measurement, was then calculated by using the following equation:

$$\mathrm{cSD} = \sqrt{\tau_{2,2}^2 - \tau_{1,2}^2/\tau_{1,1}^2}.$$

A range of two conditional standard deviations was computed and then transformed back to the original data scale. We expect that approximately 95% of the measurements of the same tumor will be within two conditional standard deviations. Values outside of this range would likely represent a change in the size of the tumor rather than variation in measurement.

We also performed a variance-components analysis by using the mixed-effects model $\ln(y_{ijk}) = \theta + \alpha_j + \beta_k + \varepsilon_{ijk}$, where $\ln(y_{ijk})$ is the natural log of the lesion measurement $(i[n = 32])$, j is the number of radiologists $(n = 3)$, k is the number measurements obtained $(n = 2)$, θ is the mean lesion measurement (on the natural log scale), α_j is a random effect representing the radiologist, β_k is the random effect representing time, and ε_{ijk} is the random error term.

Finally, to explore interrater variability among the three radiologists, the inter-CCC, as proposed by Barnhart et al (17), was calculated across the three radiologists and between each pair of radiologists. The inter-CCC is defined in a manner similar to that of the original CCC described above but also incorporates the correlation between the radiologists. Further, it can also be interpreted by using the same guidelines provided above (range, 1 = perfect agreement, to −1 = perfectly reversed agreement).

Results

Radiologists' Measurements

Table 1 presents descriptive statistics on the 32 lesions. There did appear to be some variation across the different readings by each radiologist, though the differences were small (Fig 1). To further explore the agreement between the repeated scans and the repeated readings, Table 2 shows the CCCs (range, 0.96–0.99); the highest possible value is 1.00. There was equally good agreement for the unidimensional and bidimensional measurements. There was slightly greater agreement on the repeat reading of the first scan as compared with the agreement between the two separate scans.

Figure 2 shows the Bland-Altman plots with the mean percentage of relative difference and the limits of agreement listed in Table 2. With the possible exception of the scan 1 repeat reading performed by reader 3, the mean differences were all nearly zero. It is important to remember that differences within the limits of agreement can be attributed to measurement error, whereas values outside of the limits of agreement would suggest a true difference. For instance, in the scan 1 repeat reading performed by reader 1, we see that differences from −11.8% to 8.6% could potentially be attributed to a variation in the radiologist's measurement rather than a true change in tumor size. Hence, a smaller limit of agreement corresponds to a higher degree of agreement.

By using a hypothetical tumor measuring 2 cm in diameter on the first scan, Table 3 explores values of the three radiologists' reproducible (scan 1 vs scan 2) and repeated (scan 1 repeat reading) measurements on the same tumor that could reasonably be expected, given the first measurement. Approximately 95% of the repeat measurements are expected to be within the range of two conditional standard deviations. By using the results of reader 1 as an example, for a tumor that had a diameter of 2 cm on the first scan, a second scan of the same tumor will yield a measurement of 1.66–2.40 cm 95% of the time, and a repeat reading of the same tumor will yield a measurement of 1.81–2.21 cm 95% of the time. Because of the smaller limits of agreement, repeatability in both measurements showed a higher degree of agreement. The final column of Table 3 converts these ranges to the percentage of relative difference from the first scan. For instance, 95% of unidimensional tumor measurements obtained from a second scan of the same tumor that measured 2 cm on the first scan will be within −16.8% and 20.1% for reader 1.

Note that the percentages of relative differences in this table have a different interpretation than do the 95% limits of agreement in Table 2. The limits of agreement simply reflect where we expect 95% of all differences between repeated measurements to be, regardless of the actual tumor size, and do not assume that one tumor measurement is already known. In contrast, the numbers in Table 3 are generated on the basis of the first measurement, in essence assuming that the first measurement is known, and then tell us what might be expected for a second measurement of the same tumor, given that we already know the value of the first measurement.

Table 4 shows the results of a variance-components analysis, which shows the variance in lesion measurements that is attributable to the radiologists, to the repeated readings, and to random error. Further, although not a goal of this study, we also present the agreement among the radiologists in Table 5, where we see very high agreement for all measurements.

Figure 2

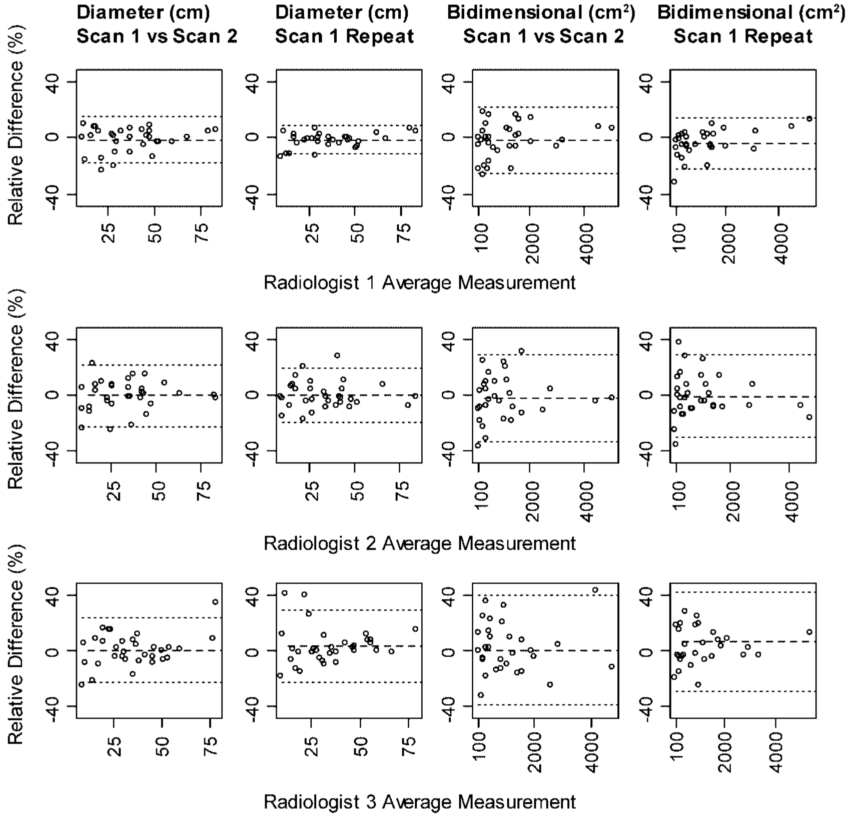

Figure 2: Bland-Altman plots of radiologists' measurements. Difference is plotted by using average of both tumor measurements for each patient. Dashed line (center) represents mean of differences. Top dotted line shows upper limit of agreement (mean difference plus 2 times standard deviation); bottom line shows lower limit of agreement (mean difference minus 2 times standard deviation). Plots show possible relationship between nodule size and relative difference in measurements (ie, the smaller the nodule, the larger the relative difference in measurements).

Semiautomated Measurements

A semiautomated method was applied to segmentation of the same 32 target lesions. Nineteen (59%) lesions on both scans were considered satisfactorily segmented by the radiologists; the remaining 13 (41%) lesions required manual correction. Figure 3 shows an example of a peripheral non–small cell lung cancer tumor that was successfully segmented by the computer on both scans.

The computer-generated measurements of the 32 lesions for both scans are shown in Table 6. Given these descriptive statistics, scans 1 and 2 appeared to yield relatively similar measurements. To better compare the measurements from the two scans, the data are displayed in Bland-Altman plots (Fig 4), with the mean relative differences and limits of agreement listed in Table 7. The mean difference across all pairs of repeated scans was nearly zero and the range of the limits of agreement was small, which suggested that the measurements were reproducible.

Table 7 also shows the CCCs for the computer-generated measurements, together with their 95% confidence intervals. All CCCs were 1.0, showing all measurements to be highly reproducible. Similarly, for the computer-generated measurements, by using a hypothetical tumor measuring 2 cm in diameter on the first scan, Table 8 shows values for a second repeated measurement on the same tumor that could reasonably be expected, given the first measurement.

Discussion

This study was designed to collect a clinical data set consisting of same-day repeat CT scans in patients with measurable non–small cell lung cancer and to evaluate the potential measurement variations in CT analysis owing to reproducibility of CT scans, as well as the repeatability of radiologists' manual measurements. Assuming that there was no biologic size change in the tumor on the two CT scans acquired within minutes of each other, size difference of the same tumor measured on the two repeat scans would allow us to explore the range of measurement variation in which a measured difference between two serial scans should be considered as a measurement error rather than a true change in size.

In the first part of this study, three radiologists independently measured 32 target pulmonary lesions on two repeat CT scans to explore reproducibility and then remeasured the lesions on the first scan for investigating the measurement repeatability. Among the radiologists' readings on two repeat scans, the best 95% limits of agreements were (−18.3%, 15.5%) and (−25.6%, 22.3%), and the worst agreements were (−22.8%, 23.0%) and (−38.9%, 39.1%), for the unidimensional and bidimensional measurements, respectively. These findings indicate that although radiologists' measurements introduced considerable variability for non–small cell lung cancer lesions, they were reproducible within the partial response category (−30% in unidimensional and −50% in bidimensional criteria) (2,3). However, the cutoff values for the progression of disease category seem to be defined as lower in both criteria (20% in unidimensional and 25% in bidimensional measurements), which coincides with the current concern about placing patients in the progression of disease category too easily (2,3). These findings need to be further validated with a larger number of radiologists.

Because variation caused by repeat readings is embedded in the measurement of repeat scans, in addition to exploring radiologists' measurement reproducibility, we also studied the repeatability of their measurements. The delay between the two repeat reading sessions was 2 days for the first and second radiologists. For the third radiologist, all measurements were performed in the same order but in one session. However, the third radiologist's results did not show better agreement on either reproducible or repeat measurements (Table 2). These findings indicate that the effect of radiologists' memory on the measurement of lesions may be limited. By remeasuring the same lesions on the first scan, radiologists' measurements revealed the range of the intrinsic variation (ie, measurement repeatability). Independently measuring tumors on each of the repeat scans and on the same scan twice revealed that agreement was only slightly greater for the latter method than for the former.

Table 3

Expected Radiologists' Repeat Measurements on a 2-cm Tumor Measured on the First Scan

Measurement, Comparison, and Reader	Tumor Size on Scan 1	Example	
		±2 cSDs	Difference*
Unidimensional (cm)	2.00		
Scan 1 vs scan 2			
1		1.66, 2.40	−16.8, 20.1
2		1.59, 2.52	−20.6, 26.0
3		1.59, 2.51	−20.4, 25.6
Scan 1 repeat			
1		1.81, 2.21	−9.5, 10.5
2		1.65, 2.42	−17.4, 21.0
3		1.57, 2.55	−21.6, 27.5
Bidmensional (cm^2)	3.14		
Scan 1 vs scan 2			
1		2.44, 4.04	−22.2, 28.5
2		2.27, 4.34	−27.6, 38.2
3		2.05, 4.81	−34.8, 53.3
Scan 1 repeat			
1		2.61, 3.78	−16.8, 20.3
2		2.32, 4.26	−26.3, 35.6
3		2.28, 4.33	−27.5, 37.9

Note.—cSD = conditional standard deviation.

* Plus or minus percentage difference.

Table 4

Variance-Components Analysis

Measurement Type and Comparison	Overall Mean (Fixed-effect Intercept)			Covariance Parameter Estimate for Random Effects		
	Estimate	Standard Error	*P* Value	Time	Radiologist	Residual
Unidimensional						
Scan 1 vs scan 2	3.418	0.070	<.001	0.309	0.005	0.006
Scan 1 repeat	3.420	0.070	<.001	0.309	0.007	0.003
Bidimensional						
Scan 1 vs scan 2	6.472	0.137	<.001	1.177	0.029	0.009
Scan 1 repeat	6.479	0.137	<.001	1.170	0.029	0.007

Table 5

Interrater Agreement Among the Radiologists

Interrater Agreement	Unidimensional		Bidimensional	
	Scan 1 vs Scan 2	Scan 1 Repeat	Scan 1 vs Scan 2	Scan 1 Repeat
Among all three Readers	0.98	0.98	0.98	0.98
Reader 1 vs 2	0.98	0.98	0.97	0.97
Reader 1 vs 3	0.99	0.99	0.99	0.99
Reader 2 vs 3	0.99	0.98	0.99	0.97

Note.—Data are the CCCs.

In the second part of this study, we used the same data set to optimally compare the reproducibility of the two repeat scans with a computer algorithm that calculated the unidimensional, bidimensional, and volumetric measurements. Our findings suggest that except for the first radiologist, the computer-generated unidimensional measurement had much narrower limits of agreement on the repeat scans, indicating a higher reproducibility of the computer over the radiologists for the unidimensional measurement. Interestingly, the computer demonstrated a very high reproducibility in the volumetric measurement: volume difference measured on the serial scans outside the range of −12.1% to 13.4% could be considered a true change in tumor volume. This range is much narrower than the cutoff values of (−65%, 40%) (or [−65%, 73%], converted from the diameter of a sphere) as suggested by Therasse et al (3) for detecting tumor response and progression by using volume in the Response Evaluation Criteria in Solid Tumors criteria.

Formally comparing unidimensional versus bidimensional or volumetric measurements was not a goal of our analysis, primarily because with 32 subjects, we were limited in our ability to conduct such an analysis. Further, it should be noted that when interpreting the results presented here, the relative percentage change in a measurement in a given dimension is not directly comparable with the relative percentage change in a measurement of a different dimension. For example, consider a hypothetical tumor with a radius (r) of 18 mm. By using calculations provided by James et al (18), the unidimensional measurement of this tumor is the diameter ($2r = 36$ mm). Assuming the tumor is spherical, the bidimensional measurement is $4r^2 = 1296$ mm^2. Now consider a second measurement of this tumor that results in a unidimensional measurement of 35.5 mm for a relative percentage change of −1.4%. The bidimensional measurement, given the unidimensional measurement of 35.5 mm, is $35.5^2 = 1260$ mm^2 for a relative percentage change of −2.7%, which is clearly different from the value of −1.4% observed with the unidimensional measurement. Therefore, one needs to be careful when comparing results across the different measurements. We could have presented all results converted to the same scale but chose not to, partially because it is not a goal of this study to compare across the

Figure 3

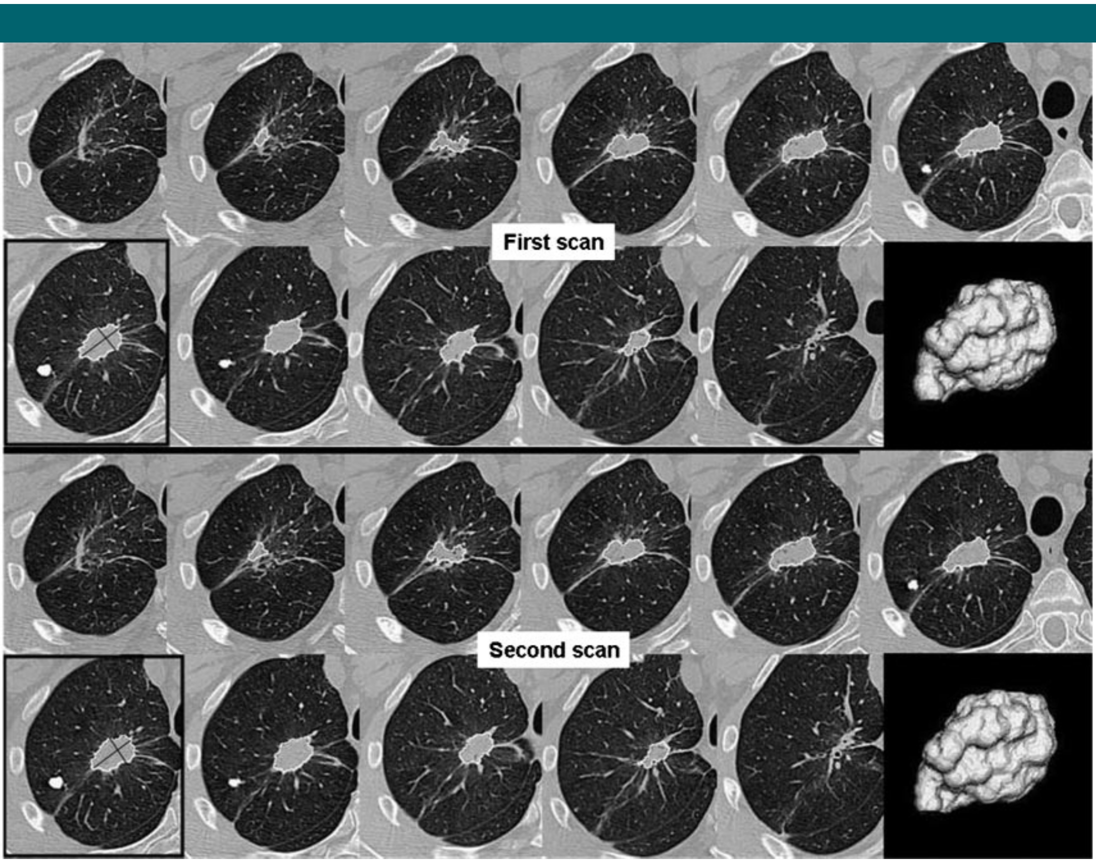

Figure 3: Computer-generated contours (white lines, superimposed on original images), two maximal perpendicular diameters (black lines, lower left image for first and second scans), and three-dimensional views (lower right image for first and second scans) of peripheral tumor on first (measurements: unidimensional = 29.7 mm, bidimensional = 507.9 mm^2, volumetric = 5564.4 mm^3) and repeat (measurements: unidimensional = 29.5 mm, bidimensional = 510.4 mm^2, volumetric = 5875.3 mm^3) scans. Every second sectional image was displayed.

different measurements and partially because these calculations assume that the tumors are spheric, which frequently may not be the case (4,19).

The relatively low success rate of the segmentation was, in part, a result of the advanced stage of non–small cell lung cancer. In this study, the tumor sizes ranged from 1.1 to 9.3 cm, with a mean of 3.8 cm. Such large masses likely attach to the surrounding structures of the soft-tissue density and can be extremely challenging to automated, accurate segmentation. As to geographic distribution of the tumors, 18 of 32 tumors were attached to surrounding structures, including the mediastinum (*n* = 3), hilum (*n* = 5), pleural effusion (*n* = 1), chest wall only (*n* = 4), and chest wall and mediastinum (hilum or diaphragm, *n* = 5). Some of the patients had obstructive pneumonia (ie, lung infection resulting from airway obstruction caused by tumor invasion), which radiographically complicated the tumor background.

The computer results should be viewed as the best possible estimates of the measurement variations or intrinsic difference between the repeat CT scans of non–small cell lung cancer. This is because we simulated an optimal scenario of computer-aided tumor measurements on serial CT scans. Assessment of a tumor at two points during the course of therapy will likely vary more than repeat scans on the same day since additional changes can occur during the time between follow-up studies, which would affect the measurements. It should also be noted that measurement variation could increase if different operators use the segmentation algorithm or if the same operator blindly initiates the same algorithm multiple times (7,8). Interoperator and intraoperator (interalgorithm) variations such as these are dependent on image context. For instance, there can be absolutely no difference in the measurements, no matter who is using the algorithm, if a tumor is surrounded by aerated lung parenchyma. On the other hand, measurement variation cannot be ignored if a tumor is attached to adjacent structures with similar attenuation as that of the tumor. Moreover, manual corrections on imper-

Table 6

Computer-generated Measurement Summary

Measurement	Data
Unidimensional (mm)	
Scan 1	38.3 ± 20.3 (10.6–93.3)
Scan 2	38.0 ± 20.0 (10.6–92.0)
Bidimensional (mm^3)	
Scan 1	1253.6 ± 1406.8 (89.0–5532.7)
Scan 2	1249.2 ± 1379.2 (86.9–5271.6)
Volumetric (mm^3)	
Scan 1	24 089.4 ± 37062.8 (223.1–155 405.2)
Scan 2	24 042.6 ± 36724.3 (229.5–152 197.4)

Note.—Data are the mean ± standard deviation; numbers in parentheses are the ranges.

Figure 4

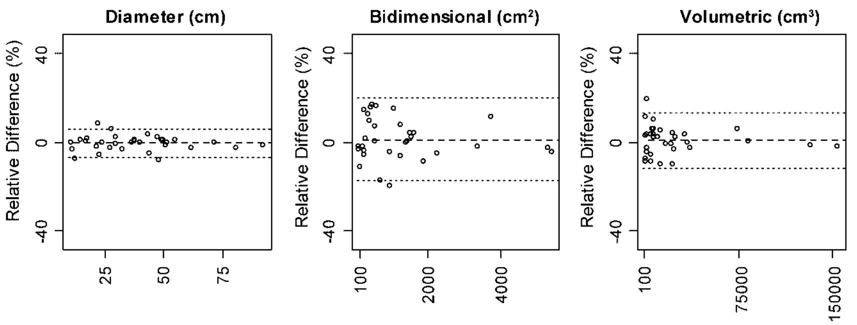

Figure 4: Bland-Altman plots of computer-generated measurements. Difference is plotted by using average of both tumor measurements for each patient. Dashed line represents mean of differences. Top dotted line shows upper limit of agreement (mean difference plus 2 times standard deviation); bottom line shows lower limit of agreement (mean difference minus 2 times standard deviation). Plots show possible relationship between nodule size and relative difference in measurements (ie, the smaller the nodule, the larger the relative difference in measurements).

Table 7

Computer-generated Measurements of Reproducibility

Measurement	Concordance Correlation*	Mean Relative Difference (%)	95% Limits of Agreement (%)
Unidimensional	1.00 (1.00, 1.00)	−0.6	−7.3, 6.2
Bidimensional	1.00 (0.99, 1.00)	1.1	−17.6, 19.8
Volumetric	1.00 (1.00, 1.00)	0.7	−12.1, 13.4

* Data are the CCC; numbers in parentheses are the 95% confidence intervals.

Table 8

Expected Computer Repeat Measurement for a 2-cm Tumor Measured on the First Scan

	Example		
Measurement	Tumor Size on Scan 1	±2 cSDs	Difference*
Unidimensional, (cm)	2.00	1.87, 2.14	−7.0, 6.5
Bidimensional (cm^2)	3.14	2.60, 3.80	−21.0, 17.5
Volumetric (cm^3)	4.19	3.69, 4.78	−14.1, 11.9

Note.—cSD = conditional standard deviation.

* Plus or minus percentage difference.

fect segmentations performed by different radiologists or by the same radiologist in different sessions will also bring variation to the measurements. Our side-by-side initiation of the algorithm and side-by-side manual correction on repeat scans helped keep the variations in the computer-aided measurements to a minimum, and optimized our ability to measure the change in actual tumor size independent of measurement effect.

The appropriate use of an imaging biomarker requires that it be both reproducible and repeatable. Despite its use as an imaging biomarker for decades, little research has been conducted regarding the reproducibility and repeatability of the CT scan and the radiologist's measurements. The lack of analysis results, in part, from the assumption that classifying a unidimensional (or bidimensional) measurement in four arbitrary categories (complete response, partial response, stable disease, or progression of disease) is generally within the acceptable range of reproducibility and repeatability. However, the cutoff values defined for partial response and progression of disease in the criteria were determined crudely and validated on the basis of then-standard diagnostic techniques (ie, physical palpation and standard radiographic analysis) in the late 1970s. Similar studies were never repeated for contemporary high-attenuation CT scanners that can acquire data volumetrically and for radiologists' techniques used to obtain measurements with electronic rulers on diagnostic monitors (the current technique standards in diagnostic radiology). Interestingly, in the development of alternate imaging biomarker techniques such as fluorine 18 fluorodeoxyglucose positron emission tomography and dynamic contrast enhancement–magnetic resonance imaging, reproducibility studies are being performed (20–23). For these reasons, we evaluated the reproducibility and repeatability of the CT scan in determining the size of pulmonary lesions in patients with non–small cell lung cancer.

The present study was performed on the basis of only 32 nodules and was designed to descriptively evaluate reproducibility and repeatability of CT measurements of solid tumors. While it is possible that with a smaller nodule, a larger relative difference in the measurements may be obtained (Figs 2, 4), it is difficult to know this from our data because of the limited number of large nodules. Furthermore, the small number of radiologists and lack of investigation of inter- and intracomputer measurement variations also limit the power of this study. Prior to being incorporated in clinical practice, our findings should be confirmed in larger, independent studies. Nevertheless, our findings will help reveal variations in the unidimensional, bidimensional, and volumetric measurements on modern CT scans and thus be valuable in helping detect biologically relevant changes in tumor size in the assessment of therapy response in non–small cell lung cancer. This provides clinicians greater precision and confidence to determine whether lung lesions actually grow or regress with therapies.

Acknowledgment: We thank our research assistant Carolina Montalvo, PhD, for her excellent effort in the coordination and collection of the data for this study.

References

1. World Health Organization. WHO handbook for reporting results of cancer treatment. Offset publication no. 48. Geneva, Switzerland: World Health Organization, 1979.
2. Miller AB, Hogestraeten B, Staquet M, Winkler A. Reporting results of cancer treatment. Cancer 1981;47:207–214.
3. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate response to treatment in solid tumors. J Natl Cancer Inst 2000;92:205–216.
4. Zhao B, Schwartz LH, Moskowitz C, Ginsberg MS, Rizvi NA, Kris MG. Computerized quantification of tumor response in lung cancer: initial results. Radiology 2006;241:892–898.
5. Winer-Muram HT, Jennings SG, Meyer CA, et al. Effects of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations. Radiology 2003;229:184–194.
6. Zhao B, Schwartz HL, Moskowitz C, et al. Effect of CT slice thickness on measurements of pulmonary metastases: initial experience. Radiology 2005;234:934–939.
7. Wormanns D, Kohl G, Klotz E, et al. Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. Eur Radiol 2004;14:86–92.
8. 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;186:989–994.
9. Thiesse P, Ollivier L, Di Stefano-Louineau D, et al. Response rate accuracy in oncology trials: reasons for interobserver variability. J Clin Oncol 1997;15:3507–3514.
10. Schwartz LH, Ginsberg MS, DeCorato D, et al. Evaluation of tumor measurements in oncology: use of film-based and electronic techniques. J Clin Oncol 2000;18:2179–2184.
11. Erasmus JJ, Gladish GW, Broemeling L, et al. Interobserver and intraobserver variability in measurement of non-small-call carcinoma lung lesions: implications for assessment of tumor response. J Clin Oncol 2003;21:2574–2582.
12. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol 1996;167:851–854.
13. Zhao B, Yankelevitz DF, Reeves AP, Henschke CI. Two-dimensional segmentation of pulmonary nodules on helical CT images. Med Phys 1999;26:889–895.
14. Zhao B, Reeves AP, Yankelevitz D, Henschke CI. Three-dimensional multicriterion automatic segmentation of pulmonary nodules of helical CT images. Opt Eng 1999;38:1340–1347.
15. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–268.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–310.
17. Barnhart HX, Song J, Haber MJ. Assessing intra, inter, and total agreement with replicated readings. Stat Med 2005;24:1371–1384.
18. James K, Eisenhauer E, Christian M, et al. Measuring response in solid tumors: unidimensional versus bidimensional measurement. J Natl Cancer Inst 1999;91:523–528.
19. Schwartz LH, Colville JA, Ginsberg MS, et al. Measuring tumor response and shape change on CT: esophageal cancer as a paradigm. Ann Oncol 2006;17:1018–1023.
20. Minn H, Zasadny KR, Quint LE, Wahl RL. Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET. Radiology 1995;196:167–173.
21. Weber WA, Ziegler SI, Thoedtmann R, Hanauske AR, Schwaiger M. Reproducibility of metabolic measurements in malignant tumors using FDG PET. J Nucl Med 1999;40:1771–1777.
22. Galbraith SM, Lodge MA, Taylor NJ, et al. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumors: comparison of quantitative and semi-quantitative analysis. NMR Biomed 2002;15:132–142.
23. Padhani AR, Hayes C, Landau S, Leach MO. Reproducibility of quantitative dynamic MRI of normal human tissues. NMR Biomed 2002;15:143–153.