Chapter 5

Pulmonary nodules detected at lung cancer screening: Interobserver variability of semiautomated volume measurements

Hester Gietema
Ying Wang
Dongming Xu
Rob van Klaveren
Harry de Koning
Ernst Scholten
Johny Verschakelen
Gerhard Kohl
Matthijs Oudkerk
Mathias Prokop

Radiology 2006;241:251-257

Abstract

Purpose

To retrospectively determine interobserver variability of semi automated volume measurements of pulmonary nodules and the potential reasons for variability.

Methods

The Dutch-Belgian lung cancer screening trial (NELSON) is a lung cancer screening study that includes men between the ages of 50 and 75 years who are current or former heavy smokers. The NELSON project was approved by the Dutch Ministry of Health and the ethics committee of each participating hospital. Informed consent was obtained from all participants. For this study, the authors evaluated 1200 consecutive low-dose computed tomographic (CT) scans of the chest obtained during the NELSON project and identified subjects who had at least one 50–500-mm^3 nodule. One local and one central observer independently evaluated the scans and measured the volume of any detected nodule by using semiautomated software. Non-calcified solid nodules with volumes of 15–500 mm^3 were included in this study if they were fully surrounded by air (intraparenchymal) and were detected by both observers. The mean volume and the difference between both measurements were calculated for all nodules. Intermeasurement agreement was assessed with the Spearman correlation coefficient. Potential reasons for discrepancies were assessed.

Results

There were 232 men (mean age, 60 years; age range, 52–73 years) with 430 eligible nodules (mean volume, 77.8 mm^3; range, 15.3–499.5 mm^3). Interobserver correlation was high (r = 0.99). No difference in volume was seen for 383 nodules (89.1%). Discrepant results were obtained for 47 nodules (10.9%); in 16 cases (3.7%), the discrepancy was larger than 10%. The most frequent cause of variability was incomplete segmentation due to an irregular shape or irregular margins.

Conclusion

In a minority (approximately 11%) of small solid intraparenchymal nodules, semiautomated measurements are not completely reproducible and, thus, may cause errors in the assessment of nodule growth. For small or irregularly shaped nodules, an observer should check the segmentation shown by the program.

INTRODUCTION

The introduction of multidetector computed tomography (CT) has led to a substantial increase in the number of incidentally detected small pulmonary nodules. In lung cancer screening trials in particular, only a small proportion of such nodules will be malignant; the vast majority will turn out to be benign [1-3]. The challenge is to correctly identify the few malignant nodules among the large number of benign ones by using noninvasive procedures, thus avoiding unnecessary invasive procedures [4].

Because of the high prevalence and often small size of these nodules, repeat CT scans are often performed to assess growth. The early assessment of growth has been established as the best way to discriminate between malignant and benign lesions [5-9]. Volume measurements obtained with automated segmentation tools have been shown to be more reproducible than diameter measurements [10]. Currently, various software tools are commercially available for that purpose. We have been using one of the commercially available software packages [11-15] for a multicenter lung cancer screening trial [16] and have found that volume measurements obtained with this semi automated nodule segmentation program sometimes differed between the first and second readings of the same CT data set by different observers. The purpose of this study, therefore, was to retrospectively determine interobserver variability of semi automated volume measurements of pulmonary nodules and potential reasons for the variability.

MATERIAL AND METHODS

STUDY GROUP

Data were collected from the NELSON project, a randomized Dutch-Belgian multicenter lung cancer screening trial that includes men aged 50–75 years who are current or former heavy smokers. Subjects were included if they smoked a minimum of 16 cigarettes per day for 25 years or 11 cigarettes per day for 30 years and were able to hold their breath for at least 20 seconds. Subjects who quit more than 10 years ago and/or had a history of cancer were excluded from participation.

The NELSON project is a population study that was approved by the Dutch Ministry of Health and the ethics committee of each participating hospital. Informed consent for participation and use of the collected data was obtained from all participants. Our study was performed by using CT data obtained at one of the participating centers (see below).

Within the NELSON project, all nodules with volumes of more than 15 mm^3 (corresponding mean diameter, 3.1 mm) were prospectively recorded in a database. Noncalcified solid nodules with volumes of more than 500 mm^3 (mean diameter, larger than 9.8 mm) were considered suspicious for carcinoma, and subjects with such nodules were referred to a pulmonologist for further work-up. Solid nodules with volumes of 15–50 mm^3 (mean diameter, 3.1–4.6 mm) were followed up at standard screening intervals (1 and 3 years). Noncalcified solid nodules with volumes of 50–500 mm^3 were considered indeterminate and followed up with a repeat CT examination after 3–4 months to determine growth. Nonsolid nodules, part-solid nodules, pleural-based nodules, and nodules attached to a vessel were excluded from this interobserver variability study because the software we used for volume measurements is not yet designed for calculating reliable volume estimates for these kinds of nodules.

For this study, we included the first 1200 consecutive participants scanned between April and December 2004 at one of the participating centers (University Medical Center, Utrecht, the Netherlands). We identified all individuals with at least one solid nodule with a volume of 50–500 mm^3. All patients were referred to short-term follow-up according to the trial set-up. We included all patients with nonc-alcified solid nodules that were not attached to vessels or pleura with volumes of 15–500 mm^3. This volume range included the smaller nodules of 15–50 mm^3 found in this patient group because we wanted to acquire data for a wider range of nodule sizes.

CT Scanning

CT was performed with a 16–detector row scanner (Mx8000 IDT; Philips Medical Systems, Best, the Netherlands). All scans were performed in a spiral mode with 16 x 0.75-mm collimation and 15-mm table feed per rotation (pitch, 1.3). We used a caudocranial scan direction and scanned the entire chest in approximately 10 seconds. Contrast material was not injected. Exposure settings were 30 mAs at 120 kVp for patients weighing 80 kg or less and 30 mAs at 140 kVp for those weighing more than 80 kg. This corresponds to CT dose index values of 2.2 and 3.5 mGy, respectively. We reconstructed 1.0-mm-thick transverse images at 0.7-mm increments by using the smallest field of view to include the outer rib margins at the widest dimension of the thorax. A soft kernel was used for reconstruction (B filter; Philips Medical Systems).

Nodule Detection and Volume Measurements

CT scans were evaluated exclusively by using digital workstations (Leonardo workstation; Siemens Medical Solutions, Erlangen, Germany) with U.S. Food and Drug Administration–approved, commercially available software for semi automated volume measurements (LungCare; Siemens Medical Solutions). Potential nodules were identified by a human observer (H.A.G., with 1 year of experience in radiology) on transverse thin-slab maximum intensity projections (slab thickness, 5 mm), which were reviewed with standardized window level and width settings of 1500 and –500 HU, respectively.

After the observer marks a candidate nodule with a mouse click, the program automatically defines a volume of interest around the candidate nodule, which can be further analyzed by using volume-rendered displays or multiplanar reformations. The candidate nodule can then be either approved or discarded.

The evaluation of a nodule with a second mouse click initiates an automated volume measurement program, which includes the following steps that are performed in the background [4]. First, a fixed-attenuation threshold of –400 HU is applied, and a three-dimensional connected "structure of interest" is extracted. This structure of interest consists of the nodule and, if present, connected structures such as vessels or parts of the chest wall. Subsequently, a small spherical three-dimensional template that originates from the click point is gradually expanded; its cross-correlation with the segmented nodule is computed for each step. The peak value of the cross-correlation curve is determined, and an empirical cutoff value close to the peak value is used to separate the nodule from potential adjacent structures.

In this manner, an optimum three-dimensional template is generated that represents the nodule in the most optimal way. This template includes as much of the nodule as possible without the inclusion of surrounding structures. Finally, the nodule is segmented by fusing the optimum three-dimensional template and the structure of interest; this is followed by spatial reasoning to remove adjacent structures [4]. The segmented nodule is then shown in yellow on the volume-rendered display of the volume of interest. If the proposed overlay does not match the nodule, the observer can interactively change which point on the cross-correlation curve is taken as the best proposal for a three-dimensional template (by using the "modify" option). In this way, the user can shrink or grow the overlaid volume to a user-defined extent. In addition, three-dimensional cutoff planes can be set, defining the space that must be excluded from the overlay volume.

The program calculates the volume; the diameters along the x-, y-, and z-axes; and the minimum and maximum diameters of the segmented nodule to two decimal places.

NODULE EVALUATION

After the first reading by one local observer (H.A.G.), scans were transferred for central reading to University Hospital Groningen, where another observer (Y.W. or D.X., both with 6 years of experience in radiology) performed the second reading. The same software algorithm was used at both centers. All nodule measurements were performed once locally and once at the central reading area without modification of the output of the program. Finally, the segmented volumes were stored in a central database (the NELSON Management System).

Nodules for which volume measurements yielded different results underwent further work-up. Two observers (H.A.G. and M.P.)—both from the local institution—classified the morphology of these nodules with regard to their outer contours (smooth, irregular, polylobulated, or spiculated) and forms (round, oval, or elliptic). This classification was performed in consensus and on the basis of visual assessment. For each of these nodules, the same two observers visually compared the two segmented volumes (marked in yellow) and the appearance of the nodule on the volume-rendered images to qualitatively assess the discrepancies in the segmented volumes. This routine procedure was performed for all nodules as part of the NELSON project.

DATA ANALYSIS

Differences in volume were calculated by subtracting the volume measured by the local observer from that measured by the central observer. Differences in volume measurements of one nodule were normalized with respect to the underlying nodule size. This was done separately for each nodule by calculating the ratio between the difference in volume measurements (numerator) and the mean volume (denominator). Agreement in measured volume was shown visually by using Bland-Altman plots with 95% confidence intervals (CIs) (limits of agreement) [17]. Because nodule size showed a non-normal distribution, inter-measurement agreement was determined by calculating the Spearman correlation coefficient.

We determined the number of nodules for which there were discrepant volume measurements. The degree of variation was calculated as the difference between the first and second measurements relative to the mean volume measured by both observers. We did this for the entire group of nodules and for the following three nodule-volume subgroups: 15–50 mm^3, greater than 50 to 200 mm^3, and greater than 200 to 500 mm^3. Differences between subgroups were calculated by using one-way analysis of variance.

In addition, we calculated descriptive statistical parameters for the relative differences between the two measurements for (a) all nodules and (b) only those nodules in which discrepancies in volume were found. The difference in the frequency of discrepancies in nodule size between the three nodule size categories was determined with the Spearman correlation coefficient for non–normally distributed populations. A P value of less than .05 was considered to indicate a statistically significant difference.

To calculate the degree of variation in those nodules for which discrepant volume measurements were obtained, we used the positive value of the relative difference between both measurements. All statistical analyses were performed with software (Excel 2000, Microsoft, Redmond, Wash; and SPSS, version 12, SPSS, Chicago, Ill.).

RESULTS

NODULES INCLUDED

From among patients who underwent the first 1200 baseline scans performed in our hospital as part of the NELSON project, we identified 232 men who had at least one nodule with a volume of 50–500 mm^3. In these 232 men (mean age, 60 years; age range, 52–73 years), we identified 450 nodules with volumes of 15–500 mm^3 that had been detected and measured by both observers. Of these 450 nodules, 430 fulfilled the inclusion criteria of being solid and fully surrounded by lung tissue. Of the 20 nodules that were excluded, one was excluded because it was not solid, six were excluded because they were attached to a vessel, and 13 were excluded because they were attached to the costal pleura. The mean volume (± standard deviation) of these 430 nodules was 77.8 mm3 ± 71.7 (range, 15.3–499.5 mm^3). One hundred eighty-eight nodules (43.7%) had a volume of 15–50 mm3, 213 (49.5%) had a volume of 50–200 mm^3, and 29 (6.7%) had a volume of 200–500 mm^3.

MEASUREMENTS

Spearman correlation coefficient analysis revealed that inter-measurement (interobserver) agreement was excellent (r = 0.99) (Figure 1). Identical volumes were measured by both observers in 383 of the 430 nodules (89.1%; Figure 2). Six nodules showed a relative difference of less than –10%, nine showed a relative difference of –0.1% to –10%, 22 showed a relative difference of 0.1%–10%, and 10 showed a relative difference of more than 10%. This means that the positive difference in volume measurements was less than 10% in 31 nodules (7.2%) and more than 10% in 16 (3.7%). In five nodules (1.2%), the positive difference was more than 25%. The mean positive difference between the measurements for all 430 nodules was 1.2% ± 4.3. If we consider only those 47 nodules for which we found interobserver variability, these values increase to 10.2% ± 8.4 (Figure 3). Results of analysis according to nodule size demonstrated that the percentage of nodules for which there were discrepant volume measurements increased with larger nodule size: Although only 7.4% (14 of 188 nodules) with volumes of 15–50 mm^3 were affected, the percentage increased to 13.1% (28 of 213 nodules) for nodules with volumes of 50–200 mm^3 and to 17.2% (five of 29 nodules) for nodules with volumes of 200–500 mm^3 ($P < 0.001$). In the latter group of five nodules, the mean positive difference in measured volume was 16.5% ± 9.9.

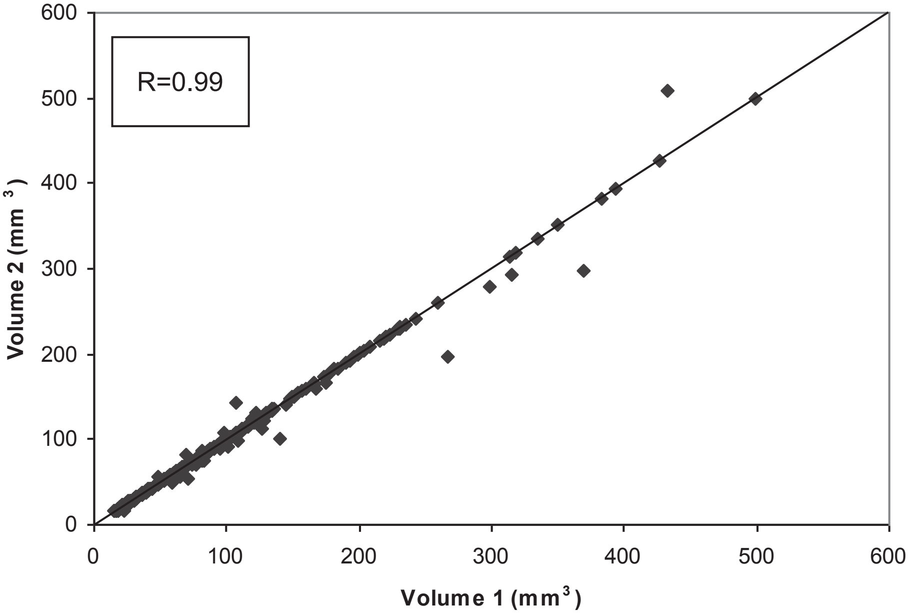
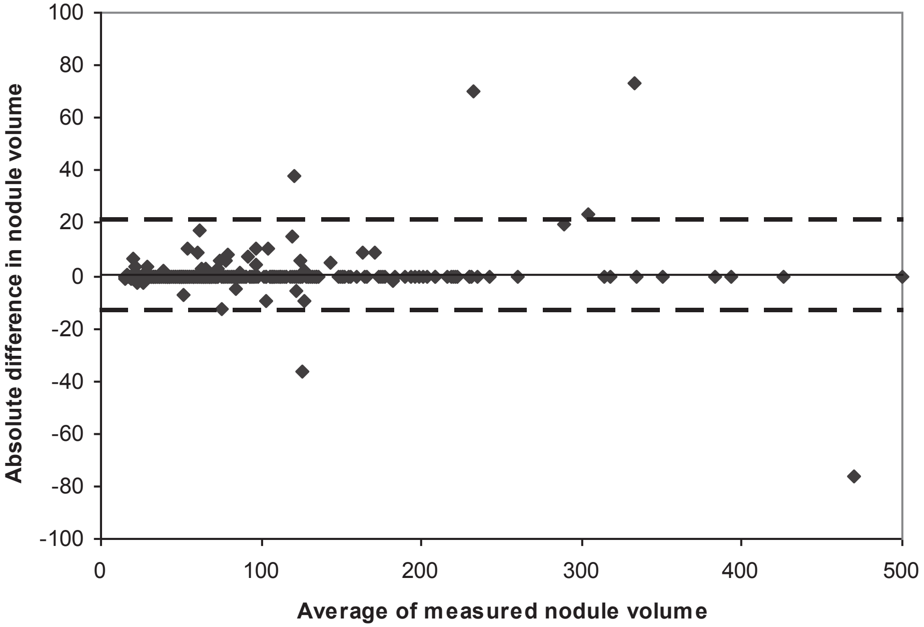
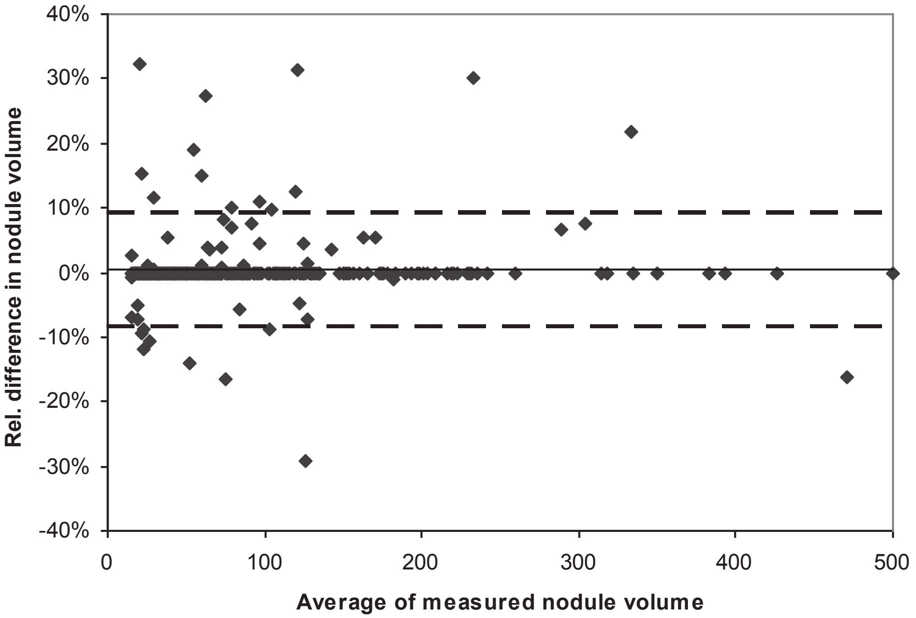


Figure 1

Volumetric results for each nodule. Volume 1 was measured by the local observer while volume 2 was measured by the central observer. The line indicates identical volumes measured by both readers.


100
80
60
40
20
0
-20
-40
-60
-80
-100
Absolute difference in nodule volume
0
100
200
300
400
500
Average of measured nodule volume

2A



2B

Figure 2

Bland and Altman plot for all 430 nodules included in the study. Absolute (A) and relative (B) interobserver variability is shown as a function of the mean nodule volume. The 95% confidence interval is indicated by a dashed lines; the bias (average difference) is stippled line. Note that although the mean difference is close to zero, a substantial interobserver variability can be seen as shown by the 95% confidence interval .

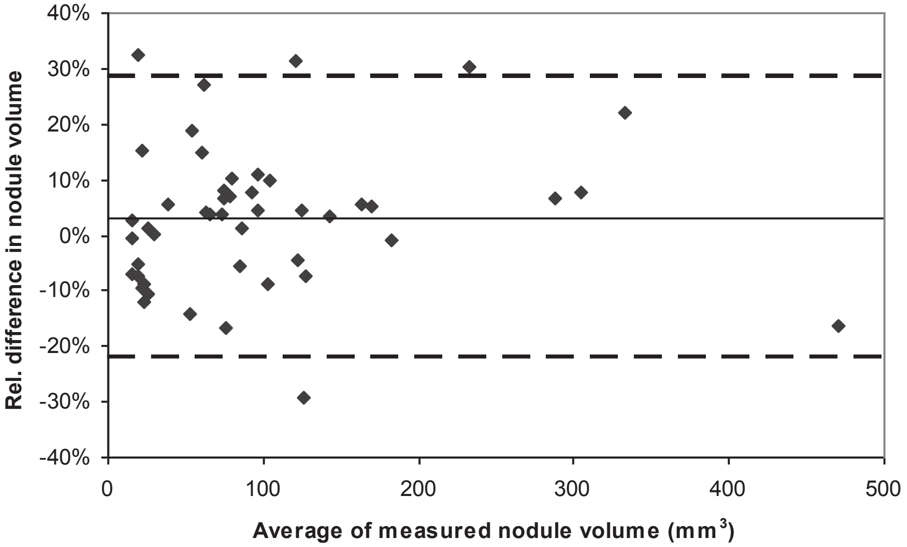


Figure 3

Bland and Altman plot for the 47 nodules in which there was a difference in volume measurements between central and local reading. Interobserver variability is shown as a function of the mean nodule volume. The 95% confidence interval is indicated by a dashed lines; the bias (average difference) is continuous line.

REASONS FOR DISCREPANCIES

We identified the following potential reasons for discrepancies in volume measurements between observers. In very small nodules with a volume of less than 30 mm^3, the entire nodule that is visible on the volume-rendered image is not actually included in the threshold used for analysis. Instead, the outer voxels, which contain part nodule and part surrounding lung, are excluded. This was the cause of discrepant measurements in 12 nodules. The user, however, can interactively modify the segmented volume to improve segmentation.

In larger nodules, an irregular shape may cause problems because the separation between a nodule and its surroundings might be dependent on how well the template matches the actual nodule. As a result, parts of the nodule may not be included in the initial segmentation. Among the remaining 35 nodules for which measurements varied, we found an irregular margin in 27 (Figure 4), a lobulated shape in three, spiculated margins in three, and an elongated shape in two (Figure 5).

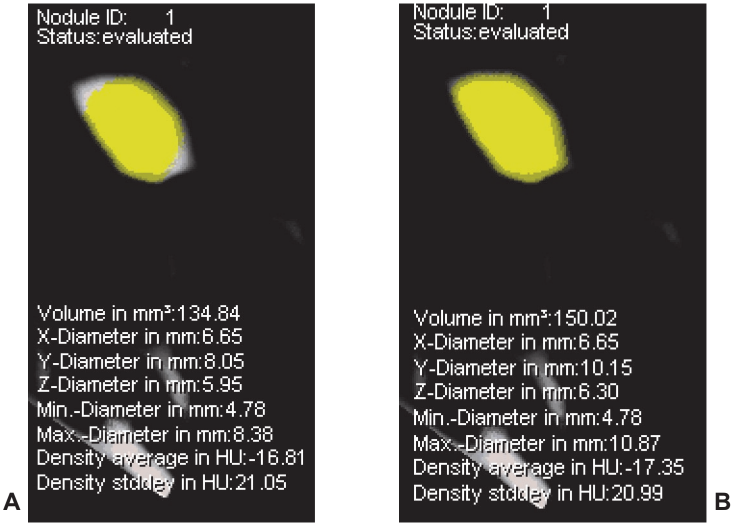


Figure 4

Nodule with a polygonal shape that was incompletely segmented during one of the two measurements. Nodule volume varied from 134.8 mm^3 (A) to 150.0 mm^3 (B).

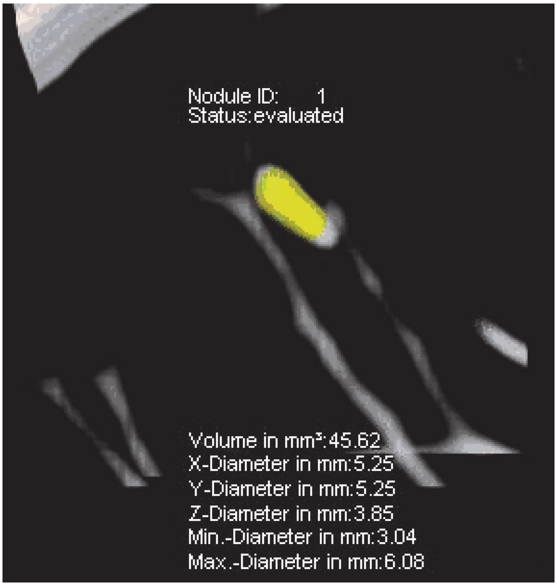


Figure 5

Elongated nodule in which there was incomplete segmentation (indicated in yellow), which can lead to interobserver variability.

DISCUSSION

Although semi automated measurements were highly reproducible for the vast majority of nodules, our results showed that there is a subset of nodules (approximately 11%) in which measured volume varies. The use of growth rates to detect malignant nodules, as is done in most lung cancer screening studies [1-3], necessitates reproducible volumetric measurements. It is known that interobserver variability is one of the factors that can influence the reproducibility of volume measurements [4;11;18].

The observation that there is a subset of nodules for which measurement varies between observers basically indicates that a very small measured difference in volume should not automatically imply real growth as the only source. When a repeat scan is performed after 6 weeks, the volume of a malignant nodule (with a volume doubling time of 300 days) that shows slow but substantial growth will increase about 10%. Referral of all subjects in whom nodules have shown a 10% increase in volume to a pulmonologist would imply that about 4% of the stable nodules will turn out to be false-positive findings. In lung cancer screening trials, in which thousands of nodules are being detected, this will concern a substantial number of nodules. A 3-month interval was recently recommended by the Fleischner Society for the follow-up of nodules with diameters of more than 6 mm (corresponding volume, 216 mm^3) [19]. After 3 months, a malignant nodule with a volume doubling time of 300 days will have increased 23% in volume. With use of a 25% increase in volume as the threshold for referral to a pulmonologist, the number of false-positive findings will decrease to 1.2%. When a repeat scan is performed after 6 months, no stable nodules will be depicted as showing substantial growth.

Currently, the combination of fully automated nodule detection and consecutive volume measurements is not yet commercially available. However, semiautomated software in which a nodule is digitally marked by an observer to initiate an automated volume measurement should already yield highly reproducible estimates because the only nonautomated part of the procedure is a manually indicated starting point for nodule segmentation. Results of in vitro validation studies of this software with artificial (rounded) nodules [14] demonstrated that the segmentation approach used with the software tools is very precise for rounded nodules, showing that even a change of 100 µm can be detected on CT images. Because in vitro reproducibility was 100%, there seemed to be little use in meticulously checking the segmentation of simple intraparenchymal nodules.

Semiautomated volume estimates of rounded nodules with a smooth margin appear to be extremely reproducible in vivo. In fact, we found no such nodule among those with discrepant measurements. This was expected because the software uses a spherical template to separate the nodule from surrounding structures. In our lung cancer screening trial, however, a substantial amount of nodules had a more irregular shape or margins, and some nodules had spiculation. These problems were seen most often in the group of intermediate-sized nodules with volumes of 50–500 mm^3. The segmentation of such nodules often turns out to be incomplete and variable, due in part to the partial volume effect on the margins of these nodules, which leads to a lowered attenuation in part of the nodule and exclusion of this part of the nodule from the segmentation. The irregular shape can lead to problems with the spherical template: Depending on the location of the seed point, there may be different peak values for matching the spherical template with the real nodule, and, thus, the measured volume may vary. This is important because this group of irregular nodules can be expected to include the most malignant ones. With the current semiautomated software, however, it is possible to modify the segmented volumes and, consequently, adapt the segmented part to the shape of the nodule. In this way, the observer can match the segmentation shown as a yellow overlay with the visual assessment of the nodule. In cases of ill-defined margins, it can be difficult to identify the correct border lines that separate the entire nodule from its surroundings.

Although we focused on software from one particular vendor for this study, most other volume measurement software for lung nodules also use seed points to indicate a nodule and require some type of paradigm to separate a nodule from neighboring structures. This makes it probable that other programs may also be affected by the factors described earlier.

A potential limitation of this study was the fact that the on-site reviewer had only 1 year of radiology experience. This reader, however, was well trained for the specific task of identifying and selecting nodules. Because this was not a nodule detection study but a study in which semiautomatic software was used to calculate the volume of nodules, the influence of user experience should be minimal. Another potential limitation of this study was the fact that it was based on one specific software program; thus, our results are strictly applicable to this specific software only. Any other software with which any manual user input is required will also be subject to potential interobserver variations. The amount of these variations, however, will be dependent on the software.

In conclusion, the results of our study demonstrate that although modern software used to measure lung nodule volume yields very high interobserver correlation ($r = 0.99$), volume measurements may vary in approximately 11% of cases. Because the detection of minor differences in nodule volume is important for determining whether a nodule is growing, it is essential to improve reproducibility even further. These tools, however, cannot be used in a fully automated approach. In fact, it is important to check the segmentation for completeness, particularly when a nodule has an irregular shape or irregular margins. Early published in vitro data [14] gave hope that checking the segmentation does not appear necessary; our results, however, suggest that a meticulous check of the segmentation should be performed in a clinical (in vivo) setting: In case of an incomplete segmentation, one should try to reposition the initial seed point to ensure visually complete segmentation and, when necessary, modify the segmented volume to match segmentation to the visual assessment of the nodule of interest.

REFERENCE LIST

(1) Diederich S, Wormanns D, Heindel W. Lung cancer screening with low-dose CT. Eur J Radiol 2003; 45(1):2-7.

(2) Henschke CI, McCauley DI, Yankelevitz DF et al. Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 1999; 354(9173):99-105.

(3) Swensen SJ, Jett JR, Hartman TE et al. Lung cancer screening with CT: Mayo Clinic experience. Radiology 2003; 226(3):756-761.

(4) 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(1):86-92.

(5) Diederich S, Thomas M, Semik M et al. Screening for early lung cancer with low-dose spiral computed tomography: results of annual follow-up examinations in asymptomatic smokers. Eur Radiol 2004; .

(6) Henschke CI, Naidich DP, Yankelevitz DF et al. Early lung cancer action project: initial findings on repeat screenings. Cancer 2001; 92(1):153-159.

(7) Takashima S, Sone S, Li F et al. Indeterminate solitary pulmonary nodules revealed at population-based CT screening of the lung: using first follow-up diagnostic CT to differentiate benign and malignant lesions. AJR Am J Roentgenol 2003; 180(5):1255-1263.

(8) Usuda K, Saito Y, Sagawa M et al. Tumor doubling time and prognostic assessment of patients with primary lung cancer. Cancer 1994; 74(8):2239-2244.

(9) Wormanns D, Diederich S. Characterization of small pulmonary nodules by CT. Eur Radiol 2004; 14(8):1380-1391.

(10) Revel MP, Lefort C, Bissery A et al. Pulmonary nodules: preliminary experience with three-dimensional evaluation. Radiology 2004; 231(2):459-466.

(11) Goo JM, Lee JW, Lee HJ et al. Automated lung nodule detection at low-dose CT: preliminary experience. Korean J Radiol 2003; 4(4):211-216.

(12) Hasegawa M, Sone S, Takashima S et al. Growth rate of small lung cancers detected on mass CT screening. Br J Radiol 2000; 73(876):1252-1259.

(13) Reeves AP, Kostis WJ. Computer-aided diagnosis for lung cancer. Radiol Clin North Am 2000; 38(3):497-509.

(14) Wormanns D, Fiebich M, Saidi M et al. Automatic detection of pulmonary nodules at spiral CT: clinical application of a computer-aided diagnosis system. Eur Radiol 2002; 12(5):1052-1057.

(15) Zhao B, Gamsu G, Ginsberg MS et al. Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm. J Appl Clin Med Phys 2003; 4(3):248-260.

(16) van Klaveren RJ, Habbema JD, de Koning HJ et al. [Screening for lung cancer in the Netherlands: the role of spiral CT scan]. Ned Tijdschr Geneeskd 2001; 145(11):521-526.

(17) Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1(8476):307-310.

(18) Kostis WJ, Yankelevitz DF, Reeves AP et al. Small pulmonary nodules: reproducibility of three-dimensional volumetric measurement and estimation of time to follow-up CT. Radiology 2004; 231(2):446-452.

(19) MacMahon H, Austin JHM, Gamsu G et al. Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society. Radiology 2005; 237(2):395-400.