Marie-Pierre Revel, MD
Alvine Bissery, MSc
Marie Bienvenu, MD
Laetitia Aycard, MD
Catherine Lefort, MD
Guy Frija, MD

Index terms:
Lung neoplasms, CT, 60.12115
Lung neoplasms, diagnosis, 60.31, 60.32, 60.33
Lung, nodule, 60.281

Published online
10.1148/radiol.2312030167
Radiology 2004; 231:453–458

Abbreviations:
PACS = picture archiving and communication system
2D = two-dimensional

[1] From the Department of Radiology (M.P.R., M.B., L.A., C.L., G.F.) and Clinical Investigation Center 9201 (A.B.), Assistance Publique des Hôpitaux de Paris/INSERM, Georges Pompidou European University Hospital, 20 Rue Leblanc, 75015 Paris, France. Received January 31, 2003; revision requested April 21; final revision received October 2; accepted November 17. **Address correspondence to** M.P.R. (e-mail: *marie-pierre.revel@hop.egp.ap-hop-paris.fr*).

See also the other article by Revel et al in this issue.

Author contributions:
Guarantors of integrity of entire study, M.P.R., G.F.; study concepts, M.P.R., G.F.; study design, M.P.R., A.B.; literature research, M.P.R., C.L.; clinical studies, M.B., L.A.; data acquisition, M.P.R., M.B., L.A.; data analysis/interpretation, A.B.; statistical analysis, A.B.; manuscript preparation, M.P.R.; manuscript definition of intellectual content and editing, M.P.R., C.L.; manuscript revision/review and final version approval, M.P.R., C.L., G.F.

Are Two-dimensional CT Measurements of Small Noncalcified Pulmonary Nodules Reliable?[1]

PURPOSE: To evaluate the intra- and interreader agreement of two-dimensional computed tomographic (CT) measurements of pulmonary nodules less than 2 cm in diameter.

MATERIALS AND METHODS: Three readers independently made three serial measurements of each of 54 pulmonary nodules measuring 3–18 mm that had been observed on standard-dose multisection CT images obtained in 24 patients who ranged in age from 36 to 81 years (mean age, 54.6 years). There were 14 women (58%), who ranged in age from 43 to 81 years (mean age, 58.9 years), and 10 men (42%), who ranged in age from 36 to 65 years (mean age, 48.5 years). The largest transverse cross-sectional diameter of each nodule was measured at picture archiving and communication system, or PACS, workstations by using high-spatial-resolution reconstructed CT images and identical window settings. Intra- and interreader agreement were determined by using methods described by Bland and Altman: the coefficient of repeatability for intrareader agreement, and methods derived from the 95% limits of agreement defined by Bland and Altman for interreader agreement.

RESULTS: The repeatability coefficients were 1.70, 1.32, and 1.51 mm for readers 1, 2, and 3, respectively. The 95% limits of agreement for the difference among readers were −1.73 and 1.73.

CONCLUSION: Two-dimensional CT measurements are not reliable in the evaluation of small noncalcified pulmonary nodules.

Incidental discovery of pulmonary nodules is very frequent during computed tomography (CT) of the chest. In lung cancer screening studies (1,2), pulmonary nodules have been identified in 23%–66% of subjects. Most nodules identified incidentally or during screening are morphologically indeterminate. Some may correspond to stage I lung carcinoma and need to be further investigated because diagnosis and treatment at this stage is associated with a 5-year survival rate of 60%–70%, compared with a global survival rate of only 15% among patients with lung carcinoma (3).

Various approaches can be used to characterize noncalcified pulmonary nodules, including positron emission tomography (PET), contrast material–enhanced CT, and CT-guided percutaneous biopsy.

PET is not applicable for the evaluation of all such nodules because its spatial resolution is limited and nodules less than 7–8 mm in diameter cannot be accurately assessed (4). In theory, contrast-enhanced CT can be used to evaluate nodules larger than 5 mm, but in practice, the lower size limit is about 10 mm; moreover, the specificity of this technique is only about 60% (5).

Because it is an invasive procedure, CT-guided percutaneous biopsy cannot be used as a first-line strategy, even if complications are relatively infrequent with it. In addition, its diagnostic accuracy is lower for smaller pulmonary nodules than for larger ones (6). For these reasons, small nodules are generally monitored by means of serial CT examinations, with the aim of detecting a size increase suggestive of malignancy. The Early Lung Cancer

Action Project group recommends that follow-up CT be performed 3 months after initial identification of nodules between 5 and 10 mm in diameter; if no growth is detected, CT should be repeated 6, 12, and 24 months later (1). Biopsy is indicated if growth is detected.

The purpose of our study was to evaluate the intra- and interreader agreement of two-dimensional (2D) CT measurements of pulmonary nodules less than 2 cm in diameter.

MATERIALS AND METHODS

According to our institutional guidelines, our institutional review board does not require its approval for our type of study; informed consent is also not required.

Nodule Selection and Imaging

Patients included in this evaluation were nonconsecutive patients who were identified, with a keyword search in our picture archiving and communication system (PACS) for 2001 and 2002, as having solid pulmonary nodules less than 2 cm in diameter. We included only those patients for whom CT images were obtained through each nodule with 2.50-mm or thinner collimation. Patients with ground-glass nodules or part-solid nodules were not included in this evaluation.

When we decided to conduct this study, we were able to identify 24 patients (with 54 nodules) who met these criteria and ranged in age from 36 to 81 years (mean age, 54.6 years). There were 14 women (58%), who ranged in age from 43 to 81 years (mean age, 58.9 years), and 10 men (42%), who ranged in age from 36 to 65 years (mean age, 48.5 years).

The number of nodules per patient ranged from one to six. Thirteen (54%) of the 24 patients had one nodule, five (21%) had two nodules, two (8%) had four nodules, one (4%) had five nodules, and three (12%) had six nodules (percentages may not add up to 100% due to rounding).

The situations in which the 54 nodules were detected were as follows: Twenty-three nodules had been found in 10 patients at CT performed to confirm conventional radiographic identification of pulmonary nodules. Five of these 10 patients were heavy smokers who had chronic obstructive pulmonary disease, while the other five patients were nonsmokers. Twenty nodules had been found in six patients during follow-up for extrathoracic cancer, six nodules had been found incidentally in four patients who had been referred for evaluation of suspected pulmonary embolism, three nodules had been found in two patients being evaluated for chronic obstructive pulmonary disease, one had been found in a patient with sarcoidosis, and the last had been identified during CT follow-up after catheter ablation of foci of ectopic paroxysmal atrial fibrillation.

Standard-dose CT images had been acquired with multi–detector row spiral CT scanners (LightSpeed, GE Medical Systems, Milwaukee, Wis; or Volume Zoom, Siemens, Erlangen, Germany) with four detector rows. The parameters used depended on the indication for CT: Collimation was 1.25 or 2.50 mm (4 × 1.25 mm or 4 × 2.50 mm), pitch was 1.2–1.5, rotation time was 0.5–0.8 second, and exposure parameters were 80–120 mAs (depending on the patient's weight) and 120–140 kV.

The acquisition field of view ranged from 290 to 390 mm, depending on the patient's size and shape. The acquisition matrix was 512 × 512, and the pixel size thus ranged from 0.56 to 0.76 mm. The mean size of the nodules was 8.5 mm ± 3.6 (SD). Twelve (22%) of the 54 nodules were less than 5 mm in diameter, 28 (52%) were 5 or more but less than 10 mm in diameter, 12 (22%) were 10 or more but less than 15 mm in diameter, and two (4%) were between 15 and 18 mm in diameter. There were three irregular oval nodules (6%), four spiculated nodules (7%), and one lobulated nodule (2%). The 46 remaining nodules (85%) were regular in shape and round ($n = 28$) or oval ($n = 18$).

Nodule Evaluation

The largest transverse cross-sectional diameter of each nodule was measured independently by three radiologists (M.P.R., M.B., L.A.), each of whom made three consecutive measurements of each nodule during the same session, with an interval of several minutes between each measurement. For instance, when a patient had several nodules, the readers were asked to measure all the nodules at each of the three readings. Analyses of patients with single nodules were pooled into groups of three or four patients, and the readers were asked to measure all the nodules at each reading as if the group represented a single patient with several nodules. This was meant to introduce a delay between the sequential analyses of a single nodule. The values of the three measurements were written down on three different score sheets.

Readers 1, 2, and 3, respectively, had 7, 2, and 4 years of experience in chest CT.

Measurements were made at PACS workstations (Impax 4.1; Agfa HealthCare, Mortsel, Belgium) with black-and-white 1,280 × 1,024-pixel screens (Siemens, Karlsruhe, Germany). Identical window settings were used by all three readers, and measurements were made on high-spatial-resolution-algorithm–reconstructed CT images by manually positioning electronic calipers. The radiologists were advised to zoom in on the nodules for more accurate analysis. Spiculations were included in the determination of the largest transverse cross-sectional diameter of the four spiculated nodules.

If a nodule was visible on several adjacent images, the image showing the largest transverse cross-sectional diameter was selected.

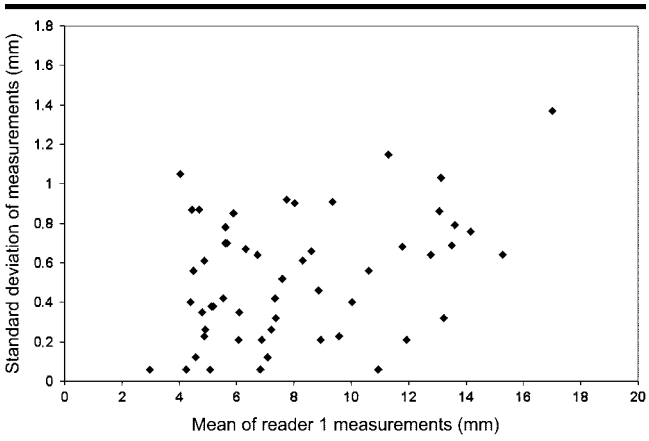
In patients with multiple nodules, the nodules were numbered craniocaudally, and in patients in whom more than one nodule was present at the same craniocaudal level, the nodules were numbered "outside to inside." Numbering was performed separately for each lung, starting with the right lung.

All the measurements (three values for each nodule and for each radiologist) were reported in separate tables for statistical analysis; nine measurements were thus obtained for each nodule.

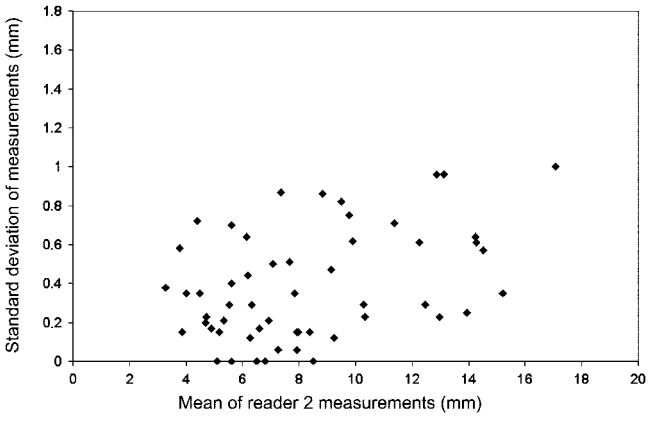
Statistical Analysis

We focused on the variability of 2D measurements of each nodule for each reader (intrareader agreement) and among the three readers (interreader agreement). We evaluated intrareader agreement for all 52 nodules, including irregular nodules, and then reevaluated the repeatability coefficient after excluding irregular nodules.

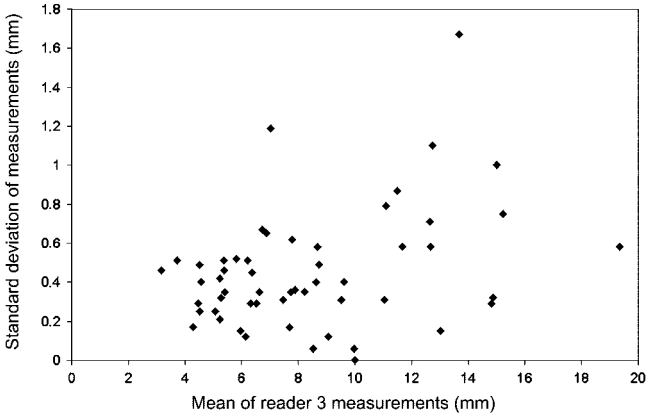
Assessment of intrareader agreement.—We used an extension of the repeatability coefficient, as defined by the British Standards Institution, for more than two repeated measurements of a given nodule (7). The SD of repeated measurements of a given nodule is used to assess the measurement error. The SD of repeated measurements is known as within-subject SD, or *sw*. The repeatability coefficient is then defined as 2.77 × *sw*, given a 5% error rate, when the assumption that the SD is unrelated to the size of the nodule is true (8). From a clinical point of view, this means that a difference of less than 2.77 × *sw* between two successive mea-



a.



b.



c.

Figure 1. Scatterplots show SDs versus mean nodule sizes for **(a)** reader 1, **(b)** reader 2, and **(c)** reader 3. Evaluating the magnitude of the SD of repeated measurements is a way of evaluating the measurement error for each reader. All three scatterplots show that the SD is not dependent on nodule size; in particular, it does not increase with nodule size. This lack of any clear graphic relationship authorized us to use the repeatability coefficient, as defined by the British Standards Institution (7), in this study.

surements of the same nodule cannot be distinguished from measurement error and thus cannot be considered to represent an actual increase in size (9).

Assessment of interreader agreement.—The method used to determine interreader agreement was very similar to that used for intrareader agreement. We used the method of Rousson et al (10), which is derived from the 95% limits of agreement described by Bland and Altman (9) for two arbitrary measurements. The intent was to determine the limits of agreement within which 95% of the differences between two measurements, made by two arbitrary readers, are expected to lie. From a clinical point of view, this interval corresponds to the range of differences caused by measurement error rather than a change in nodule size.

With this method, the readers are not specified and are assumed to be representative of all readers. Consequently, on average, differences between measurements will be nil. Limits of agreement are therefore symmetric around zero. In other words, the objective is not to determine the error made by two specific readers but rather to determine the measurement error made by two arbitrary readers taken from a population of readers. For this purpose, we used the means of the three values for each nodule obtained by each of the readers in this study.

RESULTS

Intrareader Agreement

The independence between the SD of the three measurements and mean size for each nodule was verified graphically for the three readers (Fig 1). The values of sw were 0.61, 0.48, and 0.54, corresponding to repeatability coefficients of 1.70, 1.32, and 1.51 mm, for readers 1, 2, and 3, respectively. In other words, to be 95% sure that a nodule had increased in size, the increase in diameter observed at follow-up CT would have to exceed 1.70, 1.32, and 1.51 mm for readers 1, 2, and 3, respectively.

When only measurements of the nodules with regular borders were considered—after exclusion of the measurements of the eight irregular, lobulated, or spiculated nodules—the repeatability coefficients were 1.60, 1.28, and 1.39 for readers 1, 2, and 3, respectively.

Interreader Agreement

The scatterplots in Figure 2 illustrate the poor agreement among the three readers.

The 95% limits of agreement of the difference between readers were −1.73 and 1.73. In clinical terms, this means that if two arbitrarily chosen readers measure the same stable nodule, in 95% of cases the differences between their measurements will lie between −1.73 and 1.73 mm. In other words, to state with 95% confidence that a nodule has truly increased in size when the measure-

ments are made by two different radiologists, a size change of more than 1.73 mm would be required.

DISCUSSION

Performing 2D measurements at follow-up CT is currently the principal method used to monitor noncalcified pulmonary nodules, especially those measuring between 5 and 10 mm. Indeed, other approaches, such as PET and contrast-enhanced CT, become less accurate with decreasing nodule size.

In previous studies, the lung cancer volume doubling time was observed to be between 30 and 490 days in a series of 67 patients and has generally been estimated to be around 100 days (11,12). Twenty-two percent of stage I lung carcinomas doubled in volume after 465 days or more in a study by Winer-Muram et al (13). However, indolent tumors might have been overrepresented in the relatively elderly population evaluated in that study.

The doubling time can be estimated from the difference in nodule diameter between baseline and follow-up CT and the time interval between the two examinations by using a simple exponential growth model that assumes uniform three-dimensional tumor growth.

The Early Lung Cancer Action Project group recommended repeat CT examination at 3, 6, 12, and 24 months for stable nodules measuring between 5 and 10 mm (1). However, this assumes that 2D measurements are reliable in terms of intrareader agreement (agreement of measurements made by the same reader) or interreader agreement (agreement of measurements made by different readers). In the present study, our aim was to evaluate 2D measurement error. We evaluated measurement of a single dimension on 2D images; although measurement of two orthogonal dimensions might have reduced the 2D measurement error, it would have increased the number of measurements needed per reader from three to six per nodule. We chose the maximal transverse cross-sectional diameter, which is the most commonly used nodule measurement at most institutions.

How to organize the readings was the most difficult aspect of the study design. The choice of three separate reading sessions seemed too different from daily practice. When one is measuring a nodule on two different CT scans, measurements are made consecutively for reasons of comparability. The question that led to the present study was: Is it possible to reliably estimate the size variation of a nodule with manual measurements of transverse cross-sectional diameters? The first condition of this question is that the measurements must be repeatable, and, to determine if this is the case, the measurements must be made consecutively—otherwise, they cannot be made identically. However, if the measurements are made with hardly any time interval between them, the reader will remember the previous value and tend to reproduce it, thus minimizing variability. This is why we decided to have the readers make

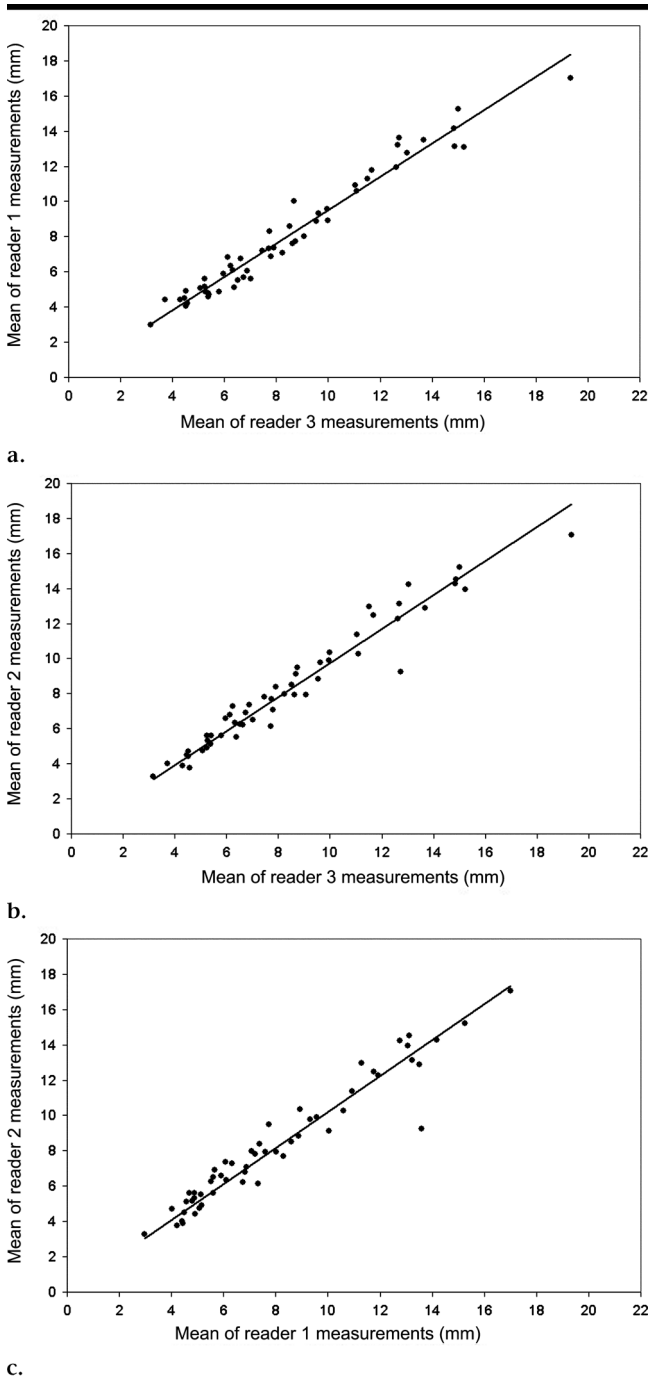


Figure 2. **(a–c)** Scatterplots illustrate agreement of the measurements for all possible pairs of readers. The poor agreement among the three readers is demonstrated by points lying outside the line of equality.

the three measurements of each nodule during the same reading session, with an interval of several minutes between each measurement, and to group the nodules together in groups of three or four and have the readers make consecutive first measurements of all these nodules, with this first reading session followed by the second and then the third reading session.

Wormanns et al (14) observed good interobserver agreement for categorization of pulmonary nodules in three size classes at spiral CT. They also found good agreement regarding exact nodule size and concluded that spiral CT enabled reproducible determinations of pulmonary nodule size (14). However, this second conclusion was not authorized by their statistical approach, as they used the Pearson correlation coefficient, which is not appropriate for calculating agreement (15). Indeed, with the Pearson method, a perfect correlation can be obtained even if one reader's values are consistently 50% higher than those of a second reader.

We found that both intra- and interreader agreement for 2D measurement of nodule size on CT scans was poor. The most consistent of the three readers had a minimum measurement error of 1.32 mm, meaning that there was only a 5% likelihood that a difference of 1.32 mm or less between two serial measurements by this reader would correspond to an actual change in size. Likewise, when two serial measurements were made by two different readers in our study, an apparent change in size of less than 1.73 mm had only a 5% chance of corresponding to a real change in size. This is a large margin of error in that the measurement error is 10% or more of the nodule diameter, introducing an even larger error in the resulting estimates of volume and doubling time. This poor level of intra- and interreader agreement was observed despite the fact that the CT parameters were optimized and standardized: Measurements were made directly on PACS screens, with identical window settings and high-spatial-resolution-algorithm–reconstructed images; in addition, the readers were strongly advised to zoom in on the nodules for optimal analysis.

Although, as expected, intrareader agreement was better than interreader agreement, it was still inadequate for reliably identifying nodule size changes and making subsequent patient care decisions. The best reader had a variability of 1.32 mm, meaning that stable nodules could be mistaken for growing lesions even in ideal working conditions. Indeed, an apparent size increase from 5.0 to 6.3 mm at a 3-month interval that resulted from measurement error would correspond to a doubling time of 105 days, which is typical of malignant lesions. With nodules measuring 10 and 15 mm at baseline, the 90-day doubling times corresponding to this degree of error would be, respectively, 170 and 250 days—doubling times that again are in keeping with malignant growth. The 2D measurement error should not exceed 0.4 mm for a stable 10-mm nodule: A size increase from 10.0 to 10.4 mm after 3 months corresponds to a doubling time of 530 days, whereas the generally accepted upper limit of doubling times for malignant pulmonary lesions is 500 days.

Although this was a single-center study, the same degree of inaccuracy would probably be encountered elsewhere with interpretations that involve similar multisection CT and PACS workstation equipment. The situation might be even worse when low-dose CT is used, because the lower signal-to-noise ratio with low-dose CT potentially makes it more difficult to identify nodule borders.

Staron and Ford (16) found that repeated measurements of cross-sectional area by a single observer varied by about ±5% to ±20%, depending on the size of the object. Likewise, Winer-Muram et al (13) reported that the within-observer error seen with different volume-estimating methods increased as tumor size decreased.

We observed no linear relationship between SD and nodule size, possibly because not enough nodules were studied or because the nodule size range was too limited. However, 70% of the nodules in the present study measured between 5 and 15 mm in diameter—a size range at which CT follow-up is generally required because only 1% of nodules smaller than 5 mm and as many as 80% of nodules larger than 20 mm are malignant (1).

Another drawback of 2D CT measurements in this setting is that growth is a three-dimensional phenomenon. Yankelevitz et al (17), in a study of techniques for assessing the growth rate of pulmonary nodules in three dimensions, found that some malignant nodules showed asymmetric growth that was not detected by using 2D techniques.

Our study had several limitations: We did not include all types of nodules (especially ground-glass and part-solid nodules). We believed that these nodule types, which occur less frequently, would be more difficult to measure, and that this would negatively affect the results of the study. This is why we preferred to focus on the reliability of 2D measurements of solid nodules. In addition, we did not evaluate the influence of nodule shape on the reliability of 2D measurements; we were therefore unable to determine whether reliability was worse for measurements of irregular or spiculated nodules, which represented only 15% of the nodules in this series. However, when we excluded measurements of irregular nodules at statistical analysis, our results were not really modified in that the repeatability coefficients remained quite similar.

We did not evaluate the influence of pixel size (determined by the acquisition field of view) on measurement error. This would offer important information but would require a specific evaluation. For each measurement of a single nodule, readers were asked to select the CT image that showed the nodule's largest transverse cross-sectional diameter, but the number of the CT section on which the reader performed the measurements was not recorded for each nodule. Thus, we cannot be sure that all nine measurements were performed at the same level, and not performing all nine measurements at the same level would tend to increase the measurement error.

Another limitation was that all three readers knew the purpose of the study, and this may have influenced the way in which they made the measurements. However, this would have tended to minimize the measurement error, which would not really have posed a problem in that our objective was to estimate the minimal 2D measurement error. Thus, even with readers who knew they were participating in a repeatability study, the 2D measurement error was 1.32 mm for the best reader.

Two-dimensional measurements at CT appear to be unreliable in the evaluation of small noncalcified pulmonary nodules, especially in view of the poor intrareader agreement observed in this study. Measurement error could lead to erroneous growth estimations during follow-up CT examinations, with a risk that unwarranted invasive investigations will be performed or, conversely, that malignant growth will not be identified.

The observed lack of 2D measurement reliability favors the use of volumetric measurements of small nodules performed with direct software calculation instead of estimates of volume that are based on 2D measurements.

Acknowledgments: We thank Professor Gilles Chatellier, MD, of Clinical Investigation Center 9201, Assistance Publique des Hôpitaux de Paris/INSERM, Georges Pompidou European University Hospital, Paris, France for his advice on the statistical analysis of the data; David Young for editorial assistance; and Joelle Bauvillard for her help in preparing the manuscript.

References

1. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early Lung Cancer Action Project: Overall design and findings from baseline screening. Lancet 1999; 354:99–105.
2. Swensen SJ, Jett JR, Sloan JA, et al. Screening for lung cancer with low-dose spiral computed tomography. Am J Respir Crit Care Med 2002; 165:508–513.
3. Boring CC, Squires TS, Tong T. Cancer statistics, 1993. CA Cancer J Clin 1993; 43:7–26.
4. Coleman RE, Laymon CE, Turkington TG. FDG imaging of lung nodules: a phantom study comparing SPECT, camera-based PET, and dedicated PET. Radiology 1999; 210:823–828.
5. Swensen SJ, Viggiano RW, Midthun DE et al. Lung nodule enhancement at CT: multicenter study. Radiology 2000; 214:73–80.
6. Li H, Boiselle PM, Shepard JO, Trotman-Dickenson B, McLoud TC. Diagnostic accuracy and safety of CT-guided percutaneous needle aspiration biopsy of the lung: comparison of small and large pulmonary nodules. AJR Am J Roentgenol 1996; 167:105–109.
7. British Standards Institution. Precision of test methods, part I: guide for determination of repeatability and reproducibility for a standard test method. London, England: BSI, 1979. BS 5497, part I.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307–310.
9. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8:135–160.
10. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Stat Med 2002; 21:3431–3446.
11. Steele JD, Buell P. Asymptomatic solitary pulmonary nodules: host survival, tumor size, and growth rate. J Thorac Cardiovasc Surg 1973; 65:140–151.
12. Geddes DM. The natural history of lung cancer: a review based on rates of tumour growth. Br J Dis Chest 1979; 73:1–17.
13. Winer-Muram HT, Jennings SG, Tarver RD, et al. Volumetric growth rate of stage I lung cancer prior to treatment: serial CT scanning. Radiology 2002; 223:798–805.
14. Wormanns D, Diederich S, Lenstchig MG, Winter F, Heindel W. Spiral CT of pulmonary nodules: interobserver variation in assessment of lesion size. Eur Radiol 2000; 10:710–713.
15. Halligan S. Reproducibility, repeatability, correlation and measurement error. Br J Radiol 2002; 75:193–194; discussion, 194–195.
16. Staron RB, Ford E. Computed tomographic volumetric calculation reproducibility. Invest Radiol 1986; 21:272–274.
17. Yankelevitz DF, Reeves AP, Kostis WJ, Zhao B, Henschke CI. Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation. Radiology 2000; 217:251–256.