Lung Cancer: Interobserver Agreement on Interpretation of Pulmonary Findings at Low-Dose CT Screening[1]

David S. Gierada, MD
Thomas K. Pilgram, PhD
Melissa Ford, PhD
Richard M. Fagerstrom, PhD
Timothy R. Church, PhD
Hrudaya Nath, MD
Kavita Garg, MD
Diane C. Strollo, MD

Purpose: To evaluate agreement among radiologists on the interpretation of pulmonary findings at low-dose computed tomographic (CT) screening examinations for lung cancer.

Materials and Methods: Institutional review board approval and informed consent were obtained. HIPAA guidelines were followed. Sixteen radiologists from the 10 National Lung Screening Trial screening centers of the National Cancer Institute's Lung Screening Study network reviewed image subsets from 135 baseline low-dose screening CT examinations in 135 trial participants (89 men, 46 women; mean age, 62.7 years ± 5.4 [standard deviation]). Interpretations were classified into one of four of the following categories: noncalcified nodule 4 mm or larger in greatest transverse dimension (positive screening result); noncalcified nodule smaller than 4 mm in greatest transverse dimension (negative screening result); calcified, benign nodule (negative screening result); or no nodule (negative screening result). A recommendation for follow-up evaluation was obtained for each case. Interobserver agreement was evaluated by using the multirater κ statistic and by using response frequencies and descriptive statistics.

Results: Multirater κ values ranged from 0.58 (for agreement among all four classifications; 95% confidence interval: 0.55, 0.61) to 0.64 (for agreement on classification as a positive or negative screening result; 95% confidence interval: 0.62, 0.65). The average percentage of reader pairs in agreement on the screening result per case (percentage agreement) was 82%. There was wide variation in the total number of abnormalities detected and classified as pulmonary nodules, with differences of up to more than twofold among radiologists. For cases classified as positive, multirater κ for follow-up recommendations was 0.35.

Conclusion: Interobserver agreement was moderate to substantial; potential for considerable improvement exists.

Clinical trial registration no. NCT00047385

[1] From the Mallinckrodt Institute of Radiology, Washington University School of Medicine, 510 S Kingshighway Blvd, St. Louis, MO 63105 (D.S.G., T.K.P.); Westat Corporation, Rockville, Md (M.F.); the National Cancer Institute, Bethesda, Md (R.M.F.); the University of Minnesota Medical School, Minneapolis (T.R.C.), the University of Alabama at Birmingham School of Medicine (H.N.); the University of Colorado Health Sciences Center, Denver (K.G.); and the University of Pittsburgh School of Medicine, Pa (D.C.S.).
From the 2005 RSNA Annual Meeting. Received December 8, 2006; revision requested February 20, 2007; revision received April 4; accepted May 4; final version accepted June 1. Supported by the National Cancer Institute Contract N01-CN-25516. **Address correspondence to** D.S.G. (e-mail: *gieradad@wustl.edu*).

Radiology

Although screening for early detection of lung cancer is not currently recommended by any major medical organization, the ability to routinely identify tumors smaller than 1 cm with low-radiation-dose spiral computed tomographic (CT) scanning has led to much enthusiasm for CT screening as a potential means of reducing mortality from this usually fatal disease. Consequently, the efficacy of low-dose CT screening has been under intense investigation (1–9).

At low-dose CT screening, small, nodular, usually benign but indeterminate pulmonary lesions are identified frequently; therefore, the rate of positive examination findings requiring a follow-up evaluation can be high (10). Findings at repeat screening CT (11,12) suggest that there also may be a considerable nodule miss rate. Thus, in addition to the technical capabilities of CT to depict lung abnormalities, screening examination results and subsequent actions taken depend on the interpretation of the radiologist. Despite this, to our knowledge, the issue of variability in the interpretation of screening CT scans has received little attention.

The primary tasks involved in the interpretation of images obtained with lung screening CT are to identify focal pulmonary lesions; discriminate potentially neoplastic, noncalcified nodules from benign lesions such as calcified granulomas, scars, inflammatory processes, and pleural plaques; and measure the size of noncalcified nodules to help assess the risk of malignancy. Each task involves some degree of subjectivity, and variability among readers is inherent at each of these steps of the screening process. Thus, our study was performed to evaluate agreement among radiologists on the interpretation of pulmonary findings at low-dose CT screening examinations for lung cancer.

Advances in Knowledge

- Interobserver agreement for classification of screening findings as measured with the κ statistic (κ = 0.58–0.64) was similar to agreement found in previous studies for mammography and other CT interpretive tasks.
- Relatively wide variation was seen among some reader pairs in the percentage of studies considered to be positive and in the overall number of nodules detected (up to twofold or greater differences for both); these variations may be related to variation in lesion detection, lesion classification as a nodule or nonnodule, or lesion measurement.

Implications for Patient Care

- With some lesions, classification of screening findings as positive or negative is not a straightforward task and may depend on the individual judgment of the radiologist.
- Identification of reliable, objective criteria for distinguishing definitively benign from indeterminate lesions may help improve interobserver agreement.

Materials and Methods

Participants

All participants were enrolled in the National Lung Screening Trial (NLST) (*http://www.cancer.gov/nlst,* clinicaltrials.gov identifier NCT00047385), a multicenter research protocol approved by the institutional review boards of all participating centers. Informed consent was obtained, including consent to use deidentified CT images and data for research purposes. Health Insurance Portability and Accountability Act guidelines were followed. All protected health information and site identifiers were removed electronically.

The primary aim of the NLST is to compare the lung cancer mortality rates of high-risk individuals randomly assigned to undergo either three annual low-dose screening CT scans or three annual posteroanterior screening chest radiographic examinations (13). In the NLST, 53 472 volunteers between the ages of 55–74 years with a smoking history of 30 pack-years or more have been randomly assigned to low-dose CT or chest radiographic screening arms. The trial is sponsored by the National Cancer Institute and is being conducted at the 10 screening centers of the Lung Screening Study (LSS) (7,8) trial network and at 23 screening centers in the American College of Radiology Imaging Network (14). Enrollment and baseline screening occurred between September 2002 and April 2004.

Screening CT Examination Selection Technique

A total of 135 screening CT examinations in 135 trial participants were selected retrospectively from the LSS-NLST database of the first 8365 baseline screening CT examinations performed at the 10 LSS-NLST screening centers (see Appendix). The selection was stratified according to the findings recorded by the radiologists who originally interpreted the screening CT examination results for the trial (Table). We randomly selected 75 of the 135 examinations from among those that contained at least one noncalcified nodule 4 mm or larger in greatest transverse dimension recorded in the NLST database. According to the NLST protocol (and for our study), such examinations had to be classified as having a positive screening result. The other 60 screening CT examinations were randomly chosen from

Published online before print
10.1148/radiol.2461062097

Radiology 2008; 246:265–272

Abbreviations:
ELCAP = Early Lung Cancer Action Project
LSS = Lung Screening Study
NLST = National Lung Screening Trial

Author contributions:
Guarantor of integrity of entire study, D.S.G.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, D.S.G., T.R.C.; clinical studies, M.F., D.C.S., T.K.P., K.G.; statistical analysis, D.S.G., T.K.P., M.F., R.M.F., T.R.C.; and manuscript editing, D.S.G., T.K.P., M.F., T.R.C., K.G., D.C.S.

Clinical trial registration no. NCT00047385

Authors stated no financial relationship to disclose.

among examinations that had been classified as having a negative screening result. Twenty were chosen from among those that contained at least one noncalcified nodule smaller than 4 mm in greatest transverse dimension that had been recorded in the database; 20 were chosen from among those that contained at least one nodule containing a benign calcification that had been recorded; and 20 were chosen from among those in which no nodules had been recorded.

For screening CT examinations in the category of noncalcified nodule 4 mm or larger, the presence or absence of other recorded nodules of any size was not a selection criterion. Examinations in the category of noncalcified nodule smaller than 4 mm could contain additional recorded nodules from the noncalcified nodule smaller than 4 mm or benign calcification categories (but not from the noncalcified nodule 4 mm or larger category), while examinations selected as part of the benign calcification category had no other noncalcified nodules recorded. To allow analysis of the full range of lesion sizes, oversampling of the less frequent, relatively larger nodules was performed by randomly selecting 16 of the 75 cases in the noncalcified nodule 4 mm or larger category from the subset of lesions for which a greatest transverse dimension of 8 mm or larger had been recorded. The final cohort was composed of 89 men and 46 women (mean age, 62.7 years ± 5.4 [standard deviation]; age range, 55–74 years) who each underwent one baseline screening examination.

Spiral multidetector scanning with at least four detector rows and low radiation dose had been performed in all cases. No intravenous contrast material was administered. Acquisition parameters included 120–140 kVp and 20–60 mAs (effective) (effective tube current = tube current/pitch, where pitch = table feed rate/[number of detectors · detector collimation]). Reconstructed section thickness was 2.5 mm or less, and reconstructed sections were contiguous or overlapping in the transverse plane.

Screening CT Examination Image Subsets

To optimize reader efficiency, subsets of contiguous images were chosen from each screening CT examination to include the lesion for which the examination was selected (lesion of interest). Of the 135 subsets, there were 108 (80%) with 12 images, 20 (15%) with 14–20 images, six (4%) with 24–30 images, and one (1%) with 40 images. In 14 of the 27 subsets with more than 12 images, CT images had been reconstructed at overlapping 1.25-mm intervals; hence, more than 12 images were presented to maintain the same total slab thickness as the 12-image subsets, in which contiguous reconstructions of 2.0- or 2.5-mm section thickness had been performed. In the other 13 subsets, more than 12 images were presented so that the first or last image contained no visible abnormality or incompletely shown structure that might be interpreted as an abnormality. Care was taken to ensure that the presentation of more than 12 images was not associated with a positive screening result at CT examination. Of the 27 subsets presented with more than 12 sections, 18 (67%) had been selected because of an original positive diagnosis; this proportion was relatively similar to the 75 (56%) examinations of the entire 135-examination set that had been selected because of an original positive diagnosis.

The section level of the lesion of interest was randomly varied among the image subsets. For examinations with no nodule, the cephalocaudal level of the lung from which the image subset was selected was randomized. The readers were not informed of the composition of the test set or of the reasons for variation in the number or location of sections presented. The case presentation order was randomized and was identical for all readers.

Readers

Sixteen radiologists from all 10 LSS-NLST screening centers who regularly interpret NLST screening CT images (including H.N. and K.G.) participated as readers. The average reader experience in interpreting CT scans was 18 years ± 7 (range, 5–30 years). Ten of the radiologists were in academic practice as thoracic radiology subspecialists, and six were in private practice as general radiologists. The readers were aware of the purpose of our research study but were blinded to the criteria for selecting the screening examinations used in the study and to the original interpretations. They were instructed to interpret the findings in the same manner as when they interpret a baseline NLST scan for which there is no comparison study. All readers had viewed an NLST training slide presentation before they read any screening im-

Screening Examination Categories Used in Case Selection

Category	Additional Case Selection Features	Screening Result*	No. of Cases
Noncalcified nodule with greatest transverse dimension 4 mm or larger	With or without noncalcified nodule smaller than 4 mm or benign calcification	Positive	75
Noncalcified nodule with greatest transverse dimension smaller than 4 mm	No noncalcified nodule 4 mm or larger or benign calcification	Negative	20
Benign calcification of any size	No noncalcified nodule	Negative	20
No nodule		Negative	20

* The NLST protocol defines cases with any noncalcified nodule 4 mm or larger as positive screening results and cases with no noncalcified nodule 4 mm or larger or other abnormalities suspicious for lung cancer (such as lobar collapse or lymph node enlargement) as negative screening results.

ages for the NLST. This slide presentation, which was produced by the American College of Radiology Imaging Network–NLST group, defined and showed examples of lesions having various features, such as soft-tissue attenuation; ground-glass attenuation; smooth margins; spiculated margins; and pseudonodules such as linear bands of atelectasis or scarring, bronchiolar inflammation, and dependent atelectasis.

Image Viewing

All readers viewed the images from compact discs in Digital Imaging and Communications in Medicine format with the same models of desktop computer (Precision 340; Dell) and flat-panel liquid crystal display color monitor (SDM-P82; Sony) (15) after proper gray-scale calibration had been confirmed by viewing a test pattern. The same image browser and viewer (Sienet MagicView 300; Siemens), which had tools for image magnification, electronic caliper measurement, and adjustment of center and window display settings, were used by all readers.

Data Collection

Abnormalities were recorded by readers on a single-page condensed version of the screening form used in the NLST. Data recording differed from the NLST protocol in that readers were instructed to ignore abnormalities other than lung nodules, or any other findings suspicious for lung cancer, and to record the section number and location (by lobe) of all nodules, regardless of size or presence of calcification. The readers recorded the longest transverse dimension of each noncalcified nodule 4 mm or larger and made a recommendation for follow-up for each positive case from among the same options used for NLST readings. The follow-up options included a chest radiographic examination at an interval recommended by the reader; a low-dose unenhanced CT scan after 3, 3–6, 6, or 12 months or other interval specified by the reader; an immediate diagnostic imaging study (contrast material–enhanced CT or positron emission tomographic [PET] scan); lung biopsy; or other method specified by the reader.

Data and Statistical Analyses

All data analyses were performed by consensus of two authors (D.S.G., with 12 years of experience interpreting chest CT images and T.K.P., a statistician). As in the NLST, all studies with at least one noncalcified nodule 4 mm or larger recorded by the reader were classified as positive in the analysis. Studies with no recorded noncalcified nodules 4 mm or larger were classified as negative. Interobserver agreement was assessed on a case-diagnosis basis primarily by using multirater κ, which measures the level of agreement after taking chance agreement into account (16,17). In addition, the overall percentage agreement, or average percentage of reader pairs in agreement on a positive or negative diagnosis per case (17), and positive and negative agreement, or the average percentage of readings in which each reader pair agreed that a diagnosis was positive or negative, respectively (16), were calculated. The number of readers in agreement on a positive or negative diagnosis for each case and the individual reader response frequencies of each case-diagnosis category and each nodule category were determined. The percentage of readers who classified a lesion of interest as a noncalcified nodule 4 mm or larger was assessed as a function of mean lesion size. Finally, to determine whether there were any patterns to the disagreements that occurred, the nodule classifications recorded by each reader for the cases in which fewer than 12 radiologists agreed on a positive or negative finding were tabulated.

Agreement on follow-up recommendations for a positive screening result was assessed with multirater κ by using the cases in which 15 or 16 readers agreed with the original interpretation on the presence of a noncalcified nodule 4 mm or larger. The possible follow-up responses were categorized as "now" when an immediate diagnostic procedure was recommended (eg, diagnostic contrast-enhanced CT, PET scan, biopsy) or as "later" when a follow-up chest radiographic examination or CT scan to evaluate lesion growth was recommended. Calculations were made by using software (JMP, SAS Institute, Cary, NC; Excel; Microsoft, Redmond, Wash).

Results

Interobserver Agreement

Multirater κ values ranged from 0.58 to 0.64, depending on how the categories were grouped. Agreement was highest ($\kappa = 0.64$; 95% confidence interval: 0.62, 0.65) for the classification of cases as either a positive (noncalcified nodule 4 mm or larger) or negative (all other classifications) screening result. Values were lower ($\kappa = 0.60$; 95% confidence interval: 0.57, 0.62) for the distinction between any noncalcified nodule (noncalcified nodule ≥ 4 mm or noncalcified nodule < 4 mm) and no noncalcified nodule (benign calcification or no nodule) and for distinction among all four categories ($\kappa = 0.58$; 95% confidence interval: 0.55, 0.61). The κ values among all reader pairs for all cases ranged from 0.40 to 0.82 for the distinction between a positive and negative screening result, with an interquartile range (25th–75th percentile) of 0.59–0.70.

Positive versus Negative Screening Result

The overall percentage agreement on a positive versus negative screening result was 82% of reader pairs. All 16 readers agreed in 60 (44%) of the 135 cases, 14–15 readers agreed in 32 (24%) cases, 11–13 agreed in 26 (19%) cases, and eight to 10 agreed in 17 (13%) cases. The average positive agreement for all reader pairs was 83% (range, 64%–92%), and average negative agreement was 81% (range, 68%–92%). The individual readers varied substantially in the percentage of cases classified as positive (Fig 1a), with a mean of 53% ± 9 classified as positive (range, 33%–66%). Similarly, there was substantial variation in the total number of nodules recorded, with some readers identifying more than twice as many noncalcified nodules 4 mm or

larger as others (mean, 93 nodules ± 22) and some identifying several times as many noncalcified nodules smaller than 4 mm as others (mean, 41 nodules ± 23) (Fig 1b).

Lesion Size and Agreement

The effect of lesion size on agreement was assessed for the 75 noncalcified nodules reported as 4 mm or larger and for the 20 noncalcified nodules reported as smaller than 4 mm by the radiologist who originally read the images (lesions of interest). The percentage of readers who recorded a lesion of interest as a noncalcified nodule 4 mm or larger increased with mean lesion size (Fig 2). Complete agreement that a nodule was 4 mm or larger was seen when the average reader measurement reached 6 mm in greatest transverse dimension. Review of the 23 cases in which fewer than 12 radiologists agreed on a positive or negative interpretation revealed no dominant patterns of disagreement: in eight of the 23 cases, diagnoses were divided between noncalcified nodule 4 mm or larger, noncalcified nodule smaller than 4 mm, and no nodule (Fig 3a); in four cases, diagnoses were divided between noncalcified nodule 4 mm or larger and no nodule (Fig 3b); and in five cases, diagnoses were divided between noncalcified nodule 4 mm or larger and noncalcified nodule smaller than 4 mm. In the other six cases, diagnoses included benign calcification and one, two, or three of the other diagnoses.

Follow-up Recommendations

At least 15 readers agreed with the original reading in 38 of the 75 cases in which the original reading was a noncalcified nodule 4 mm or larger; multirater κ for the "now" versus "later" categorization of follow-up recommendations for these cases was 0.35.

Discussion

Quantification of observer agreement is an essential complement to conventional studies of diagnostic accuracy in the evaluation of a diagnostic test (18). According to criteria commonly used for the interpretation of κ values (19), interobserver agreement for the classification of screening findings in our study was moderate to substantial and was similar for positive and negative interpretations. Reader variability could have occurred at lesion detection, characterization of a lesion as a nodule or nonnodule, and lesion measurement.

Variation in measurement accounted for much of the disagreement in the classification of studies with nodules near the 4-mm size threshold as positive or negative. This was expected considering the irregular shape and indistinctness of the margins of many pulmonary nodules, which affect the location of electronic cursor placement. Other study results (20–22) have shown considerable variation in two-dimensional lung nodule size measurements. Some investigators of low-dose CT screening studies (9,12,23,24) classified noncalcified nodules of any size as positive screening results, although the measured nodule size still influenced the

Figure 2

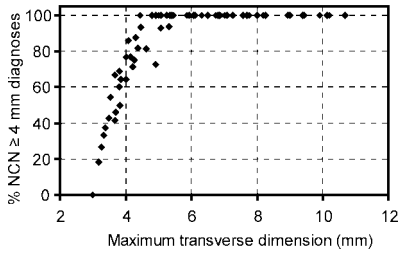

Figure 2: Graph shows percentage of readers who measured each of the 95 lesions of interest as a noncalcified nodule *(NCN)* 4 mm or larger. Dimensions are means of values recorded by readers who identified each nodule, using a value of 3 mm when a nodule was recorded as a noncalcified nodule smaller than 4 mm. Percentages are based on number of readers who identified each nodule. Mean number of radiologists who identified and classified each lesion of interest as a pulmonary nodule was 12.8 ± 4.0.

Figure 1

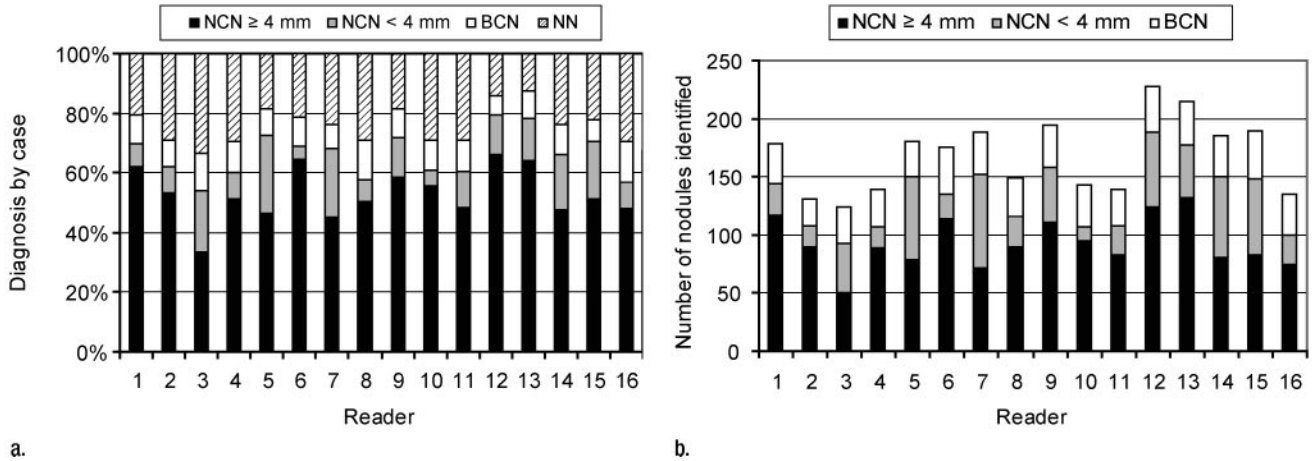

Figure 1: Bar graphs of diagnostic frequencies for individual readers show **(a)** variation in final case diagnosis and **(b)** number of nodules recorded among readers. *BCN* = benign calcified nodule, *NCN* = noncalcified nodule, *NN* = no nodule.

suspicion of malignancy and subsequent work-up. In our study, agreement on the presence or absence of any noncalcified nodule (κ = 0.60) was virtually the same as agreement on the classification of a screening result as positive or negative on the basis of a 4-mm size threshold (κ = 0.64). It was not possible to distinguish detection (lesion vs no lesion) from characterization (unimportant finding vs indeterminate nodule) disagreements because readers did not record the lesions they saw but then dismissed as benign processes (eg, scarring or inflammation).

The level of agreement in our study is lower than that in the initial Early Lung Cancer Action Project (ELCAP) cohort, which had a κ value of 0.91 for two readers from a single institution (23). The difference from our study may be due in part to the different methods. Screening positivity in the ELCAP study did not depend on the size of noncalcified nodules, so measurement variation was not a potential source of disagreement. In addition, the greater section thickness of 10 mm used in this initial ELCAP study likely limited the depiction of smaller lesions that may result in variable detectability.

Another study (25) revealed lower case-based agreement than in our study, with κ values of 0.23–0.46 among three reader pairs at a single institution. The low percentage of negative examination findings (6.8% of findings were classified as a negative finding by all three readers) in that study may have contributed to this lower agreement, because κ may be reduced if one classification category dominates (17). (However, this apparently was not an important factor for the ELCAP study, in which 76.7% of the screening examinations had negative findings.) Their review of entire screening CT examinations (25) may explain the lower agreement in part, because viewing more sections may increase the chance of some readers recording abnormalities that others would not detect or would not classify as nodules. (However, the higher agreement found in the ELCAP study also was based on the review of complete scans.) That study (25) revealed even lower agreement for a nodule-based analysis, with κ of 0.12 or less. We too observed substantial differences in the total number of nodules noted by different readers, despite frequent agreement on whether the case findings were positive or negative.

One advantage of our study compared with both of these studies (23,25) was its larger number of readers. With few readers in an agreement study, there is a greater chance that the readers will have very similar or dissimilar reading styles and will thus have high or low agreement, respectively. Indeed, we found reader pairs among whom agreement was nearly as high and nearly as low as in the two comparison studies (23,25).

The level of agreement in our study is similar to or better than that found with the classification of screening mammograms, for which κ values of 0.47 (26) and 0.58 (27) have been reported. It was also similar to that of other CT interpretive tasks. For example, κ values for CT diagnoses of cystic renal masses (28), the etiology of diffuse lung disease (29), and deep venous thrombosis (30,31) in the range of 0.5–0.6 have been reported. For the diagnosis of pulmonary embolism, κ values tend to be greater than 0.7 and higher with multidetector than with single-detector scanning but are lower for smaller segmental vessels (32).

Agreement on follow-up recommendations for positive screening results was only fair. This likely reflects the NLST practice of allowing radiologists

Figure 3

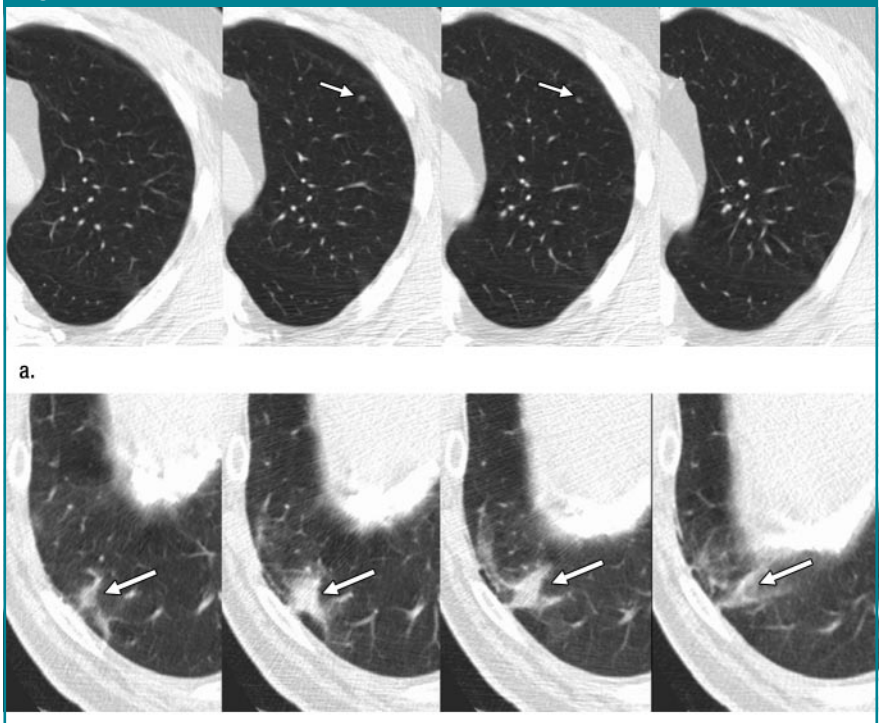

Figure 3: CT images of cases in which fewer than 12 readers agreed on diagnosis. **(a)** Small left upper lobe opacity (arrows) was classified as a noncalcified nodule 4 mm or larger by six readers, noncalcified nodule smaller than 4 mm by seven readers, and no nodule by three readers. Of those who recorded the nodule, seven readers measured it as smaller than 4 mm, three as 4 mm, and three as 5 mm. **(b)** In another case, irregular right lower lobe opacity (arrows) adjacent to calcified diaphragmatic pleural plaque was classified as noncalcified nodule 4 mm or larger by 10 readers and no nodule by six readers. Eleven readers commented regarding probable atelectasis, scar, and/or asbestos exposure (seven classified the case as having noncalcified nodule 4 mm or larger and four classified the case as having no nodule).

the discretion to make recommendations within a range of options, according to acceptable medical practice. Literature guidelines for the follow-up and management of pulmonary nodules on the basis of size criteria proposed before (33) and after (34) the NLST began have evolved as additional screening trial data (35) have become available. Our observations suggest that with guidelines based primarily on nodule size, it may be difficult to achieve consistent agreement on follow-up recommendation because of measurement variation, particularly near the size thresholds at which the recommended management changes. Furthermore, the various lesion morphologies encountered may lead to differences in the suspicion of malignancy.

Short of seamlessly inserting copies of the same imaging studies into the clinical workflow of multiple readers, any reader agreement study design inevitably creates an artificial experimental setting that limits the study in some manner. Consequently, one limitation of our study was that readers may have behaved differently than in daily practice; some may have been more careful in the testing situation, while others may have been less careful because their performance had no clinical consequences. Using subsets of images may have reduced opportunities for disagreement on each case, but it allowed assessment of findings from a relatively large number of cases by numerous radiologists in an efficient and controlled manner. Requiring readers to concurrently detect and classify abnormalities limited the ability to analyze agreement for each of these tasks independently but more closely simulated actual clinical practice than if these tasks had been divided.

Our study results illustrate that, despite the detailed depiction of lung parenchyma provided by using the screening CT protocol, the interpretation of pulmonary findings is a complex task. The variation in size, location, and morphology of lesions likely hinders the ability to obtain perfect agreement on lesion detection and classification. Although most readers agree on the majority of findings, there is substantial room for improvement. Computer-aided programs that assist in the detection of lesions may improve reader performance (36,37) and hold promise as a means of reducing observer variability. Semiautomated volumetric determination of lesion size may reduce variation related to nodule measurement (38,39). Further development and validation of objective, evidence-based nodule characterization criteria (40) and automated nodule characterization algorithms (41) also may help increase agreement at screening CT interpretation.

Appendix

The 10 screening centers of the LSS-NLST network and their National Cancer Institute contract numbers are The University of Alabama at Birmingham (N01-CN-75022); University of Colorado Health Sciences Center (N01-CN-25514); Georgetown University (N01-CN-25522); Henry Ford Hospital (N01-CN-25512); Marshfield Clinic (N01-CN-25518); University of Minnesota (N01-CN-25513); Pacific Health Research Institute (N01-CN-25515); University of Pittsburgh (N01-CN-25511); the University of Utah with a satellite center at St Luke's Meridian Medical Center in Boise, Idaho (N01-CN-25524); and Washington University in St Louis (N01-CN-25516). Coordinating and statistical services for the LSS-NLST, including the database search for the NLST participant CT screening examinations used in this study, was provided by Westat Corporation (Rockville, Md) (N01-CN-25476).

Acknowledgments: The authors thank Peter Balkin, Sanjeev Bhalla, Matthew T. Freedman, Harvey Glazer, Fernando Gutierrez, William Herbick, Subi Inampudi, Howard Mann, William Manor, Stuart Sagel, Satinder Singh, David Spizarny, John Waltz, and Patrick Wolfe for participating as readers in this study. We thank Matthew T. Freedman, David Lynch, and Paul Pinsky for their helpful reviews of an early draft of the manuscript. We acknowledge the assistance and support of LSS-NLST Quality Assurance Working Group members Ken Clark, Kathy Clingan, Glenn Fletcher, Mike Flynn, Randall Krueger, Fred Larke, Guillermo Marquez, Tom Payne, Pete Ohan, and Xizeng Wu, and the technical assistance of Kirk Smith.

References

1. Kaneko M, Eguchi K, Ohmatsu H, et al. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. Radiology 1996;201:798–802.
2. Henschke CI, Lee IJ, Wu N, et al. CT screening for lung cancer: prevalence and incidence of mediastinal masses. Radiology 2006;239:586–590.
3. Sone S, Li F, Yang ZG, et al. Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner. Br J Cancer 2001;84:25–32.
4. Swensen SJ, Jett JR, Hartman TE, et al. CT screening for lung cancer: five-year prospective experience. Radiology 2005;235:259–265.
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;14:691–702.
6. Pastorino U, Bellomi M, Landoni C, et al. Early lung-cancer detection with spiral CT and positron emission tomography in heavy smokers: 2-year results. Lancet 2003;362:593–597.
7. Gohagan J, Marcus P, Fagerstrom R, Pinsky P, Kramer B, Prorok P. Baseline findings of a randomized feasibility trial of lung cancer screening with spiral CT scan vs chest radiograph: the Lung Screening Study of the National Cancer Institute. Chest 2004;126:114–121.
8. Gohagan JK, Marcus PM, Fagerstrom RM, et al. Final results of the Lung Screening Study, a randomized feasibility study of spiral CT versus chest X-ray screening for lung cancer. Lung Cancer 2005;47:9–15.
9. MacRedmond R, Logan PM, Lee M, Kenny D, Foley C, Costello RW. Screening for lung cancer using low dose CT scanning. Thorax 2004;59:237–241.
10. Humphrey LL, Teutsch S, Johnson M. Lung cancer screening with sputum cytologic examination, chest radiography, and computed tomography: an update for the U.S. Preventive Services Task Force. Ann Intern Med 2004;140:740–753.
11. Henschke CI, Naidich DP, Yankelevitz DF, et al. Early lung cancer action project: initial findings on repeat screenings. Cancer 2001;92:153–159.
12. Swensen SJ, Jett JR, Sloan JA, et al. Screening for lung cancer with low-dose spiral com-

puted tomography. Am J Respir Crit Care Med 2002;165:508–513.

13. Church TR. Chest radiography as the comparison for spiral CT in the National Lung Screening Trial. Acad Radiol 2003;10:713–715.

14. Hillman BJ. Economic, legal, and ethical rationales for the ACRIN national lung screening trial of CT screening for lung cancer. Acad Radiol 2003;10:349–350.

15. Moore SM, Gierada DS, Clark KW, Blaine GJ. Image quality assurance in the prostate, lung, colorectal, and ovarian cancer screening trial network of the National Lung Screening Trial. J Digit Imaging 2005;18:242–250.

16. Kundel HL, Polansky M. Measurement of observer agreement. Radiology 2003;228:303–308.

17. Crewson PE. Reader agreement studies. AJR Am J Roentgenol 2005;184:1391–1397.

18. Hansell DM, Wells AU. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Radiol 2003;58:573–574.

19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.

20. Erasmus JJ, Gladish GW, Broemeling L, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. J Clin Oncol 2003;21:2574–2582.

21. Revel MP, Bissery A, Bienvenu M, Aycard L, Lefort C, Frija G. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 2004;231:453–458.

22. Bogot NR, Kazerooni EA, Kelly AM, Quint LE, Desjardins B, Nan B. Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. Acad Radiol 2005;12:948–956.

23. 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.

24. Diederich S, Wormanns D, Semik M, et al. Screening for early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers. Radiology 2002;222:773–781.

25. Leader JK, Warfel TE, Fuhrman CR, et al. Pulmonary nodule detection with low-dose CT of the lung: agreement among radiologists. AJR Am J Roentgenol 2005;185:973–978.

26. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–1499.

27. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998;90:1801–1809.

28. Siegel CL, McFarland EG, Brink JA, Fisher AJ, Humphrey P, Heiken JP. CT of cystic renal masses: analysis of diagnostic performance and interobserver variation. AJR Am J Roentgenol 1997;169:813–818.

29. Aziz ZA, Wells AU, Hansell DM, et al. HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation. Thorax 2004;59:506–511.

30. Garg K, Kemp JL, Wojcik D, et al. Thromboembolic disease: comparison of combined CT pulmonary angiography and venography with bilateral leg sonography in 70 patients. AJR Am J Roentgenol 2000;175:997–1001.

31. Cham MD, Yankelevitz DF, Shaham D, et al. Deep venous thrombosis: detection by using indirect CT venography. The Pulmonary Angiography-Indirect CT Venography Cooperative Group. Radiology 2000;216:744–751.

32. Patel S, Kazerooni EA. Helical CT for the evaluation of acute pulmonary embolism. AJR Am J Roentgenol 2005;185:135–149.

33. Aberle DR, Gamsu G, Henschke CI, Naidich DP, Swensen SJ. A consensus statement of the Society of Thoracic Radiology: screening for lung cancer with helical computed tomography. J Thorac Imaging 2001;16:65–68.

34. MacMahon H, Austin JH, Gamsu G, et al. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. Radiology 2005;237:395–400.

35. Henschke CI, Yankelevitz DF, Naidich DP, et al. CT screening for lung cancer: suspiciousness of nodules according to size on baseline scans. Radiology 2004;231:164–168.

36. Shah SK, McNitt-Gray MF, De Zoysa KR, et al. Solitary pulmonary nodule diagnosis on CT: results of an observer study. Acad Radiol 2005;12:496–501.

37. Brown MS, Goldin JG, Rogers S, et al. Computer-aided lung nodule detection in CT: results of large-scale observer test. Acad Radiol 2005;12:681–686.

38. Kostis WJ, Yankelevitz DF, Reeves AP, Fluture SC, Henschke CI. Small pulmonary nodules: reproducibility of three-dimensional volumetric measurement and estimation of time to follow-up CT. Radiology 2004;231:446–452.

39. 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.

40. Takashima S, Sone S, Li F, et al. Small solitary pulmonary nodules (< or =1 cm) detected at population-based CT screening for lung cancer: reliable high-resolution CT features of benign lesions. AJR Am J Roentgenol 2003;180:955–964.

41. Shah SK, McNitt-Gray MF, Rogers SR, et al. Computer-aided diagnosis of the solitary pulmonary nodule. Acad Radiol 2005;12:570–575.