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

**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 $\kappa$ statistic and by using response frequencies and descriptive statistics.

**Results:** Multirater $\kappa$ 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 $\kappa$ for follow-up recommendations was 0.35.

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

© RSNA, 2007

Clinical trial registration no. NCT00047385
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 interpretation of images obtained with thoracic imaging 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.

**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 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

**Advances in Knowledge**

- Interobserver agreement for classification of screening findings as measured with the $\kappa$ statistic ($\kappa = 0.58–0.64$) 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.

**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 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.

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-

<table>
<thead>
<tr>
<th>Screening Examination Categories Used in Case Selection</th>
<th>Additional Case Selection Features</th>
<th>Screening Result</th>
<th>No. of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noncalcified nodule with greatest transverse dimension 4 mm or larger</td>
<td>With or without noncalcified nodule smaller than 4 mm or benign calcification</td>
<td>Positive</td>
<td>75</td>
</tr>
<tr>
<td>Noncalcified nodule with greatest transverse dimension smaller than 4 mm</td>
<td>No noncalcified nodule 4 mm or larger or benign calcification</td>
<td>Negative</td>
<td>20</td>
</tr>
<tr>
<td>Benign calcification of any size</td>
<td>No noncalcified nodule</td>
<td>Negative</td>
<td>20</td>
</tr>
<tr>
<td>No nodule</td>
<td></td>
<td>Negative</td>
<td>20</td>
</tr>
</tbody>
</table>

* 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 presenta-
tion, which was produced by the Amer-
ican College of Radiology Imaging
Network–NLST group, defined and
showed examples of lesions having var-
ious features, such as soft-tissue attenu-
ation; ground-glass attenuation; smooth
margins; spiculated margins; and
pseudonodules such as linear bands of
atelectasis or scarring, bronchiolar in-
flammation, 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 com-
puter (Precision 340; Dell) and flat-
panel liquid crystal display color mon-
itor (SDM-P82; Sony) (13) 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 magnifica-
tion, electronic caliper measurement,
and adjustment of center and window
display settings, were used by all read-
ers.

Data Collection

Abnormalities were recorded by read-
ers 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 suspi-
cious for lung cancer, and to record the
section number and location (by lobe)
of all nodules, regardless of size or pres-
ence of calcification. The readers re-
corded the longest transverse dimen-
sion 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 or CT
scan to evaluate lesion growth was rec-
commended. Calculations were made by
using software (JMP, SAS Institute,
Cary, NC; Excel; Microsoft, Redmond,
Wash).

Results

Interobserver Agreement

Multirater χ values ranged from 0.58 to
0.64, depending on how the categories
were grouped. Agreement was highest
(χ = 0.64; 95% confidence interval:
0.62, 0.63) 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 (χ = 0.60; 95% confidence
interval: 0.57, 0.62) for the distinction
between any noncalcified nodule (non-
calcified nodule ≥ 4 mm or noncalcified
nodule < 4 mm) and no noncalcified
nodule (benign calcification or no nod-
ule) and for distinction among all four
categories (χ = 0.58; 95% confidence
interval: 0.53, 0.61). The χ values
among all reader pairs for all cases
ranged from 0.40 to 0.82 for the distinc-
tion 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 \(\pm\) 22) and some identifying several times as many noncalcified nodules smaller than 4 mm as others (mean, 41 nodules \(\pm\) 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 \(\kappa\) 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 \(\kappa\) 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
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 (e.g., 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
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 Herbrick, 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 Clingen, 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


