Published in final edited form as:
Acad Radiol. 2009 January ; 16(1): 28–38. doi:10.1016/j.acra.2008.05.022.

Assessment of Radiologist Performance in the Detection of Lung Nodules: Dependence on the Definition of "Truth"

Samuel G. Armato III, Ph.D.[1], **Rachael Y. Roberts, M.D.**[1], **Masha Kocherginsky, Ph.D.**[1], **Denise R. Aberle, M.D.**[3], **Ella A. Kazerooni, M.D., M.S.**[4], **Heber MacMahon, M.D.**[1], **Edwin J.R. van Beek, M.D., Ph.D.**[2], **David Yankelevitz, M.D.**[5], **Geoffrey McLennan, M.D., Ph.D.**[2], **Michael F. McNitt-Gray, Ph.D.**[3], **Charles R. Meyer, Ph.D.**[4], **Anthony P. Reeves, Ph.D.**[5], **Philip Caligiuri, M.D.**[1], **Leslie E. Quint, M.D.**[4], **Baskaran Sundaram, M.D.**[4], **Barbara Y. Croft, Ph.D.**[6], and **Laurence P. Clarke, Ph.D.**[6]

1*The University of Chicago*

2*University of Iowa*

3*University of California, Los Angeles*

4*University of Michigan*

5*Cornell University*

6*National Cancer Institute*

Abstract

Rationale and Objectives—Studies that evaluate the lung-nodule-detection performance of radiologists or computerized methods depend on an initial inventory of the nodules within the thoracic images (the "truth"). The purpose of this study was to analyze (1) variability in the "truth" defined by different combinations of experienced thoracic radiologists and (2) variability in the performance of other experienced thoracic radiologists based on these definitions of "truth" in the context of lung nodule detection on computed tomography (CT) scans.

Materials and Methods—Twenty-five thoracic CT scans were reviewed by four thoracic radiologists, who independently marked lesions they considered to be nodules ≥ 3 mm in maximum diameter. Panel "truth" sets of nodules then were derived from the nodules marked by different combinations of two and three of these four radiologists. The nodule-detection performance of the other radiologists was evaluated based on these panel "truth" sets.

Results—The number of "true" nodules in the different panel "truth" sets ranged from 15–89 (mean: 49.8±25.6). The mean radiologist nodule-detection sensitivities across radiologists and panel "truth" sets for different panel "truth" conditions ranged from 51.0–83.2%; mean false-positive rates ranged from 0.33–1.39 per case.

Conclusion—Substantial variability exists across radiologists in the task of lung nodule identification in CT scans. The definition of "truth" on which lung nodule detection studies are based

Corresponding Author: Samuel G. Armato III, Ph.D., Dept. of Radiology, MC 2026, The University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637, 773-834-3044, 773-702-0371 (fax), s-armato@uchicago.edu.

must be carefully considered, since even experienced thoracic radiologists may not perform well when measured against the "truth" established by other experienced thoracic radiologists.

Keywords

lung nodule; computed tomography (CT); thoracic imaging; inter-observer variability; computer-aided diagnosis (CAD)

INTRODUCTION

Studies that evaluate the lung-nodule detection performance of computer-aided diagnostic (CAD) methods or of different groups of radiologists fundamentally depend on an initial inventory of the nodules within the images. This assessment of "truth" is usually provided by a panel of experienced thoracic radiologists who review the images used in the study to identify lesions that are defined as targets of the study (1–3). Change the "truth," however, and the performance of the CAD method or radiologist under evaluation necessarily changes (4,5). The "truth" for a specific study is affected by a number of factors, including the composition of the expert panel (6), the defined targets of the study, the instructions provided to panel members, and the manner in which individual panel members interpret the defined study targets and instructions.

Lung nodules as a study target are especially subjective. The term "nodule" refers to abnormalities that span a wide spectrum, which is itself a subset of a broader spectrum of lesions that can be described as "focal abnormalities" (7). Varying interpretations of these spectra by different radiologists lead to variability in radiologists' identification of lung nodules (8). Compound variability in the definition of "nodule" with subjective qualifying attributes, such as minimum size, radiographic solidity, or actionability, and the potential for discordant interpretation is further magnified. The determination that a nodule is present at a specific location is almost always based on image features alone as interpreted by a radiologist, without independent objective verification, given the inherent limitations of obtaining lung tissue or post mortem data in humans. According to Dodd et al. (9), "dependence on expert opinion derived from the very same images used for the assessment of the imaging system or algorithm leads to an additional source of uncertainty that is not present when an independent source of 'ground truth' is available." These investigators suggest that some form of resampling of the expert panel may be useful to understand this additional uncertainty (9).

To create a publicly available database of annotated thoracic computed tomography (CT) scans as a reference standard for the medical imaging research community, the Lung Image Database Consortium (LIDC) developed a two-phase process for the interpretation of CT scans by such an expert panel. Specifically, a panel of four experienced thoracic radiologists, one from each of four different institutions, reviews the CT scans under two separate and distinct conditions (10). According to the LIDC process, the initial "blinded read phase" requires radiologists to independently mark nodules and other lesions they identify in a thoracic CT scan using a computer interface. During the subsequent "unblinded read phase," the blinded read results of all radiologists are revealed to each of the radiologists, who then independently review their marks along with the anonymous marks of their colleagues; each radiologist may choose to alter or delete any of their own marks during the unblinded read or they may place additional marks in the CT scan. This two-phase approach was developed to identify as completely as possible all lesions interpreted as nodules in a CT scan without requiring forced consensus. The blinded and unblinded read phases are intended to comprise a single, comprehensive process for establishing a robust "truth" for lung nodules in CT scans. The "truth" created through this process will be associated with the LIDC database for all who use it to train and

test CAD methodologies or to conduct any other studies that evaluate nodule-detection performance.

The LIDC two-phase process, however, deliberately deviates from the more commonly used approaches to establish "truth." Most studies incorporate a single-read panel approach, while others include a limited second round to arbitrate discordant findings. The findings of these expert panels may be combined and permuted (in a logical and scientifically sound manner) to obtain a series of non-unique "truth" sets against which the performance of the system under consideration may vary substantially (11–13).

The present study circumvents the more robust LIDC process and simulates the single-read panel paradigm by investigating the "truth" sets that may be constructed from the blinded reads for CT scans in the LIDC database. Then, rather than evaluate a CAD system, the performance of other LIDC radiologists is evaluated against these blinded-read-only "truth" sets.

This study poses several fundamental questions. When a group of two or three experienced thoracic radiologists forms a consensus panel to establish "truth" for a nodule-detection study, how would the results of the study differ if two or three other thoracic radiologists of equivalent experience were employed to establish the "truth?" How would the radiologists in the first panel fare against the "truth" established by the second panel? The purpose of this study was to analyze variability in the "truth" defined by different combinations of experienced thoracic radiologists from the expert panel and the variability in the performance of other experienced thoracic radiologists in the identification of lung nodules on CT scans based on these different definitions of "truth."

MATERIALS AND METHODS

Patient image data

A total of 25 thoracic helical CT scans were collected from a single LIDC site, in accordance with the inclusion criteria previously published (7,14). Appropriate local IRB approval was obtained for the research use of scans that had been acquired in accordance with established clinical or on-going research imaging protocols. Each CT scan had been acquired from a different patient (10 females, 15 males; age 40–75 years, median 59 years) on Aquilion (Toshiba) (n=17), Sensation 16 (Siemens) (n=5), or Emotion 6 (Siemens) (n=3) CT scanners. The tube peak potential energies used for scan acquisition were as follows: 120 kV (n=5), 130 kV (n=3), and 135 kV (n=17). Tube current ranged from 45–499 mA (mean: 228.9 mA). The slice thickness and reconstruction interval were equal for each scan at 2.0 mm (n=4) and 3.0 mm (n=21). The in-plane resolution of the 512×512-pixel sections ranged from 0.537–0.946 mm (mean: 0.682 mm). A "standard/non-enhancing" convolution kernel was used for image reconstruction. The majority of the CT scans (n=15) had been performed using intravenous contrast material.

Image evaluation

Monitors with clinically acceptable specifications were used at each site, and each monitor was calibrated with a VeriLUM Color Dual Mode Pod (IMAGE Smiths, Kensington, MD). Ambient lighting was set to simulate the clinical reading environment. Each CT scan was initially presented at a standard brightness/contrast setting without magnification, but the radiologists were allowed to adjust brightness, contrast, and magnification as appropriate to enable the most complete interpretation of the scan.

The scans identified at a single LIDC site were anonymized to remove all protected health information within the DICOM headers of the images in accordance with HIPAA guidelines (15) and electronically transferred to each of the four other LIDC sites to initiate the blinded

read process (10). One LIDC radiologist at each of the four sites independently evaluated each scan for the presence of lesions in three different categories: (1) nodules with greatest in-plane dimension ≥ 3 mm but < 30 mm, regardless of presumed histology ("nodule ≥ 3 mm"), (2) nodules < 3 mm that are not clearly benign (i.e., diffusely calcified) ("nodule < 3 mm"), and (3) other intraparenchymal lesions ≥ 3 mm ("non-nodule ≥ 3 mm") (e.g., scars, areas of consolidation, and lesions greater than 3 cm in diameter), which were noted for the sake of completeness (10). Through discussions and training, the radiologists were familiar with the subtleties of each lesion category prior to the study. The radiologists indicated the location of lesions through the placement of category-specific annotations on the images using an interactive computer interface. The spatial positions of lesions in each category as defined by each radiologist were recorded in an XML file for later analysis. This interface included measurement tools to assist the radiologists determine whether a lesion's dimension exceeded the 3-mm threshold. These cases subsequently proceeded to the unblinded read phase of the LIDC process. Since the purpose of this study was to simulate a single-read panel approach to establishing "truth," the final post-unblinded read results were not included in this study.

An LIDC site may have more than a single "LIDC radiologist" to handle the workload generated by the LIDC database, which will eventually contain nearly one thousand CT scans with lung nodules. Each LIDC radiologist is a thoracic radiologist, and each was trained by the site's primary LIDC radiologist to become familiar with the details of the LIDC process. Accordingly, reads performed for the LIDC database are considered on an institutional basis, and, for the purpose of this study, "Radiologist A" will refer to the LIDC radiologist or radiologists at one specific LIDC site. All 25 scans were evaluated by a single LIDC radiologist at each of two sites (Radiologist C and Radiologist D), while at the other two sites, the scans were distributed between two LIDC radiologists (at one site, one radiologist evaluated 21 scans and the other evaluated four scans (Radiologist A); at the other site, one radiologist evaluated 14 scans and the other evaluated 11 scans (Radiologist B)). The radiologists were aware that they were reviewing the CT scans to provide an assessment of "truth" for lung nodule studies.

The image evaluation process (the blinded reads) effectively yielded four independent sets of nodule "truth" data. To determine the physical correspondence of annotations from different radiologists, all radiologist annotations were visually reviewed and inventoried by a single LIDC principal investigator. Using the computer interface and the XML files created during the blinded reads, the annotations of all four radiologists were displayed simultaneously at the appropriate spatial locations within the images. Through differences in color and shape, the displayed annotations identified the institution of the radiologist who placed the annotation and the lesion category indicated by that radiologist. Annotations considered to represent the same physical lesion within the scan were grouped together by visual inspection of all annotations followed by a subjective determination of the three-dimensional contiguity of the lesions those annotations were intended to represent. It should be noted that the same lesion could have been assigned to different lesion categories (i.e., "nodule < 3 mm," "nodule ≥ 3 mm," or " non-nodule ≥ 3 mm") by different radiologists or not annotated at all by a subset of radiologists. This grouping of annotations defined the inventory of lesions that provided the basis for all subsequent analyses; this comprehensive inventory process identified which lesions were annotated by which radiologists and the lesion category to which the lesion was assigned by each radiologist.

Evaluation of radiologist performance based on the "truth" of the other radiologists

With four independent and equally valid assessments of "truth," analysis was confined to the "nodule ≥ 3 mm" lesion category. The nodule-identification performance of each LIDC radiologist was evaluated in the context of a panel "truth" set derived from the "truth" sets of different combinations of two or three other LIDC radiologists. A true positive was recorded

for the radiologist being evaluated if that radiologist had annotated as a "nodule ≥ 3 mm" a lesion that was included in the panel "truth" set, a false positive was recorded if that radiologist had annotated as a "nodule ≥ 3 mm" a lesion that was not included in the panel "truth" set, and a false negative was recorded if that radiologist had annotated a lesion that was included in the panel "truth" set as one of the other two lesion categories (i.e., "nodule < 3 mm" or "non-nodule ≥ 3 mm") or did not annotate the lesion at all.

The performance of each of the four radiologists was evaluated against panel "truth" sets formed by the three possible pairwise combinations of the three other radiologists through (1) a logical OR (i.e., union) of the "truth" sets of the pair of radiologists and (2) a logical AND (i.e., intersection) of the "truth" sets of the pair of radiologists. The logical OR panel "truth" set included lesions annotated as a "nodule ≥ 3 mm" by at least one radiologist of the pair, while the logical AND panel "truth" set included lesions annotated as a "nodule ≥ 3 mm" by both radiologists of the pair.

The performance of each radiologist was evaluated against the panel "truth" sets that consisted of (1) a logical OR combination of the "truth" sets of the other three radiologists, (2) a majority combination of the other three radiologists' "truth" sets, and (3) a logical AND combination of the "truth" sets of other the three radiologists. The logical OR panel "truth" set included lesions annotated as a "nodule ≥ 3 mm" by at least one of the three radiologists, the majority panel "truth" set included lesions annotated as a "nodule ≥ 3 mm" by at least two of the three radiologists, and the logical AND panel "truth" set included lesions annotated as a "nodule ≥ 3 mm" by all three radiologists.

RESULTS

Number of nodules

A total of 91 lesions were identified as "nodules ≥ 3 mm" by at least one of the four radiologists. The number of nodules identified by each of the four radiologists is shown in Table 1. Radiologist C defined the fewest lesions as nodules (n=20), and Radiologist A defined the most lesions as nodules (n=63). For the nodules that were identified by each radiologist, Figure 1 presents the numbers of those nodules that were identified by that radiologist alone, by the radiologist and one other radiologist, by the radiologist and two other radiologists, and by the radiologist and all three other radiologists. The complexities of the varied combinations of radiologists that identified each of the 91 nodules can be appreciated from the Venn diagram in Figure 2.

Variability in "truth" sets

Twenty-four panel "truth" sets were created in total: the logical OR and the logical AND sets for the six possible pairwise combinations of the four radiologists (n=12) (Figure 3) and the logical OR, the majority, and the logical AND sets for the four possible combinations of three of the four radiologists (n=12) (Figure 4). The number of "true" nodules in these "truth" sets spanned a wide range (Table 2), with the smallest number of nodules included in the panel "truth" set derived from the logical AND of Radiologists B, C, and D (n=15) and the largest number of nodules included in the panel "truth" set derived from the logical OR of Radiologists A, B, and D (n=89). The mean number of "true" nodules across all panel "truth" sets was 49.8 ±25.6.

Figure 5(a) shows a lesion that was identified as a "nodule ≥ 3 mm" by one radiologist but not by another radiologist, so that this lesion was considered a "true" nodule for the logical OR combination of these two specific radiologists but not for their logical AND combination. Figure 5(b) shows a lesion that was identified as a "nodule ≥ 3 mm" by both of these

radiologists; this lesion was considered a "true" nodule for both the logical OR and the logical AND combinations of the two radiologists. Figure 5(c) shows a lesion that was identified as a "nodule ≥ 3 mm" by one radiologist but not by two others, so that this lesion was considered a "true" nodule for the logical OR combination of these three specific radiologists but not for their majority or logical AND combinations. Figure 5(d) shows a lesion that was identified as a "nodule ≥ 3 mm" by two of these radiologists but not by the third; this lesion was considered a "true" nodule for both the logical OR and the majority combinations of the three radiologists but not for their logical AND combination. Figure 5(e) shows a lesion that was identified as a "nodule ≥ 3 mm" by all three of these radiologists; this lesion was considered a "true" nodule for the logical OR, the majority, and the logical AND combinations of the three radiologists.

Table 2 also presents the mean number of nodules across all radiologist combinations for each panel "truth" condition. The trends observed in these means are consistent with expectation. First, the mean number of nodules obtained for all logical OR combinations of individual radiologist "truth" sets exceeds the mean number of nodules obtained for all logical AND combinations of individual radiologist "truth" sets since AND (the intersection of the sets of nodules identified by each radiologist) is more restrictive than OR (the union of these sets). Second, the mean number of nodules obtained for all majority combinations of individual radiologist "truth" sets (for the combinations of three radiologists) is between the mean number of nodules obtained for all logical OR combinations and the mean number of nodules obtained for all logical AND combinations since the majority is more strict than an OR but less strict than an AND. Third, the mean number of nodules obtained for all logical OR combinations of three radiologists exceeds the mean number of nodules obtained for all logical OR combinations of two radiologists since an OR among three is more inclusive than an OR between two. Fourth, the mean number of nodules obtained for all logical AND combinations of three radiologists is less than the mean number of nodules obtained for all logical AND combinations of two radiologists since an AND among three is less inclusive than an AND between two. Fifth, the mean number of nodules obtained for all majority combinations of individual radiologists (for the combinations of three radiologists) exceeds the mean number of nodules obtained for all logical AND combinations of two radiologists since a majority among three is more inclusive than an AND between two specific radiologists.

Radiologist performance

When the "nodules ≥ 3 mm" identified by the individual radiologists were compared against the different panel "truth" sets (for panel combinations that did not include that specific radiologist), a wide range of nodule-detection sensitivities and false-positive rates resulted (Table 3). The mean sensitivities ranged from 51.0% for radiologist performance compared against the logical OR combination of three radiologists to 83.2% for radiologist performance compared against the logical AND combination of three radiologists. The mean false-positive rates ranged from 0.33 false positives per case for radiologist performance compared against the logical OR combination of three radiologists to 1.39 false positives per case for radiologist performance compared against the logical AND combination of three radiologists. Both the average sensitivities and the average false-positive rates increased as the panel "truth" set became more restrictive (Figure 6).

The nodule-detection sensitivities of individual radiologists are shown in Table 4 for the pairwise "truth" sets and in Table 5 for the triplet "truth" sets. Consistent with the trends observed for the aggregate sensitivities in Table 3, the sensitivities of individual radiologists increased as the panel "truth" sets became more restrictive. Each radiologist tended to be fairly consistent in terms of sensitivity across "truth" sets from different radiologist pairs for a given panel condition (i.e., logical OR or logical AND), with no coefficient of variation exceeding 0.10 (Table 4).

The combination of radiologist performance and variable panel "truth" sets may be appreciated by referring to Table 6, which presents the lesion categories assigned by the four radiologists to the five lesions shown in Figures 5(a)–(e). The "nodule ≥ 3 mm" category, which is the only category of interest for this study, is shown in bold. Each lesion could have three other possible category assignments: "nodule < 3 mm," "non-nodule ≥ 3 mm," or no category at all. Of immediate note is that none of these five lesions was assigned to the same category by all four radiologists. Furthermore, the inclusion of any of the five lesions in a specific panel "truth" set depends on how that panel is constructed (pair vs. triplet, the specific radiologists included, and the combination rule (OR, majority, or AND)). The lesion in Figure 5(a) was selected to demonstrate a lesion identified as a "nodule ≥ 3 mm" by one radiologist but not by another radiologist (i.e., a "true" nodule for the logical OR combination of two radiologists). From Table 6, the previous statement holds for a logical OR combination of Radiologists B and C or Radiologists A and C or Radiologists C and D; however, the lesion in Figure 5(a) would not be included in the logical OR panel "truth" set of Radiologists A and B or Radiologists A and D or Radiologists B and D. Considering the logical OR panel "truth" set of Radiologists B and C for Figure 5(a), it is interesting to note from Table 6 that this lesion would be recorded as a false negative for both Radiologist A and Radiologist D, but for different reasons: Radiologist A identified the lesion but considered it to be less than 3 mm in diameter and hence not a study target, while Radiologist D either did not observe the lesion at all or considered the structure to be beyond the scope of the three defined lesion categories (e.g., normal anatomy). Similar observations regarding differences in interpretation or "missed" lesions may be made from Table 6 for the other four lesions in Figures 5(b)–(d).

DISCUSSION

Several limitations are inherent in this study. First, the task of identifying nodules in the context of establishing "truth" for research studies differs from the identification task in the clinical setting, and the radiologists were asked to identify lesions without the benefit of accompanying clinical data. Second, pathologic information was not available for any of the lesions. Third, to define the study targets, radiologists were forced to make binary decisions as to the presence of appropriately sized nodules on the CT scans. Interestingly, these limitations are shared with many published nodule-detection studies.

The design of our study presents the potential for an interesting bias. Although radiologists were instructed to review the CT scans from the perspective of identifying "truth," these same findings were used to evaluate the "performance" of each radiologist against the findings (i.e., the "truth") of the other radiologists. The alternative study design would have included an initial session in which radiologists would be instructed to evaluate the scans for "truth," and then, after sufficient time had elapsed, a second session would have been conducted in which the radiologists would be instructed to review the same scans for the presence of lung nodules in a more routine manner and without the added burden of establishing "truth." Either scenario, however, differs from the reality of clinical practice, a fact that underlies any observer study conducted in a research setting; the psychology of the radiologist is necessarily altered. It is difficult to know whether the radiologists might have interpreted the scans differently with the knowledge that they were establishing "truth" rather than the thought that their findings were to be compared against an already existing "truth." It could be argued that the process of establishing "truth" would cause the radiologists to be more vigilant, especially in the absence of clinical information that could mitigate the need to report a specific lesion as a potential nodule. Conversely, under the alternative study design, the radiologists could demonstrate greater attention to the task with the knowledge that their performance would be compared against some reference. Despite these potential differences, the approach adopted in the present study reflects a consistent psychology that existed for both the "truthing" task and the performance evaluation task.

A total of 91 lesions were identified as "nodules ≥ 3 mm" by at least one of the four radiologists. This finding does not imply that the 25 CT scans contained only 91 nodules; had a fifth radiologist been involved, additional lesions might have been defined as "nodules ≥ 3 mm." Such a postulation further supports the conclusions that may be drawn from this study regarding the variability of "truth" assessments.

The assignment of a lesion to a specific category in the context of the panel "truth" set or the evaluated radiologist required three subjective steps: (1) identification of a lesion (Is the observed structure an abnormality or normal anatomy?), (2) determination of lesion size (Is the longest dimension of the lesion ≥ 3 mm but < 30 mm?), and (3) evaluation of lesion features (Does the lesion represent a "nodule"?). The multiple levels of inherently subjective interpretation required on the part of the radiologists help explain the observed variability in this study, and such variability, based on equally subjective aspects of image interpretation, is certainly present in clinical practice. A lesion included as a "nodule ≥ 3 mm" in any particular panel "truth" set but not identified as such by the evaluated radiologist could represent, in the context of that "truth" set, a search error or a decision-making error according to the categories of Kundel, *et al.* (16), although a full accounting of this distinction may not be extracted from the data collected.

A 3-mm size threshold separated the "nodules ≥ 3 mm" that were of interest in this study from the "nodules < 3 mm" that were not of interest. The measurement of lesion size, even with the use of electronic measurement tools, is a highly variable task both in clinical practice and in observer evaluation studies.(17–19) From the perspective of establishing "truth," a lesion marked as a "nodule ≥3 mm" in the "truth" set became a target that must be identified by the "system" (i.e., a CAD method or, in this study, the radiologist being evaluated), while a lesion marked as a "nodule < 3 mm" was not to be identified by the system. From the perspective of performance assessment, a lesion marked by the evaluated radiologist as a "nodule ≥ 3 mm" was considered a true positive if the lesion was marked as a "nodule ≥ 3 mm" in the "truth" set, a lesion marked by the radiologist as a "nodule < 3 mm" (or as a "non-nodule ≥ 3 mm" or not marked at all) was considered a false negative if the lesion was marked as a "nodule ≥ 3 mm" in the "truth" set, and a lesion marked by the radiologist as a "nodule ≥ 3 mm" was considered a false positive if the lesion was not marked as a "nodule ≥ 3 mm" by the requisite number of truth panel radiologists regardless of whether the other truth panel radiologists marked the lesion as a "nodule < 3 mm" or as a "non-nodule ≥ 3 mm" or whether the other truth panel radiologists provided no mark at all. A greater degree of variability would be expected in the "truth" sets for nodules with diameter near 3 mm and, accordingly, in the radiologists' nodule-detection performance for these nodules.

The imposition of such a size threshold is consistent with the design of most reported CAD system evaluation studies. CAD systems are typically developed to identify nodules above some minimum size (as determined by the investigators), and at some point in the algorithm the system must determine whether each nodule candidate satisfies that size threshold. Similarly, when establishing "truth," three binary decisions must be made: (1) whether a lesion is present at a specific location, (2) whether the lesion is a "nodule," and (3) whether that nodule satisfies the size threshold. These are the same decisions that were required of the radiologists in this study; indeed, analogous decisions are required of radiologists in clinical practice on a daily basis. Accordingly, a "false negative" or "false positive" could result when the opinion of the evaluated radiologist differed from that of the "truth" panel with regard to the size or the "nodularity" of an abnormality that was recognized by both the evaluated radiologist and the panel.

While the number of nodules identified by different radiologists is similar (especially for Radiologists A, B, and D) (see Figure 1), the specific nodules identified by the individual

radiologists are quite distinct. For example, while Radiologists A and B identified 63 and 62 nodules, respectively, the number of nodules contained within the logical OR combination of Radiologists A and B (n=84) is more than twice the number of nodules contained within their logical AND combination (n=41). This finding means that only about two-thirds of the nodules identified by either Radiologist A or Radiologist B also were identified by the other.

Radiologist nodule-detection sensitivity increased as the panel "truth" set criterion became more strict (e.g., from OR to AND) for two main reasons. First, the number of "actual nodules" based on the more restrictive "truth" set decreased, and so the denominator of the expression for sensitivity decreased. Second, the nodules contained within the more strict "truth" set likely represent more obvious nodules that present a greater likelihood of radiologist agreement.

Similarly, the number of false positives increased as the panel "truth" set criterion became more strict. Since the more strict "truth" set contained fewer "actual nodules," fewer of the lesions marked by the radiologist under evaluation were considered "true" nodules. The remaining marked lesions, therefore, were considered false positives. Note that since this study was one of detection rather than classification, the concept of a "true negative" did not exist, and the estimation of several quantities frequently used in rater agreement studies (e.g., specificity and the kappa statistic) was not possible.

The findings presented here challenge the certitude inherently associated with the expert-observer-defined "truth" that provides the basis for many medical image analysis studies across a diversity of imaging modalities. Similar considerations exist for nodule segmentation studies, which depend on variable definitions of "truth" based on the nodule outlines of different radiologists (17,20). For many tasks, radiologist interpretation is the closest approximation to "truth" that may be attained; the limitations of that approximation, however, must be recognized and appreciated by investigators. The two-phase approach to the definition of "truth" for nodule-detection studies developed by the LIDC was intended to reduce the variability inherent among radiologists (8). The results of the present study could have important implications for the clinical interpretation of CT scans in the context of lung nodule detection in which a single reader is responsible for what is clearly a difficult detection (and subsequent classification) task—even double and triple readings have limitations and variability that must be understood and should be taken into account when comparing the performance of CAD systems against radiologist "truth."

Acknowledgements

Supported in part by USPHS Grants U01CA091085, U01CA091090, U01CA091099, U01CA091100, and U01CA091103.

REFERENCES

1. Wormanns D, Ludwig K, Beyer F, Heindel W, Diederich S. Detection of pulmonary nodules at multirow-detector CT: Effectiveness of double reading to improve sensitivity at standard-dose and low-dose chest CT. European Radiology 2005;15:14–22. [PubMed: 15526207]
2. Leader JK, Warfel TE, Fuhrman CR, et al. Pulmonary nodule detection with low-dose CT of the lung: Agreement among radiologists. American Journal of Roentgenology 2005;185:973–978. [PubMed: 16177418]
3. Novak, CL.; Qian, J.; Fan, L., et al. Inter-observer variations on interpretation of multi-slice CT lung cancer screening studies, and the implications for computer-aided diagnosis; SPIE Proceedings; 2002. p. 68-79.
4. Ochs, R.; Kim, HJ.; Angel, E.; Panknin, C.; McNitt-Gray, M.; Brown, M. Forming a reference standard from LIDC data: Impact of reader agreement on reported CAD performance; SPIE Proceedings; 2007.

5. Paquerault S, Petrick N, Myers KJ, Samuelson FW. Impact of a computer-aided detection (CAD) system on reader performance: Assessment based on a truthing panel compared to the true gold standard. Radiology 2007;245(P):546–547.
6. Petrick, N.; Gallas, BD.; Samuelson, FW.; Wagner, RF.; Myers, KJ. Influence of panel size and expert skill on truth panel performance when combining expert ratings; SPIE Proceedings; 2005. p. 49-57.
7. Armato SG III, McLennan G, McNitt-Gray MF, et al. Lung Image Database Consortium: Developing a resource for the medical imaging research community. Radiology 2004;232:739–748. [PubMed: 15333795]
8. Armato SG III, McNitt-Gray MF, Reeves AP, et al. The Lung Image Database Consortium (LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans. Academic Radiology 2007;14:1409–1421. [PubMed: 17964464]
9. Dodd LE, Wagner RF, Armato SG III, et al. Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: Contemporary research topics relevant to the Lung Image Database Consortium. Academic Radiology 2004;11:462–475. [PubMed: 15109018]
10. McNitt-Gray MF, Armato SG III, Meyer CR, et al. The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation. Academic Radiology 2007;14:1464–1474. [PubMed: 18035276]
11. Miller, DP.; O'Shaughnessy, KF.; Wood, SA.; Castellino, RA. Gold standards and expert panels: A pulmonary nodule case study with challenges and solutions; SPIE Proceedings; 2004. p. 173-184.
12. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Investigative Radiology 1983;18:194–198. [PubMed: 6862810]
13. Jiang, Y. A Monte Carlo simulation method to understand expert-panel consensus truth and double readings. Medical Image Perception Conference XII; Medical Image Perception Conference XII 2007; The University of Iowa; Iowa City, IA.
14. Clarke LP, Croft BY, Staab E, Baker H, Sullivan DC. National Cancer Institute initiative: Lung image database resource for imaging research. Academic Radiology 2001;8:447–450. [PubMed: 11345275]
15. Department of Health and Human Services. Standards for privacy of individually identifiable health information: final rules. Federal Register 2002 2002;67:53182–53272.
16. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Investigative Radiology 1978;13:175–181. [PubMed: 711391]
17. Meyer CR, Johnson TD, McLennan G, et al. Evaluation of lung MDCT nodule annotation across radiologists and methods. Academic Radiology 2006;13:1254–1265. [PubMed: 16979075]
18. Reeves AP, Biancardi AM, Apanasovich TV, et al. The Lung Image Database Consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements. Academic Radiology 2007;14:1475–1485. [PubMed: 18035277]
19. Schwartz LH, Ginsberg MS, DeCorato D, et al. Evaluation of tumor measurements in oncology: Use of film-based and electronic techniques. Journal of Clinical Oncology 2000;18:2179–2184. [PubMed: 10811683]
20. Ross JC, Miller JV, Turner WD, Kelliher TP. An analysis of early studies released by the Lung Imaging Database Consortium (LIDC). Academic Radiology 2007;14:1382–1388. [PubMed: 17964461]

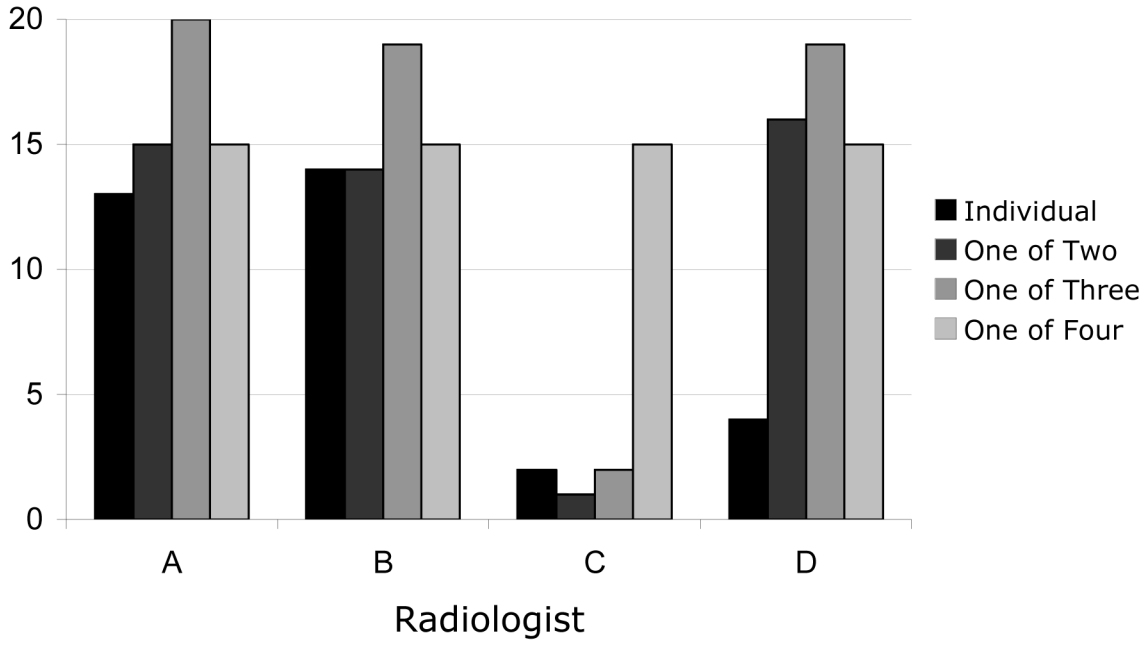


Figure 1.
The number of lesions that were identified as a "nodule ≥ 3 mm" by each radiologist individually, by each radiologist and one other radiologist, by each radiologist and two other radiologists, and by each radiologist and the three other radiologists (i.e., lesions that all four radiologists identified as nodules). The sum of the four bars for each radiologist corresponds to the data in Table 1.

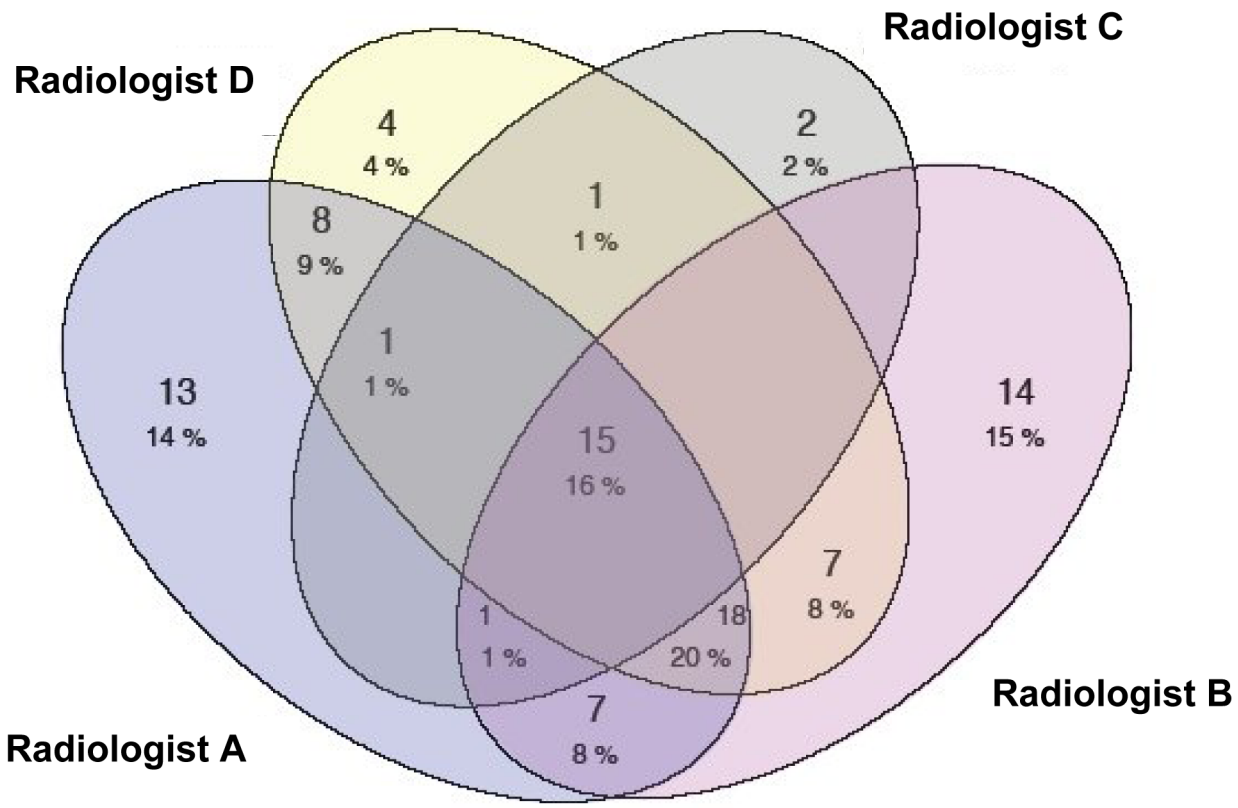


Figure 2.
Venn diagram of the different combinations of radiologists that identified the 91 lesions that were defined as "nodules ≥ 3 mm" by at least one radiologist.

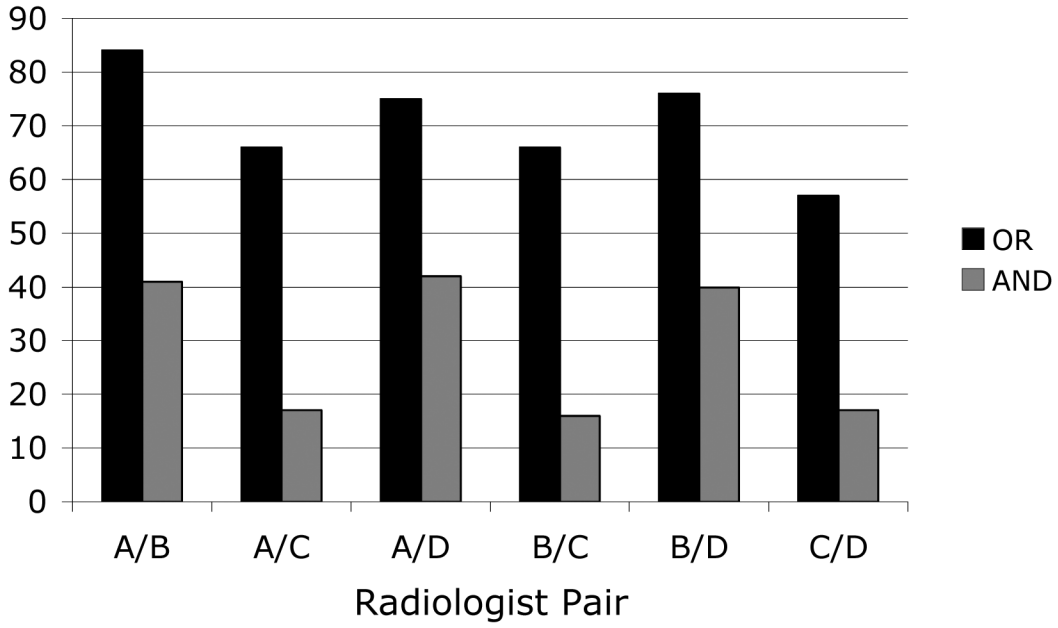


Figure 3.
The number of lesions identified as "nodule ≥ 3 mm" in the panel "truth" sets created from pairwise combinations of the four radiologists' individual reads combined through a logical OR operation and combined through a logical AND operation.

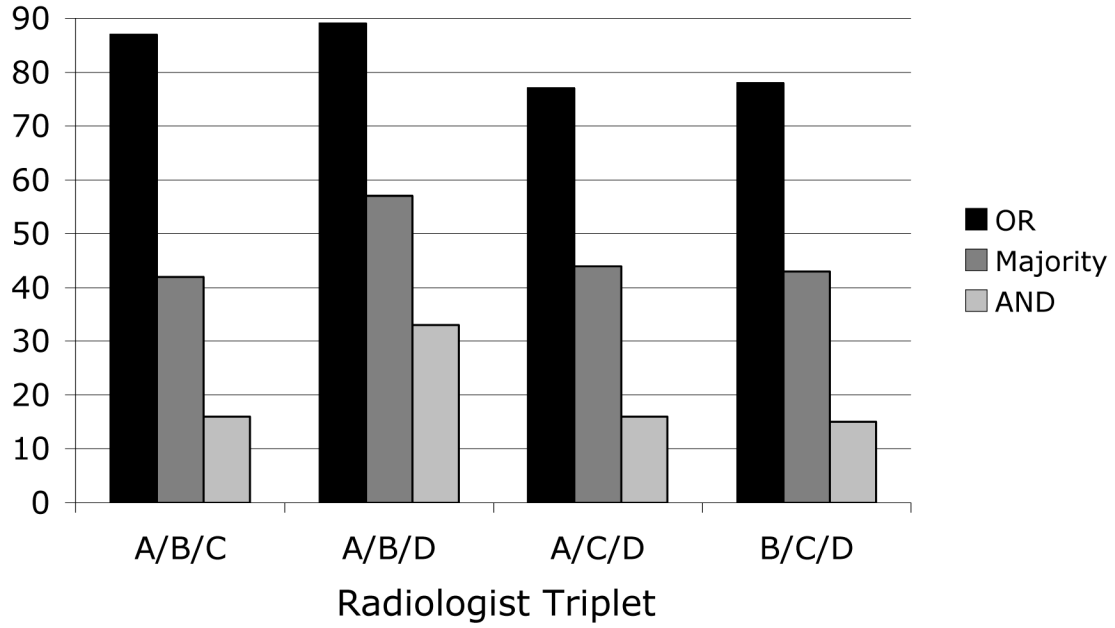


Figure 4.
The number of lesions identified as "nodule ≥ 3 mm" in the panel "truth" sets created from triplet combinations of the four radiologists' individual reads combined through a logical OR operation, a majority approach, and a logical AND operation.

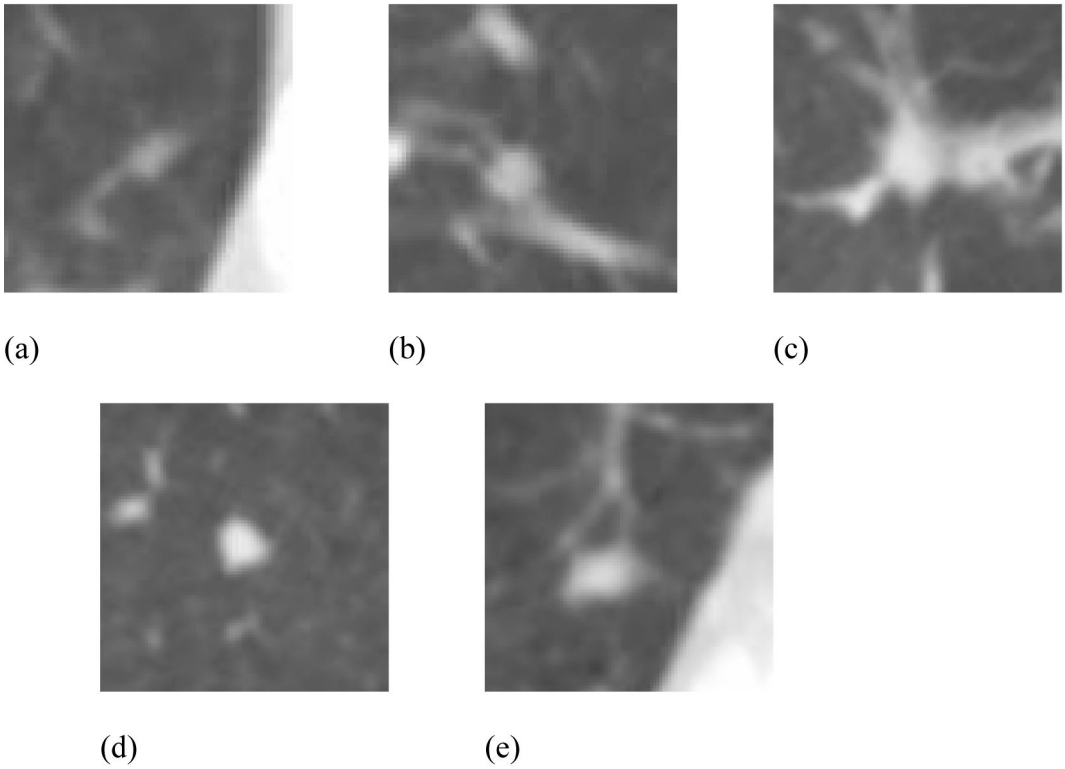
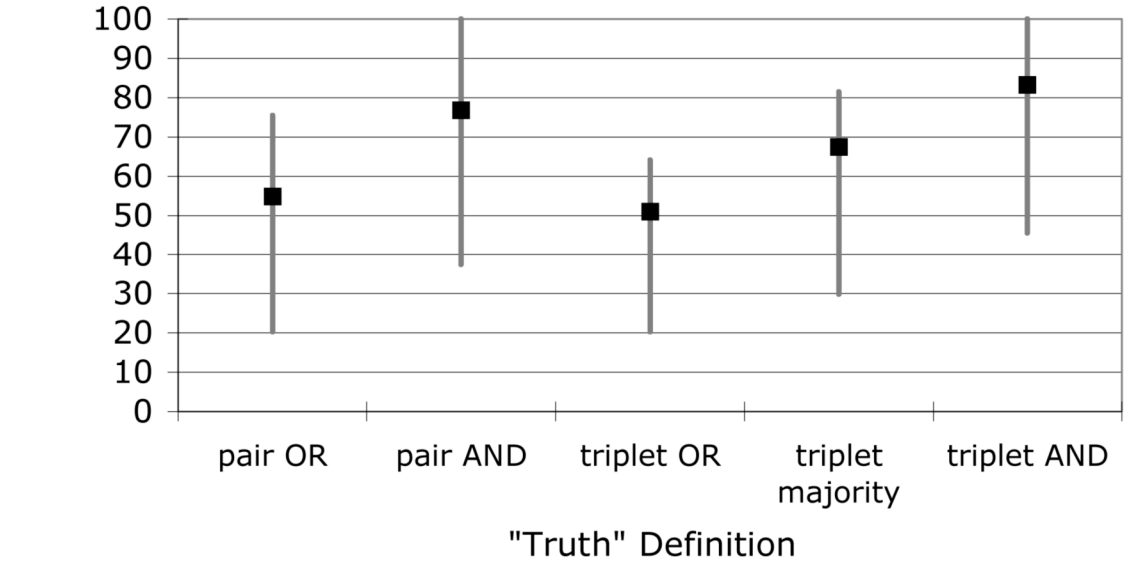
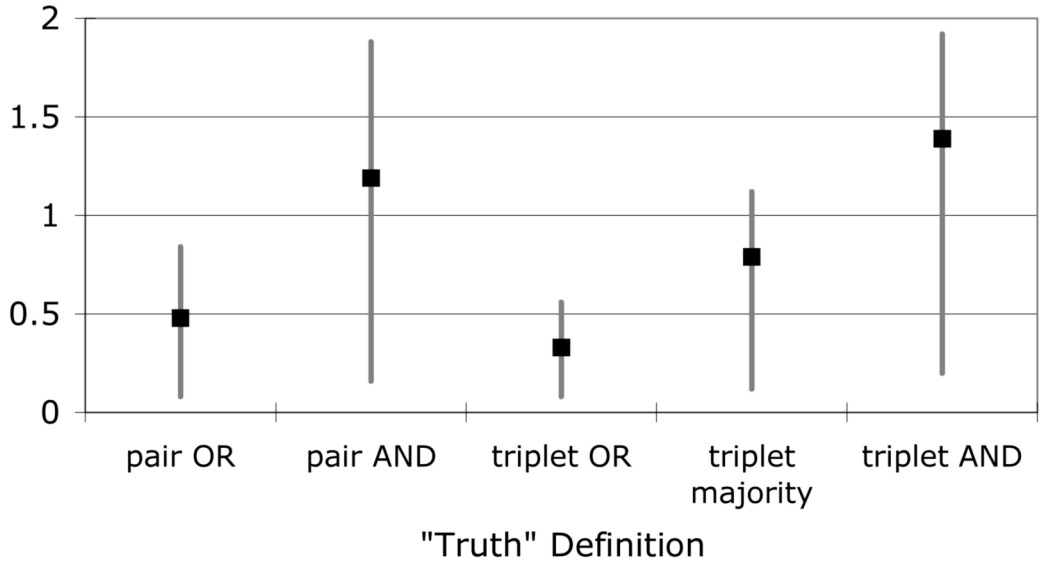


Figure 5.
Examples of lesions that were identified as (a) a "nodule ≥ 3 mm" by one radiologist but not by another (a "true" nodule for the logical OR combination of these two specific radiologists), (b) a "nodule ≥ 3 mm" by both of these radiologists (a "true" nodule for both the logical OR and the logical AND combinations), (c) a "nodule ≥ 3 mm" by one radiologist but not by two others (a "true" nodule for the logical OR combination of these three specific radiologists), (d) a "nodule ≥ 3 mm" by two of these radiologists but not by the third (a "true" nodule for both the logical OR and the majority combinations), (e) a "nodule ≥ 3 mm" by all three of these radiologists (a "true" nodule for the logical OR, the majority, and the logical AND combinations).



(a)



(b)

Figure 6.
The means and ranges (across radiologist combinations) of (a) radiologist nodule-detection sensitivities and (b) radiologist false-positive rates based on the different panel “truth” sets.

Table 1

The number of nodules identified by each radiologist.

	Number of nodules	Mean ± SD
Radiologist A	63	49.8±20.2
Radiologist B	62	
Radiologist C	20	
Radiologist D	54	

Table 2

The number of nodules contained in the panel "truth" sets obtained from different combinations of radiologists under different conditions.

Panel "truth" set		Number of nodules	Mean ± SD
Radiologist pairs (OR / AND)	Radiologists A/B	84 / 41	70.7±9.5 / 28.8±13.4
	Radiologists A/C	66 / 17	
	Radiologists A/D	75 / 42	
	Radiologists B/C	66 / 16	
	Radiologists B/D	76 / 40	
	Radiologists C/D	57 / 17	
Radiologist triplets (OR / Majority / AND)	Radiologists A/B/C	87 / 42 / 16	82.8±6.1 / 46.5±7.0 / 20.0±8.7
	Radiologists A/B/D	89 / 57 / 33	
	Radiologists A/C/D	77 / 44 / 16	
	Radiologists B/C/D	78 / 43 / 15	

Table 3

Radiologist nodule-detection sensitivities and false-positive rates across radiologists and panel "truth" sets for each panel "truth" condition.

Panel "truth" condition	Sensitivity			False-positive rate (FPs/case)		
	Min	Max	(mean±SD)	Min	Max	(mean±SD)
Radiologist pairs (logical OR)	20.2%	75.4%	54.8±19.9%	0.08	0.84	0.48±0.31
Radiologist pairs (logical AND)	37.5%	100.0%	76.7±24.1%	0.16	1.88	1.19±0.70
Radiologist triplets (logical OR)	20.2%	64.1%	51.0±20.7%	0.08	0.56	0.33±0.25
Radiologist triplets (majority)	29.8%	81.4%	67.4±25.1%	0.12	1.12	0.79±0.47
Radiologist triplets (logical AND)	45.5%	100.0%	83.2±25.4%	0.20	1.92	1.39±0.81

Table 4

Individual radiologist nodule-detection sensitivities (in %) based on different pairwise panel "truth" sets combined through a logical OR / AND.

	Radiologist Pair							
	A/B	**A/C**	**A/D**	**B/C**	**B/D**	**C/D**	**Mean**	**CV**
Radiologist A	—	—	—	63.6 / 100.0	65.8 / 82.5	75.4 / 94.1	68.3 / 92.2	0.09 / 0.10
Radiologist B	—	62.1 / 94.1	64.0 / 78.6	—	—	71.9 / 88.2	71.9 / 88.2	0.08 / 0.09
Radiologist C	20.2 / 39.0	—	24.0 / 38.1	—	23.7 / 37.5	—	23.7 / 37.5	0.09 / 0.02
Radiologist D	58.3 / 80.5	65.2 / 94.1	—	63.6 / 93.8	—	—	63.6 / 93.8	0.06 / 0.09

CV: coefficient of variation

Table 5

Individual radiologist nodule-detection sensitivities (in %) based on different triplet panel "truth" sets combined through a logical OR / majority / AND.

	Radiologist Triplet			
	A/B/C	**A/B/D**	**A/C/D**	**B/C/D**
Radiologist A	—	—	—	64.1 / 81.4 / 100.0
Radiologist B	—	—	62.3 / 77.3 / 93.8	—
Radiologist C	—	20.2 / 29.8 / 45.5	—	—
Radiologist D	57.5 / 81.0 / 93.8	—	—	—

Table 6

The lesion categories assigned by the four radiologists to the five lesions shown in Figure 5. The "nodule ≥ 3 mm" category, which is the only category of interest for this study, is shown in bold.

	Radiologist A	Radiologist B	Radiologist C	Radiologist D
Figure 5(a)	Nodule < 3 mm	Nodule < 3 mm	**Nodule ≥ 3 mm**	
Figure 5(b)	**Nodule ≥ 3 mm**	**Nodule ≥ 3 mm**	**Nodule ≥ 3 mm**	
Figure 5(c)	**Nodule ≥ 3 mm**			Non-nodule ≥ 3 mm
Figure 5(d)	**Nodule ≥ 3 mm**	**Nodule ≥ 3 mm**		**Nodule ≥ 3 mm**
Figure 5(e)	**Nodule ≥ 3 mm**	**Nodule ≥ 3 mm**	**Nodule ≥ 3 mm**	Nodule < 3 mm