Published in final edited form as:

Eur Radiol. 2010 March ; 20(3): 549–557. doi:10.1007/s00330-009-1596-y.

Computer-aided detection (CAD) of lung nodules in CT scans: radiologist performance and reading time with incremental CAD assistance

Justus E. Roos✉,

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

David Paik,

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

David Olsen,

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

Emily G. Liu,

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

Lawrence C. Chow,

Department of Radiology, Oregon, Health and Science University, MCL340,3181 SW Sam Jackson Park Rd, Portland, OR 97201, USA

Ann N. Leung,

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

Robert Mindelzun,

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

Kingshuk R. Choudhury,

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

David P. Naidich,

Department of Radiology, New York University Medical Center, 550 First Avenue, New York, NY 10016, USA

Sandy Napel, and

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

Geoffrey D. Rubin

Department of Radiology, Stanford University Medical Center, 300 Pasteur Drive, Room S-072, Stanford, CA 94305-5105, USA

Abstract

Objective—The diagnostic performance of radiologists using incremental CAD assistance for lung nodule detection on CT and their temporal variation in performance during CAD evaluation was assessed.

Methods—CAD was applied to 20 chest multidetector-row computed tomography (MDCT) scans containing 190 non-calcified 3-mm nodules. After free search, three radiologists independently evaluated a maximum of up to 50 CAD detections/patient. Multiple free-response ROC curves were generated for free search and successive CAD evaluation, by incrementally adding CAD detections one at a time to the radiologists' performance.

Results—The sensitivity for free search was 53% (range, 44%–59%) at 1.15 false positives (FP)/patient and increased with CAD to 69% (range, 59–82%) at 1.45 FP/patient. CAD evaluation initially resulted in a sharp rise in sensitivity of 14% with a minimal increase in FP over a time period of 100 s, followed by flattening of the sensitivity increase to only 2%. This transition resulted from a greater prevalence of true positive (TP) versus FP detections at early CAD evaluation and not by a temporal change in readers' performance. The time spent for TP (9.5 ± 4.5 s) and false negative (FN) (8.4 ± 6.7 s) detections was similar; FP decisions took two- to three-times longer (14.4 ± 8.7 s) than true negative (TN) decisions (4.7 ± 1.3 s).

Conclusions—When CAD output is ordered by CAD score, an initial period of rapid performance improvement slows significantly over time because of non-uniformity in the distribution of TP CAD output and not to a changing reader performance over time.

Keywords

Multidetector-row computed tomography; MDCT; Computer-aided detection; CAD; Pulmonary nodules; Diagnostic performance

Introduction

Computer-aided detection (CAD) systems are capable of detecting lung nodules, including those missed by one or more radiologists, while at the same time creating false-positive (FP) results [1–11]. Das et al. [12] showed in a recent study comparing two commercially available CAD systems using multidetector-row computed tomography (MDCT) datasets, a mean sensitivity increase from individual readings of 76–84% (pooled data) using CAD as a second reader. The average number of FP findings per patient was reported as seven with both CAD systems.

To date, most reports use commercially available detectors with fixed performance thresholds (single 'black box' CAD output), but the successful discrimination of lesions from background tissue by a CAD algorithm is not an absolute determination. CAD algorithms typically compute a quantity or "CAD score" that corresponds to the probability

that a given locus within the imaging data is a lesion. When a lower CAD score is selected as the threshold for lesion inclusion in the CAD output, then sensitivity will increase at the expense of greater FP detections, moving to the right on the free-response receiver-operating characteristic (FROC) curve [13]. When CAD algorithms report detections as a set, a threshold has already been applied to the CAD score, which fixes the algorithm at a specific operating point on the FROC curve. Manufacturers select this threshold on commercial CAD detectors and thus provide a static output that assumes a preferred threshold for optimal CAD performance, regardless of reader characteristics, pre-test probability of significant lung lesions, and CT scan quality.

When investigators cannot vary the performance threshold of CAD, we cannot know if the detectors are optimised for maximal diagnostic performance and if a performance threshold optimised for one reader would be optimal for other readers. In more generalised terms, without the ability to systematically adjust performance parameters on commercial detectors, analyses of the complexity and confounding dependencies between CAD performance profiles and reader performance have been limited.

To our knowledge, the balance between sensitivity and specificity at varying CAD performance thresholds and the impact of CAD output on the reliability and interaction time of human users have not been described. Using an open CAD system that was developed in an interdisciplinary academic research setting [14] has allowed us to systematically alter performance thresholds and evaluate a variety of interactions between CAD output thresholds, reader-CAD interaction times and reader performance. Measurement of the absolute gain in sensitivity using CAD has been reported by others and thus was of secondary interest. The purpose of this work was (1) to assess the variability with which radiologists use CAD to improve nodule detection, (2) to determine how diagnostic performance changes as a function of the incremental use of CAD and (3) to evaluate temporal trends of radiologists' performance using CAD.

Materials and methods

Patients and image acquisition parameters

The patient database was derived from a previous study comparing the performance of radiologists and CAD that included chest CT of 20 outpatients (15 male and five female patients aged 15–91; mean 64), who were referred because of a clinical suspicion of pulmonary nodules [6]. Three patients had extrathoracic malignancy, one with known pulmonary metastases; the remainder had or were suspected of having at least one of the following: cardiac disease, COPD, inflammatory lung disease. Analysis of clinically acquired CT data was approved by our institutional review board and the requirement for informed consent was waived. MDCT (Siemens Medical Systems, Erlangen, Germany) was acquired in a single breath-hold without the use of intravenous contrast. Imaging parameters were as follows: 1-mm detector configuration, pitch 1.5–1.75, rotation time 0.5 s, 120 kVp, 200–300 mA, 512×512 image matrix. Transverse 1.25-mm thick sections were reconstructed at 0.6-mm intervals by using a high-resolution reconstruction kernel.

CAD

All CAD outputs derived from our CAD detector [14] were characterised by their location within the CT and by a CAD score. The CAD score is a numerical value that is derived for every voxel in the CT and approximates the likelihood of a detection being a true nodule. The loci with the 50 highest CAD scores were identified for each patient and sorted by descending CAD score. Thus, the CAD candidate with the highest likelihood of being a true pulmonary nodule was shown first to the reader during their CAD evaluation. This large number of CAD candidates per case provided a basis for the analysis of the incremental benefit and temporal trends through CAD evaluation and it was not intended to represent a clinically optimised operating point of CAD detector output.

For all subsequent analyses, a proximity criterion of a 1-cm radius from the centre of the standard of reference nodule was used to determine if a CAD hit was a true-positive (TP) or FP CAD detection.

Image interpretation

Three faculty radiologists, with 9, 14 and 24 years of CT experience, independently interpreted the 20 CT results. Readers 1 and 3 were specialists in body CT and reader 2 specialised in thoracic imaging. Interpretations were performed using a dedicated, non-commercially available lung nodule evaluation platform (LNEP) [6] that featured automated and simultaneous display of two orthogonal cross sections (transverse and coronal plane) in stacked cine mode together with a volume rendering of a cuboidal region of interest. Additional pan and zoom, window levelling (default window level was -750 HU and window width 1,500 HU), sliding thin-slab MIP display and cartwheel capabilities were available for image interpretation. Readers were free to use these tools at their own discretion. The LNEP was installed on a PC workstation (Dell E520, 2.1 GHz, Intel Core 2 Duo, 4 GB RAM) with a 20-inch LCD display (Dell, UltraSharp, 1,600×1,200 pixels). Ambient light conditions, which were similar to our clinical reading conditions, were kept constant for all reading sessions and readers.

CT images were presented to the readers for a two-phase interpretation. The readers were instructed first to identify all non-calcified pulmonary nodules with a diameter of 3-mm or more without time restriction or the aid of CAD (free search). The coordinates and a reader confidence rating from 1 to 5 were recorded by LNEP. The confidence rating was defined as: 5, definitely a nodule; 4, probably a nodule; 3, possibly a nodule; 2, unlikely to be a nodule; 1, very unlikely to be a nodule.

Following free search, the readers were presented with a list of up to 50 CAD detections per case. The CAD list was pre-filtered in order to remove candidates that were already identified by the readers during their free search. Confidence ratings were assigned to each CAD candidate with the additional possibility of assigning a zero value to CAD candidates that were definitely not considered to be a nodule.

Both interpretation phases were held in multiple readings sessions with blocks not exceeding one hour of successive readings in order to minimise any reader fatigue. The order with which the cases were presented to the readers was randomised. The readers were aware that

neither the range of numbers of nodules per case nor the fact that the CAD candidates were ordered according to a descending CAD score.

Expert-derived standard of reference

A consensus panel of two thoracic radiologists, who were not involved in the observer study, with 25 and 14 years of experience in reading thoracic CT established the standard of reference, following independent free search and a directed analysis of all nodule candidates identified by the three radiologists and CAD. The greatest dimension of each nodule was measured with digital callipers.

Evaluation

The interpretations of the three independent radiologists were compared with the reference standard after free search and after CAD analysis using three methods.

1. Global reader performance was assessed using mean sensitivities, calculated for each radiologist. A confidence level of 3 or more was used as a threshold for positivity.
2. A series of FROC curves were calculated following free search and following assessment of each of the 50 CAD detections. Sensitivity was plotted against the average FP detections per patient at five confidence intervals: 5, 4, 3, 2 and 1. The change in position of the five operating points on the FROC curve was recorded with each successive CAD candidate evaluation, and the change in the diagnostic performance of the readers relative to incremental use of CAD was analysed.

To assess the impact of the presentation order of CAD detections on successive reader performance, a time invariant interpretation model was created. It allowed us to assess the effect of time-dependent phenomena (e.g. tedium-induced fatigue) on performance by assuring a consistent response over the entire input range of CAD candidates. The model was created to respond to CAD candidates with a result that was weighted by the average confidence of three readers for both nodules and non-nodules detected by CAD. Two candidate orderings were input into the model: (1) the same distribution of nodule and non-nodule CAD candidates presented to the readers during their CAD evaluation, and (2) a uniform distribution of nodule and non-nodule CAD candidates.

3. To evaluate temporal trends of radiologists' performance using CAD, the duration of free search per case was recorded by the study coordinator using a stopwatch and the duration of each CAD evaluation was recorded automatically by the LNEP. To examine time-based influences on reader performance, sensitivity and specificity were measured based upon all preceding detections for each successive CAD candidate rank. To examine longer time-scale trends and to adjust for unevenly spaced time intervals between consecutive CAD evaluations, we smoothed both the rank and the temporal data using an averaging window of ± 3 data points.

Statistical analysis

A two-sided test for relative sensitivities, equivalent in large samples to McNemar's test [15], was performed at different FP rates per patient to test for significance between free search and free search plus CAD performance among the readers. Sensitivities at 0.5, 1.0 and 2.0 FP/patient were calculated using linear interpolation between FROC operating points. *P* values obtained by this test were adjusted upwards by a factor of 3 using Bonferroni correction to account for multiple comparisons at three different FP rates. Independence of readings across nodules was assumed, despite the fact that multiple nodules are sampled per case and independence between readings may be violated. However, assuming positive correlation for readings within cases, it can be shown that the *P* value of the naive test for relative sensitivity is an overestimation of the true significance [16]. Hence the resulting tests will reject the null hypothesis of no difference more conservatively if the readings are correlated.

Inter-reader agreement was quantified with kappa statistics [17]. Paired *t*-tests were used to evaluate the relative durations of the readers' interpretations. The chi-squared test was used to evaluate the pairwise comparison of readers' decisions on detected CAD candidates. For all statistical tests, a *P* value of less than 0.05 was considered to be significant.

Results

The reference standard contained 382 pulmonary nodules; 308 nodules were not calcified and 74 were calcified. One hundred and ninety non-calcified nodules were ≥ 3 mm. The median/mean number of nodules was 3.5/9.5 per CT (range 0–66 per CT) and the median/mean nodule diameter was 4.3 mm/5.2 mm (range 3–17.6 mm).

Diagnostic performance before and following CAD interpretation

CAD detected 74% (141 of 190) of the nodules. Eighteen percent (25 of 141) of these nodules were not found by any of the readers during free search. In contrast, 14% (27 of 190) of the nodules were detected by at least one reader on free search but not by CAD. Fourteen percent of all nodules were either undetected (11.4%) or rejected (2.6%) by all three readers.

Reader performance for detecting ≥ 3 -mm non-calcified nodules during free search and CAD evaluation is plotted in Fig. 1. Reader 1 achieved the highest sensitivity during free search. Following CAD interpretation, reader 2 achieved the highest sensitivity but with a substantially higher FP rate. Reader 3 had the lowest sensitivity during free search, but in spite of achieving greater performance improvement than reader 1 following CAD interpretation, he had the lowest performance overall. When interpolating a constant 0.5 FP rate per patient, the relative reader performance, both before and after CAD interpretation, was the same.

Relative reader sensitivities at different FP rates are listed in Tables 1 and 2. Free search performance between reader 2 and the other two readers was not significantly different (*P* = 0.07), but was significant between readers 1 and 3 (*P* = 0.01). After CAD interpretation,

the difference between readers 2 and 3 became significant ($P = 0.05$) but differences between reader 1 and the other two readers did not achieve significance ($P = 0.17$).

Table 3 summarises readers' decisions on CAD candidates representing nodules or non-nodules. The readers rejected TP CAD candidates heterogeneously, with readers 1 and 3 rejecting significantly more than reader 2 ($P < 0.001$). However, the readers were similar in their behaviour with a strong rejection of most of the FP CAD candidates ($P = 0.85$). Kappa values for reader pairs were 0.40 (R1–R2), 0.45 (R1–R3) and 0.40 (R2–R3) after free search and changed following CAD interpretation to 0.53 (R1–R2), 0.60 (R1–R3) and 0.42 (R2–R3).

Successive performance changes during CAD interpretation

Figure 2 shows FROC plots for free search and for successive use of CAD. For all three readers, CAD evaluation initially resulted in a sharp rise in sensitivity with any added CAD detection of approximately 14% with a simultaneously minimal increase in FP rate of 0.08 per patient. This was followed by an abrupt flattening of reader performance with only 2% sensitivity increase at a FP rate increase of now 0.22 per patient. The transitional knee-like bend in the FROC points with successive CAD interpretation occurred at confidence intervals 4, 3, 2 and 1. A “knee” was not observed with confidence level 5. The transitional “knee” occurred after a mean cumulative CAD evaluation time of approximately 100 s corresponding to a mean of 15 assessed CAD candidates.

Figure 3 demonstrates that the CAD score-ordered distribution of nodules versus non-nodules was skewed with nodules accounting for 63% of the CAD candidates at a CAD rank of 1 and dropping to 11% at a CAD rank of 10.

Figure 4 shows the performance of CAD interpretation using our time-invariant model. Successive FROC points demonstrated a “knee” if the input was set to the CAD score ranked distribution of nodules and non-nodules from the reader trial. Modelling a non-skewed, uniform distribution resulted in a linear distribution of FROC points with successive CAD interpretation and disappearance of the “knee”.

Temporal trends of radiologists' performance using CAD

Table 4 summarises the durations of free search and CAD evaluation. Readers averaged 71% of their interpretation in free search and 29% for CAD evaluation. Reader 1 was slowest and reader 2 fastest among the readers in both free search ($P < 0.001$) and CAD interpretation ($P < 0.001$). The durations for interpretation of free search and CAD evaluation between readers 1 and 3 were not significantly different ($P > 0.46$). The gain in sensitivity per minute of interaction was highest for reader 2, followed by readers 3 and 1.

Figure 5 demonstrates interpretation time averaged across the three readers and stratified into TP, FN, TN and FP reader decisions. Readers spent more time evaluating CAD candidates that were nodules (8.7 ± 4.6 s) than those that were not nodules (4.8 ± 1.3 s, $P < 0.0001$). Although the evaluation time for FN (= rejection of real nodule) reader decisions (8.4 ± 6.7 s) was similar to that of TP (= acceptance of real nodule) reader decisions (9.5 ± 4.5 s, $P = 0.47$), the evaluation time for FP (= acceptance of non-nodule) reader decisions

(14.4 s $\pm$ 8.7 s) was approximately twice as long ($P < 0.015$). Alternatively, TN (= rejection of non-nodule) reader decisions (4.7 s $\pm$ 1.3 s) required the least time ($P = 0.0002$) of all reader decisions.

Figure 6 shows that over time the specificity of readers stayed within the range of 98–100% and shows the sensitivity with greater time-dependent variability but no consistent trend from 69–100%.

Discussion

Our findings confirm the potential of CAD to improve radiologists' diagnostic performance in the detection of pulmonary nodules on chest MDCT. However, the magnitude of the performance increase with the use of CAD is influenced by the quality of the reader interpretation of CAD output as well as by the inherent technical properties of the algorithm. The diagnostic performance of CAD evaluation was variable among readers and did not follow the patterns of reader performance during free search. The time-varying performance observed during CAD evaluation seems to be attributable to the ordering of candidate evaluation by CAD score, resulting in a non-uniform presentation of TP and FP CAD detections with respect to time rather than as a result of variable reader performance during CAD evaluation.

Although a diversity of detection algorithms, image acquisition parameters, study designs and image databases impede a direct comparison of current CAD studies, our results (Fig. 1) describe a similar increase in sensitivity in lung nodule detection as reported in recent publications [2–6, 8, 9, 12, 18].

While there is ongoing research into reducing FP detections by CAD systems [19–21], studies have not yet addressed the fact that readers surprisingly tend to reject some TP CAD candidates and thus potentially reduce the benefit of CAD. In this study, approximately every second TP CAD candidate was rejected by two of the readers, while the other reader only rejected one in seven TP CAD candidates. This diversity in acceptance of TP CAD candidates was responsible for the variable benefit readers achieved from the use of CAD. Similarly diverse behaviour was observed in the acceptance of FP CAD detections. In general, good reader performance during free search did not automatically ensure good performance when evaluating CAD output. One explanation for this variability is the subjectivity of readers' decisions when classifying pulmonary findings. Armato et al. [22] recently described substantial variability across very experienced thoracic radiologists in the determination of 'truth' for the first 30 cases of the Lung Nodule Database Consortium (LIDC). Only 18.5% of 443 lesions identified during unblinded reading were categorised in the same way by all four radiologists. Our results indicate a similar degree of classification heterogeneity.

Overall, readers spent approximately one-third of their total reading time evaluating CAD candidates; however, the duration of CAD interpretation did not correlate with the diagnostic benefit of CAD. To our knowledge, data have not been published that assess time spent on correct versus incorrect reader decisions during CAD evaluation. In our study,

readers spent more time evaluating nodules than non-nodules. Although correct and incorrect decisions took about the same amount of time, the incorrect acceptance of non-nodules took two- to three-times longer than their correct rejection. A possible explanation is that many FP CAD hits are obvious non-nodular structures, whereas FP CAD detections with some nodular imaging features require more manipulation to distinguish them from true nodules.

If a reader were to operate with a constant FP rate during CAD evaluation, one might intuitively expect to see linearly increasing sensitivity as more CAD candidates are presented, resulting in FROC operating points that move upwards and to the right with a constant slope. In contrast, we observed a knee-like bend in the trajectory of the FROC operating points. We hypothesised that the skewed distribution of CAD output with preferential ordering of true nodules to low CAD ranks led to an early improvement in sensitivity with little increase in FP results. However, as the prevalence of TP CAD detections fell and that of FP detections increased, readers had fewer opportunities to increase sensitivity and more opportunities for FP acceptance. Modelling a uniform distribution of TP and FP CAD candidates confirmed this hypothesis, resulting in linear FROC curves and disappearance of the “knee” phenomenon. Thus, the “knee” appearance can be explained by the tendency of the CAD algorithm to front-load its output with TP CAD candidates and not by time-varying reader performance. This conclusion is supported by the absence of major temporal trends in readers’ accuracy through the CAD evaluation.

It is obvious that the large number of CAD candidates provided in this study (up to 50 CAD candidates per case) does not represent a clinically optimised CAD output, but it laid the foundation for determining when a “knee” in performance occurs, which could be valuable in clinical practice, as it represents a point at which the efficiency (balance of TP versus FP detections) of CAD interpretation diminishes. This particular transition occurred at around 100 s of CAD output evaluation, which corresponds to a mean of 15 assessed CAD candidates. Confirmation of the consistency of this transition could provide a principled means of establishing prospectively the number of CAD candidates to be presented and the duration of CAD evaluation, based upon a desired balance between TP and FP acceptance. It also may indicate how the performance characteristics of a CAD system (sensitivity relative to FP frequency) can impact on reader performance.

There were several limitations to our study. Without histological proof, lung nodules are radiological observations and thus tend to be interpreted variably among even experienced thoracic radiologists [22]. We believe, however, that our multi-phase interpretation of the sum of all detections by a consensus panel helped to minimise any effect of existing reader variability on the ‘truth’ in this study. Further, we assessed only one CAD candidate distribution. Other distributions, such as ordering candidates by location, as well as assessment of morphological features of CAD candidates being erroneously rejected or accepted are part of an ongoing investigation. While recent publications recommend follow-up of 4-mm lung nodules when detected incidentally [23], we chose a 3-mm size threshold for data analysis to encompass nodules that would be relevant in higher risk populations. While the number of patients was limited to 20, our analysis treated each CAD detection independently, providing a comparable sample size to other published data analysing CAD

for lung nodule detection [1, 4, 8, 12, 22, 24]. Finally, a potential learning bias of the readers during their CAD evaluation could not be completely excluded.

In summary, we have demonstrated that radiologists derive variable benefit from CAD and that performance during CAD interpretation does not mirror performance during free search. However, CAD has the potential to equalise performance among readers by reducing individual detection errors of lung nodules on chest CT. Analysis of CAD-dependent temporal variations in performance may facilitate the establishment of optimised thresholds for the duration of CAD interaction.

References

1. Beigelman-Aubry C, Raffy P, Yang W, Castellino RA, Grenier PA. Computer-aided detection of solid lung nodules on follow-up MDCT screening: evaluation of detection, tracking, and reading time. *AJR Am J Roentgenol*. 2007; 189:948–955. [PubMed: 17885070]
2. Beyer F, Zierott L, Fallenberg EM, Juergens KU, Stoeckel J, Heindel W, Wormanns D. Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader. *Eur Radiol*. 2007; 17:2941–2947. [PubMed: 17929026]
3. Godoy MC, Cooperberg PL, Maizlin ZV, Yuan R, McWilliams A, Lam S, Mayo JR. Detection sensitivity of a commercial lung nodule CAD system in a series of pathologically proven lung cancers. *J Thorac Imaging*. 2008; 23:1–6. [PubMed: 18347512]
4. Jankowski A, Martinelli T, Timsit JF, Brambilla C, Thony F, Coulomb M, Ferretti G. Pulmonary nodule detection on MDCT images: evaluation of diagnostic performance using thin axial images, maximum intensity projections, and computer-assisted detection. *Eur Radiol*. 2007; 17:3148–3156. [PubMed: 17763856]
5. Marten K, Engelke C. Computer-aided detection and automated CT volumetry of pulmonary nodules. *Eur Radiol*. 2007; 17:888–901. [PubMed: 17047961]
6. Rubin GD, Lyo JK, Paik DS, Sherbondy AJ, Chow LC, Leung AN, Mindelzun R, Schraedley-Desmond PK, Zinck SE, Naidich DP, Napel S. Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. *Radiology*. 2005; 234:274–283. [PubMed: 15537839]
7. Saba L, Caddeo G, Mallarini G. Computer-aided detection of pulmonary nodules in computed tomography: analysis and review of the literature. *J Comput Assist Tomogr*. 2007; 31:611–619. [PubMed: 17882043]
8. White CS, Pugatch R, Koonce T, Rust SW, Dharaiya E. Lung nodule CAD software as a second reader: a multicenter study. *Acad Radiol*. 2008; 15:326–333. [PubMed: 18280930]
9. Yuan R, Vos PM, Cooperberg PL. Computer-aided detection in screening CT for pulmonary nodules. *AJR Am J Roentgenol*. 2006; 186:1280–1287. [PubMed: 16632719]
10. Das M, Muhlenbruch G, Heinen S, Mahnken AH, Salganicoff M, Stanzel S, Gunther RW, Wildberger JE. Performance evaluation of a computer-aided detection algorithm for solid pulmonary nodules in low-dose and standard-dose MDCT chest examinations and its influence on radiologists. *Br J Radiol*. 2008; 81:841–847. [PubMed: 18941043]
11. Lee JY, Chung MJ, Yi CA, Lee KS. Ultra-low-dose MDCT of the chest: influence on automated lung nodule detection. *Korean J Radiol*. 2008; 9:95–101. [PubMed: 18385555]
12. Das M, Muhlenbruch G, Mahnken AH, Flohr TG, Gundel L, Stanzel S, Kraus T, Gunther RW, Wildberger JE. Small pulmonary nodules: effect of two computer-aided detection systems on radiologist performance. *Radiology*. 2006; 241:564–571. [PubMed: 17057074]
13. Zheng B, Leader JK, Abrams G, Shindel B, Catullo V, Good WF, Gur D. Computer-aided detection schemes: the effect of limiting the number of cued regions in each case. *AJR Am J Roentgenol*. 2004; 182:579–583. [PubMed: 14975949]
14. Paik DS, Beaulieu CF, Rubin GD, Acar B, Jeffrey RB Jr, Yee J, Dey J, Napel S. Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT. *IEEE Trans Med Imaging*. 2004; 23:661–675. [PubMed: 15191141]

15. Cheng H, Macaluso M. Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results. *Epidemiology*. 1997; 8:104–106. [PubMed: 9116081]
16. Eliasziw M, Donner A. Application of the McNemar test to non-independent matched pair data. *Stat Med*. 1991; 10:1981–1991. [PubMed: 1805322]
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]
18. Brochu B, Beigelman-Aubry C, Goldmard JL, Raffy P, Grenier PA, Lucidarme O. Computer-aided detection of lung nodules on thin collimation MDCT: impact on radiologists' performance. *J Radiol*. 2007; 88:573–578. [PubMed: 17464256]
19. Suzuki K, Armato SG 3rd, Li F, Sone S, Doi K. Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. *Med Phys*. 2003; 30:1602–1617. [PubMed: 12906178]
20. Ge Z, Sahiner B, Chan HP, Hadjiiski LM, Cascade PN, Bogot N, Kazerooni EA, Wei J, Zhou C. Computer-aided detection of lung nodules: false positive reduction using a 3D gradient field method and 3D ellipsoid fitting. *Med Phys*. 2005; 32:2443–2454. [PubMed: 16193773]
21. Roy AS, Armato SG 3rd, Wilson A, Drukker K. Automated detection of lung nodules in CT scans: false-positive reduction with the radial-gradient index. *Med Phys*. 2006; 33:1133–1140. [PubMed: 16696491]
22. Armato SG 3rd, McNitt-Gray MF, Reeves AP, Meyer CR, McLennan G, Aberle DR, Kazerooni EA, MacMahon H, van Beek EJ, Yankelevitz D, Hoffman EA, Henschke CI, Roberts RY, Brown MS, Engelmann RM, Pais RC, Piker CW, Qing D, Kocherginsky M, Croft BY, Clarke LP. The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. *Acad Radiol*. 2007; 14:1409–1421. [PubMed: 17964464]
23. MacMahon H, Austin JH, Gamsu G, Herold CJ, Jett JR, Naidich DP, Patz EF Jr, Swensen SJ. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. *Radiology*. 2005; 237:395–400. [PubMed: 16244247]
24. Brown MS, Goldin JG, Rogers S, Kim HJ, Suh RD, McNitt-Gray MF, Shah SK, Truong D, Brown K, Sayre JW, Gjertson DW, Batra P, Aberle DR. Computer-aided lung nodule detection in CT: results of large-scale observer test. *Acad Radiol*. 2005; 12:681–686. [PubMed: 15935966]

Fig. 1.

The mean sensitivities for the three readers after their individual free search (*w/o CAD*) and the use of CAD (*w CAD*). *Numbers in parentheses* represent the mean FP rate per patient. Non-calcified pulmonary nodules (≥ 3 mm) were detected with a sensitivity of 59%, 57% and 44% during free search, which increased with CAD up to 67%, 82% and 60%, respectively. The FP rate was between 0.6 and 2.1 FP/patient for their free search and 0.85 and 2.7 FP/patient with CAD. For a balanced comparison, performance with a constant FP rate of 0.5 FP/patient was plotted

Fig. 2.

Free-response ROC after free search and incremental use of CAD. Free-response ROC plots show the mean sensitivity versus FP detections per patient for all and each individual reader. *Numbers inside the brackets* represent the reader confidence levels. Sensitivities achieved after free search are represented by the *grey dot-dash line*; the increase in sensitivity with successive use of CAD (increments of one CAD candidate) is shown as *black points*. Note how the incremental use of CAD resulted in an initial sharp rise in sensitivity with minimal increase in FP detections followed by an abrupt flattening of reader performance describing a knee-like bend with the exception of confidence level 5, where the sensitivity increase does not come with additional FPs. This phenomenon was observed in all three readers

Fig. 3.

Distribution of CAD candidates representing nodules and non-nodules over 20 subjects. Stacked area plot demonstrates the trend of the percentage of CAD candidates representing nodules (*black area*) or non-nodules (*grey area*) over time reflected by increasing CAD ranks. A clear trend of 'early' CAD candidates (low CAD rank) being rich in nodules and 'late' CAD candidates having fewer nodules can be observed. This positively skewed or 'frontloaded' distribution shows an approximately threefold drop in nodule number and a twofold increase in non-nodule number following the first ten of the 50 evaluated CAD candidates

Fig. 4.

Time-invariant model: influence of nodule distribution on reader performance. Free-response ROC plots show the mean sensitivity versus FP detections per patient for the statistical model. *Numbers inside the brackets* represent the confidence level intervals. The reader sensitivities achieved after free search are represented by the *gray dot-dash line*; the increase in sensitivity with the incremental use of CAD (increments of one CAD candidate) is shown as *white boxes*. The *black line* represents the model's output to the same nodule distribution as shown to the readers. Note the similar knee like bend between the model and the reader trial. The *grey line* represents changing the model's input to a uniform nodule distribution. In this case, the knee like bend is no longer observed

Fig. 5.

Reader speed for correct and incorrect decisions during CAD evaluation. **a** The mean time spent for acceptance (*TP Nodules*) or rejection (*FN Nodules*) of CAD candidates representing true nodules. **b** The mean time spent for acceptance (*FP Non-nodules*) and rejection (*TN Non-nodules*) of CAD candidates representing non-nodules. The darker lines show correct reader interpretations, while the lighter lines show incorrect interpretations. Plots were smoothed with a moving average of ± 3 CAD ranks

Performance over Time w CAD

Fig. 6.

Temporal changes in diagnostic performance during CAD evaluation. Temporal trends of sensitivity and specificity do not show a particular temporal trend over the duration of the CAD evaluation. The late dip in sensitivity is caused by the fact that the number of TP nodules gets very small towards the higher CAD ranks and thus a change of a single nodule leads to a larger change in sensitivity. The specificity, however, does not change as a result of the large number of non-nodules towards the end of the CAD evaluation. Plots were smoothed with a moving average of ± 3 CAD ranks.

Table 1

Relative sensitivities at three FP/rates for free search and CAD (R1 reader 1, R2 reader 2, R3 reader 3, FP/Pt false positives per patient)

FP/Pt	1			2		
	Free search	CAD	P	Free search	CAD	P
R1	0.58	0.63	0.48	0.60	0.67	0.18
R2	0.51	0.66	0.004	0.54	0.71	<0.001
R3	0.42	0.54	0.03	0.44	0.60	0.005
						0.005

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

P values for pair-wise reader comparisons during free search and CAD evaluation

Pairs	Free search			CAD		
	FP/Pt			FP/Pt		
	0.5	1	2	0.5	1	2
R1-R2	0.54	0.8	0.39	0.99	0.99	0.17
R1-R3	0.01	0.01	<0.001	0.28	0.37	0.17
R2-R3	0.3	0.19	0.07	0.05	0.05	<0.001

Reader (*R*) decisions on CAD candidates representing nodules or non-nodules (*TN* true negative = correct rejection of non-nodule, *FP* false positive = incorrect acceptance of non-nodule, *TP* true positive = correct acceptance of nodule, *FN* false negative = incorrect rejection of nodule)

Table 3

Decision type	R1	R2	R3	R1	R2	R3	R1	R2	R3	%
All patients										
Decision on CAD candidates that were non-nodules										
TN	756	751	760	37.8	37.55	38	99.3	98.4	99.9	
FP	5	12	1	0.25	0.6	0.05	0.7	1.6	0.1	
Decision on CAD candidates that were nodules										
TP	15	47	29	0.75	2.35	1.45	41.7	85.5	46.0	
FN	21	8	34	1.05	0.4	1.7	58.3	14.5	54.0	

Table 4

Mean duration of free search and CAD evaluation per case (*Min* mean time in minutes with standard deviation in parentheses, *sens* sensitivity, % *sens/min* percental increase in sensitivity per 1 min)

	All Readers				Reader 1				Reader 2				Reader 3			
	Min	% of time	% sens/min		Min	% of time	% sens/min		Min	% of time	% sens/min		Min	% of time	% sens/min	
Free Search	9.7 (5.3)	71	5.5		12.3 (5.5)	73	4.8		4.9 (2.0)	63	11.6		11.8 (4.0)	74	3.7	
CAD	3.9 (1.6)	29	4.1		4.6 (2.0)	27	1.7		2.9 (0.9)	37	8.6		4.2 (1.5)	26	3.8	
Total	13.6 (6.0)	100	5.1		16.9 (5.7)	100	4.0		7.8 (2.3)	100	10.5		16.0 (4.5)	100	3.8	