The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/89546

Please be advised that this information was generated on 2019-10-07 and may be subject to change.
Computer-aided Detection of Lung Cancer on Chest Radiographs: Effect on Observer Performance

Bartjan de Hoop, MD
Diederik W. De Boo, MD
Hester A. Gietema, MD, PhD
Frans van Hoom, MD
Banafsche Mearadji, MD
Laura Schijf, MD
Bram van Ginneken, PhD
Mathias Prokop, MD, PhD
Cornelia Schaefer-Prokop, MD, PhD

Purpose:
To assess how computer-aided detection (CAD) affects reader performance in detecting early lung cancer on chest radiographs.

Materials and Methods:
In this ethics committee–approved study, 46 individuals with 49 computed tomographically (CT)-detected and histologically proved lung cancers and 65 patients without nodules at CT were retrospectively included. All subjects participated in a lung cancer screening trial. Chest radiographs were obtained within 2 months after screening CT. Four radiology residents and two experienced radiologists were asked to identify and localize potential cancers on the chest radiographs, first without and subsequently with the use of CAD software. A figure of merit was calculated by using free-response receiver operating characteristic analysis.

Results:
Tumor diameter ranged from 5.1 to 50.7 mm (median, 11.8 mm). Fifty-one percent (22 of 49) of lesions were subtle and detected by two or fewer readers. Stand-alone CAD sensitivity was 61%, with an average of 2.4 false-positive annotations per chest radiograph. Average sensitivity was 63% for radiologists at 0.23 false-positive annotations per chest radiograph and 49% for residents at 0.45 false-positive annotations per chest radiograph. Figure of merit did not change significantly for any of the observers after using CAD. CAD marked between five and 16 cancers that were initially missed by the readers. These correctly CAD-depicted lesions were rejected by radiologists in 92% of cases and by residents in 77% of cases.

Conclusion:
The sensitivity of CAD in identifying lung cancers depicted with CT screening was similar to that of experienced radiologists. However, CAD did not improve cancer detection because, especially for subtle lesions, observers were unable to sufficiently differentiate true-positive from false-positive annotations.

© RSNA, 2010
Computer-aided Detection of Lung Cancer on Chest Radiographs

**Study Population**

All chest radiographs used in this study were retrospectively collected from participants from two centers (Utrecht and Groningen, the Netherlands) of the Dutch-Belgian Randomized Lung Cancer Screening, or NELSON, trial (12). This trial was approved by the Ministry of Health and by the ethics committee of each participating hospital. Participants were aged between 50 and 75 years and were current or former heavy smokers, reflecting a population with high risk of developing lung cancer.

In this population of 4938 participants at the two centers, chest radiographs may be ordered for preoperative routine work-up or for follow up of screening-detected lesions and also for clinical causes unrelated to screening. We included all chest radiographs obtained between April 2004 and January 2008 in this group of subjects under the following conditions: In patients in whom a pulmonary malignancy was detected at screening CT and was histologically confirmed (cases), chest radiography had to be performed within 6 weeks after screening CT; in the other subjects (control subjects), chest radiography had to be performed within 2 months of screening CT and no nodules larger than 5 mm in diameter had to be present at the screening CT. In total, the control subjects had 43 nodules that were smaller than 5 mm at CT. These nodules had an average diameter of 3.9 mm (range, 3.1–5.0 mm) and did not show malignant growth during a minimal 2-year follow-up in the CT lung cancer screening study. None of the control subjects developed lung cancer during this follow-up period. Chest radiographs for which the radiology report mentioned pulmonary abnormalities other than chronic obstructive pulmonary disease were excluded.

**Acquisition of Images**

All chest radiographs were obtained by using a cesium iodide amorphous silicon detector.
THORACIC IMAGING: Computer-aided Detection of Lung Cancer on Chest Radiographs

de Hoop et al

flat-panel-detector unit (DigitalDiagnost; Philips, Best, the Netherlands). Images were processed by using nonlinear multifrequency-band processing (13); parameters recommended by the manufacturer were used. For all patients, posteroanterior and lateral projections were available.

The screening CT examinations were performed with 16 × 0.75-mm collimation at 30 mAs and 120–140 kV, depending on weight. Sections of 1 mm thickness were reconstructed every 0.7 mm.

Standard of Reference

In the cancer-positive cases, the exact location of each nodule on a chest radiograph was determined by two observers who did not participate as a reader (B.d.H., radiology researcher with 3 years experience in reading CT lung cancer screening studies). In case of doubt, he consulted an independent chest radiologist (M.P.). This chest radiologist also judged whether lesions were retrospectively visible on a chest radiograph. Both had access to chest radiographs, as well as screening CT scans. Lesions that were, even with the knowledge of the CT findings, not visible on the chest radiograph were excluded from analysis.

Findings of all screening CT examinations were evaluated for nodules according to the criteria set by the lung cancer screening program (14). Volumetric software (Lung Care 5 VB10A-W; Siemens Medical Solutions, Erlangen, Germany) was used to assess nodule volume. This volume was used to calculate the diameter on the basis of the assumption of a perfect sphere.

CAD System

We used a commercially available CAD system (Onguard 5.0; Riverain, Miamisburg, Ohio). The software highlights regions suspicious for containing a focal lung lesion by placing a circle of 5 cm in diameter around the suspicious area (Fig 1). Images are automatically processed in the background so that results are immediately available on demand when the chest radiograph is being read by a radiologist. The program only analyzes the posteroanterior or anteroposterior projection. According to the manufacturer, the algorithm was optimized to detect nodules of 9–30 mm in diameter, although in practice, it also marks larger and smaller nodules.

Stand-Alone CAD Performance

To assess stand-alone performance of the CAD system, annotations were labeled TP if the suspicious lesion was located at least partially within the central 50% of the circular CAD annotation.

Observer Study

Images were evaluated on Digital Imaging and Communications in Medicine-calibrated liquid crystal display monitors (MFGD 3220D; Barco, Kortrijk, Belgium) with a matrix size of 2048 × 1536. Options for magnification and adaptation of window settings were available. All chest radiographs were anonymized. Posteroanterior and lateral images were available for evaluation. Chest radiographs were shown in alphabetical order on the basis of patient name to six independent observers. The observers varied in their level of experience: one general radiologist with 6 years of experience (observer A), one chest radiologist with more than 20 years of experience (observer B), and four radiology residents with experience that varied from 1 to 4 years (observers C–F). Observers knew that the study group was chosen from a lung cancer screening trial and they were also told that some patients might have more than one malignant lesion. Two of the observers (reader B and E) had used the CAD system before during other reader studies, but none of the readers had routine experience. To familiarize the observers with the CAD system, five cancer cases that were not included in the observer study were shown to the observers without and with CAD annotations before the start of the study.

Each chest radiograph was first evaluated without and subsequently with CAD results, and observer readings were recorded separately. On a per-patient basis, the observers were asked to document all potentially malignant focal abnormalities seen on the chest radiograph on a separate paper printout with respect to the anatomic lesion locations.
and the readers’ confidence scores by using a four-point scale (score of 1: potential lesion, very low degree of suspicion; score of 2: dubious lesion; score of 3: probable lesion; and score of 4, definite lesion). Observers were allowed to mark multiple suspicious lesions on each chest radiograph. They were instructed, however, to ignore nodules smaller than 5 mm in diameter. The researcher (B.d.H.) and the experienced radiologist (M.P.) who had not been involved in the readings analyzed all paper printouts, with the chest radiographs and CT scans being available. The readers’ markings were considered TP if the centers of the markings were within the boundaries of the nodules on the chest radiograph. Locations that did not match with a lesion were classified as FP.

**Data Analysis**

Free-response receiver operating characteristic (FROC) analysis of the observer study was performed as described by Swensson (15) on a per-marker basis. Jackknife FROC, especially developed to analyze observer free-response tasks (16-18), was used to analyze the FROC data. Jackknife FROC software (JAFROC, version 2.3a; http://www.devchakraborty.com) (16,19) was used to compute a figure of merit (FOM). The FOM is defined as the probability that lesions (including unmarked lesions) are rated higher than nonlesion marks on control chest radiographs (17), or, in other words, that lesions are given a higher confidence rating for the presence of malignancy than normal findings. Normal images with no marks and unmarked lesions are assigned a zero rating. The level of significance was corrected for multiple comparisons by using Bonferroni correction.

Sensitivity was calculated as the number of TP markings divided by the total number of malignancies. All observer markings, even those that were scored with low confidence, were included to calculate the sensitivity and the FP rate.

Since it is controversial whether application of CAD as a second reader also allows for discharge of candidates seen without CAD (20), we also evaluated a situation in which the observers could only increase their suspicion with CAD while preserving all lesion locations seen without CAD.

In an effort to understand the effect of lesion conspicuity on our results, we performed a separate jackknife FROC analysis on conspicuous nodules, defined as lesions that were detected by three or more readers.

To test for demographic differences between the cases and the control subjects, we compared both groups with respect to sex by using a $\chi^2$ test and age by using a Student t test. $P$ values less than .05 were considered to indicate a significant difference.

**Results**

**Sample Characteristics**

A total of 46 participants with 49 histologically proved pulmonary malignancies met the criteria for the cancer-positive cases. Sixty-five subjects met the criteria for control cases. Indications for acquisition of the chest radiograph in the control group were exclusion of acute cardiovascular disease ($n = 18$), chronic obstructive pulmonary disease ($n = 18$), screening for lung abnormalities because of rheumatoid arthritis ($n = 10$), preoperative screening ($n = 10$), unexplained fever ($n = 4$), chronic cough ($n = 3$), malaise ($n = 1$), and trauma ($n = 1$).

Cases did not differ significantly from the control subjects with respect to age and sex (Table 1). Tumor diameter ranged from 5.1 to 50.7 mm (median, 12.0 mm), with two lesions being larger than 30 mm. Conspicuity of malignancies was very variable: Ten of 49 (20%) malignancies were detected by all six observers without the use of CAD. Furthermore, 11 malignancies were detected by five observers, two were detected by four observers, six were detected by three observers, five were detected by two observers, and seven were detected by only one observer without the use of CAD. Eight (16%) malignancies were not detected by any of the observers without or with the use of CAD. None of the 43 small benign nodules in the control group was marked by either the CAD system or any of the observers.

**CAD Stand-Alone Performance**

The CAD stand-alone sensitivity was 61% (30 of 49), with on average 2.4 FP annotations (range, zero to five) per chest radiograph. CAD depicted three malignancies that were initially not detected by any of the observers. The diameter of the CAD-depicted malignancies ranged from 7.0 to 50.7 mm.

**Observer Performance without CAD**

Without CAD, the FOM was 0.72 for radiologists and 0.58 for residents (Table 2, Fig 2). The radiologists had an average sensitivity of 63%, with 0.23 FP annotations per chest radiograph. The residents had an average sensitivity of 49%, with 0.45 FP annotations per chest radiograph. Twenty-seven lesions were detected by at least three observers. In this subselection of more conspicuous lesions, the average FOM was 0.93 for radiologists and 0.76 for residents, with an average sensitivity of 96% for radiologists and 75% for residents.

**Observer Performance with CAD When Lowering of Confidence Scores Was Allowed**

When the readers were allowed to change their ratings depending on CAD...
Observer Performance with CAD When Lowering of Confidence Scores Was Not Allowed

When readers were only allowed to increase their confidence scores after having CAD results, average FOM decreased from 0.72 to 0.70 for radiologists (P < .001) and from 0.58 to 0.57 for residents (P = .6). Average sensitivity remained virtually unchanged, 61% and 51%, respectively. Specificity improved, from 0.23 to 0.19 FP annotations per chest radiograph for radiologists and from 0.45 to 0.36 FP annotations for residents.

In the subselection of conspicuous lesions, average FOM remained 0.93 for radiologists, but significantly improved for residents (from 0.76 to 0.82, P < .001). Sensitivity remained 96% for radiologists, but improved from 75% to 84% for residents.

Figure 2: Alternative FROC curves for detection of pulmonary malignancies by (a, c) residents and (b, d) radiologists. Separate analysis for all lesions (a, b) and more conspicuous lesions seen by more than two observers (c, d) was performed. The FOM, which is area under the alternative FROC curve, improved significantly for detection of more conspicuous lesions by residents if they were allowed to freely adjust their level of confidence after being provided with the CAD output. The remaining alternative FROC curves did not significantly change with use of CAD.

Figure 2

![Alternative FROC curves](image)

suggestions, average FOM for the radiologists did not change (0.72, P = .98). Average FOM for the residents increased from 0.58 to 0.61, but the improvement was not significant (P = .08) (Table 2). With CAD, the average sensitivity of radiologists and residents remained virtually unchanged, 61% and 51%, respectively. Specificity improved, from 0.23 to 0.19 FP annotations per chest radiograph for radiologists and from 0.45 to 0.36 FP annotations for residents.

In the subselection of conspicuous lesions, average FOM remained 0.93 for radiologists, but significantly improved for residents (from 0.76 to 0.82, P < .001). Sensitivity remained 96% for radiologists, but improved from 75% to 84% for residents.
63%–65%) for radiologists and from 49% to 55% (range, 41%–69%) for residents, but the average number of FP annotations per chest radiograph also increased from 0.23 to 0.31 and from 0.45 to 0.54, respectively.

**Interaction between CAD and Readers**

Together, the six observers placed a total of 66 new markings after having CAD results: 12 for TP CAD annotations and 54 for FP CAD annotations. The number of additionally detected malignancies following TP CAD annotations ranged from zero to six for the various observers (Table 3). The residents benefited more from CAD than did the radiologists, but they also accepted more FP CAD annotations, on average one per 11 chest radiographs versus one per 19 chest radiographs for the radiologists.

Observers A, B, C, D, E, and F, respectively, dismissed 23, 4, 35, 35, 3, and 17 of their own initial markings because CAD had not annotated these regions (Table 4). The number of malignancies initially not seen by the observers but correctly annotated by CAD varied between five and 16 per observer. Eighty percent (47 of 59) of these TP CAD annotations were rejected by the observers (Table 3). An example is shown in Figure 3.

The average confidence levels were generally low for new TP markings, new FP markings, and markings that were initially called but later dismissed after seeing CAD annotations, with confidence levels of 1.9, 1.8, and 1.6, respectively.

**Discussion**

In this study we assessed how recently released, commercially available CAD software affected reader performance in detecting early lung cancer on chest radiographs. Stand-alone sensitivity of CAD was virtually identical to that of experienced radiologists: 61% in a dataset where 16% of the nodules were not detected by any of the observers. However, the number of FP annotations per chest radiograph was, on average, 10 times higher with CAD than with the two radiologists. The number of CAD-annotated malignancies that were initially not detected by observers varied between five and 16 per observer, out of a total of 49 malignancies, which indicates a vast potential for CAD to improve reader performance. Still, no significant improvement in observer performance could be demonstrated with use of CAD as a second reader in the detection of nodules on chest radiographs.

---

**Table 2**

<table>
<thead>
<tr>
<th>Variable</th>
<th>FOM</th>
<th>Sensitivity (%)</th>
<th>FP Markings per Chest Radiograph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radiologist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.73</td>
<td>63</td>
<td>0.25</td>
</tr>
<tr>
<td>B</td>
<td>0.71</td>
<td>63</td>
<td>0.22</td>
</tr>
<tr>
<td>Average</td>
<td>0.72</td>
<td>63</td>
<td>0.23</td>
</tr>
<tr>
<td>Resident</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.47</td>
<td>29</td>
<td>0.59</td>
</tr>
<tr>
<td>D</td>
<td>0.60</td>
<td>69</td>
<td>0.75</td>
</tr>
<tr>
<td>E</td>
<td>0.62</td>
<td>37</td>
<td>0.13</td>
</tr>
<tr>
<td>F</td>
<td>0.62</td>
<td>51</td>
<td>0.32</td>
</tr>
<tr>
<td>Average</td>
<td>0.58</td>
<td>49</td>
<td>0.45</td>
</tr>
</tbody>
</table>

*FN = False-negative, TN = true-negative.

---

**Table 3**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Radiologist</th>
<th>Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>No. of TP CAD annotations</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>annotations initially not</td>
<td></td>
<td></td>
</tr>
<tr>
<td>detected by observers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of rejected TP CAD</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>annotations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—CAD correctly annotated 30 malignancies. Most TP CAD annotations were rejected by the observers.

---

**Table 4**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Radiologist</th>
<th>Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Effect of CAD*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN to TP markings</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FP to TN markings</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN to FP markings</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>TP to FN markings</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

*FN = False-negative, TN = true-negative.
An interesting observation is that in the current study, CAD did not improve observer performance. The reason is not that the observers disregarded the CAD annotations; on the contrary, in total, 66 CAD annotations were accepted and 117 initial observer markings were removed because CAD did not annotate the corresponding region. The 66 accepted annotations, pooled over all observers, were 12 TP CAD annotations of lesions initially missed and 54 FP CAD annotations. Among the 117 removed markings were 11 TP lesions. This shows that the observers had difficulties differentiating TP from FP CAD annotations.

This principle has previously been described in a chest radiograph nodule detection study in which eye-tracking was used. In that study, only a minority of the lesions were missed due to inefficient search. The dominant cause of unreported nodules proved to be incorrect decision-making (21). This has also been described in a study that used CAD for detection, as well as classification of suspicious regions (22). The detection function of that CAD system annotated suspicious regions, but only slightly increased the number of lung cancers detected by the observers. Similar to our study, cancers initially missed by the observers but correctly annotated by CAD were frequently rejected by the observers. The authors report that the missed cases were mainly subtle lesions. The reported improvement in radiologists’ performance was mainly due to the classification function that computed the likelihood of malignancy for regions indicated by the observer. Using this information, the observer could then change his or her initial decision.

All malignancies included in our study were depicted with CT during lung cancer screening. Malignancies detected during CT screening are usually in an early stage and consequently more difficult to recognize on chest radiographs (23), a fact that is reflected in this study by the relatively low sensitivity of the observers. In a previous study (24) analyzing CAD, pulmonary malignancies that were inadequately visible on chest radiographs were excluded from the analysis. A very high area under the ROC curve of 0.92 was reported without the use of CAD. Observer performance was reported to improve significantly with CAD, and detection became almost flawless. When we excluded subtle lesions from our database and repeated the analysis, we also found excellent performance for radiologists (FOM, 0.93) and significant improvement in FOM to 0.82 for residents. These results show that classification is less problematic in nonsubtle lesions and the benefit of CAD is larger in more conspicuous cases, although such obvious lesions are less likely to be missed in the first place by experienced radiologists.

Figure 3: (a) Chest radiograph and (b) CT scan of correctly CAD-annotated adenocarcinoma (arrow). Both radiologists detected the tumor without CAD, but none of the four residents marked the region, even after seeing CAD results.
We showed that to improve observer performance for subtle lesions, observers need to learn to better differentiate between TP and FP CAD annotations. Observer training to recognize FP CAD annotations or a change in how CAD presents results might lead to this goal. In that respect, the lack of training of our observers might have contributed to the low positive effect of CAD. On the other hand, it is, to date, unknown how much training would be necessary and how strong such learning effects would be.

CAD systems of the future may not only provide annotations, but also assign likelihood that an annotation is a true lesion. Alternatively, CAD may just display the likelihood of a suspicious abnormality, because observers knew that the prevalence of cases was higher than in a normal screening situation. In practice, lower detections rates for the observers may therefore be likely. How far that will affect their attitude toward positive CAD markings is unknown. Finally, although we did not find significant improvement in FOM with the use of CAD, we did find a strong trend for the residents. All residents showed an equal or higher FOM with the use of CAD, with an improvement that reached a \( P \) level of .08. It is likely that this improvement would have yielded statistical significance when more cases or observers had been included. More research is needed to confirm this trend for the use of this CAD system by residents.

We conclude that the detection rate of pulmonary malignancies on chest radiographs is comparable for current CAD software and experienced radiologists. However, the positive predictive value of CAD was limited by the high FP rate. Because observers were unable to sufficiently differentiate TP from FP annotations, CAD did not significantly change nodule detection performance. CAD significantly improved detection of more conspicuous lesions by less experienced observers. For subtle lesions, however, additional measures are needed to be able to take advantage of lesions that were missed by observers but were annotated by CAD. Special training of readers might help them differentiate TP from FP CAD annotations. As an alternative, CAD findings might be presented so that they also provide an estimation of the probability of malignancy.

References

12. van Iersel CA, de Koning HJ, Draisma G, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice...


