Clinical evaluation of a deep-learning-based computer-aided detection system for the detection of pulmonary nodules in a large teaching hospital

C.O. Martins Jarnalo a,*, P.V.M. Linsen a, S.P. Blazís b, P.H.M. van der Valk a, D.B.M. Dickerscheid b

a Department of Radiology, Albert Schweitzer Ziekenhuis, Dordrecht, the Netherlands
b Department of Clinical Physics, FP, the Netherlands

AIM: To evaluate a deep-learning-based computer-aided detection (DL-CAD) software system for pulmonary nodule detection on computed tomography (CT) images and assess its added value in the clinical practice of a large teaching hospital.

MATERIALS AND METHODS: A retrospective analysis was performed of 145 chest CT examinations by comparing the output of the DL-CAD software with a reference standard based on the consensus reading of three radiologists. For every nodule in each scan, the location, composition, and maximum diameter in the axial plane were recorded. The subgroup of chest CT examinations (n = 97) without any nodules was used to determine the negative predictive value at the given clinical sensitivity threshold setting.

RESULTS: The radiologists found 91 nodules and the CAD system 130 nodules of which 80 were true positive. The measured sensitivity was 88% and the mean false-positive rate was 1.04 false positives/scan. The negative predictive value was 95%. For 23 nodules, there was a size discrepancy of which 19 (83%) were measured smaller by the radiologist. The agreement of nodule composition between the CAD results and the reference standard was 95%.

CONCLUSIONS: The present study found a sensitivity of 88% and a false-positive rate of 1.04 false positives/scan. Together with the measured negative predictive value of 95% the system performs very well; however, these rates are still not good enough to replace the radiologist, even for the specific task of nodule detection. Furthermore, a surprisingly high rate of overestimation of nodule size was observed, which can lead to too many follow-up examinations.

© 2021 The Royal College of Radiologists. Published by Elsevier Ltd. All rights reserved.
Introduction

Lung cancer is the most common malignancy and causes the most cancer-related deaths each year worldwide, around 1.76 million in 2018. Annually over 13,000 people in the Netherlands are diagnosed with lung cancer, which makes it the fourth most common form of cancer in both men and women. With the recent publication of the Nelson trial results, demonstrating reduced mortality with computed tomography (CT) lung cancer screening, more attention has been drawn to CT-based lung pulmonary cancer screening. Early detection of lung cancer, which in most cases implies a lower stage of disease, results in a significantly better prognosis. After the recent Nelson results, it is expected that more countries will adopt lung cancer screening, and as a result, the number of pulmonary CT examinations will increase enormously in the coming years. All those CT examinations will detect a large number of pulmonary nodules, which will most likely result in numerous follow-up CT examinations according to the British Thoracic Society (2015) and the 2017 Fleischner Society guidelines.

Thoracic pulmonary nodule detection is one of the most time-consuming components of the assessment of a chest CT examination; therefore, there could be a role in the pulmonary nodule detection for computer-aided detection (CAD), which could result both in reducing workload but more importantly in improved accuracy.

CAD of pulmonary nodules has been a major field of research in medical image analysis for many years. Since the early publications, there have been a lot of developments in the area of CAD pulmonary nodule detection. In the last couple of years advances in deep-learning neural network techniques have made it possible to create CAD systems with very high sensitivity whilst keeping the number of false positives acceptably low; however, these CAD systems have not been evaluated scientifically in an average hospital in a normal clinical workflow as was undertaken in the present study. Because of this setting, data were also collected on different iterative reconstruction level strengths, kernels, CT systems, and section thickness, which has been published elsewhere.

In Albert Schweitzer Ziekenhuis, Dordrecht, a teaching hospital, has been using a commercially available deep-learning-based CAD (DL-CAD) system integrated into the clinical workflow for approximately 1.5 years. The radiology department in the hospital performs around 9,500 chest CT examinations a year. Therefore, even a minimal improvement in pulmonary nodule detection could have a significant influence on the early detection of possible lung cancer. The purpose of this study was to evaluate the performance of a commercially available DL-CAD system integrated into the clinical workflow of a large teaching hospital. It was hypothesised that the use of the DL-CAD system will result in high sensitivity in the setting of this non-academic hospital.

Materials and methods

Study population

Initially, a retrospective analysis was performed of 145 patients aged from 18 to 91 years comprising 79 men and 66 women (mean age 62 years), who had undergone a chest CT between December 2018 and May 2020. The CT examinations were selected randomly, both contrast-enhanced and unenhanced CT examinations were included; however, CT examinations with more than five nodules were excluded to simplify the annotation of nodule location and to avoid confusing nodules with each other. Reading by a radiologist was performed within 3 months after acquisition so that any potential additional nodules would not have an impact on patient management. CT examinations had been performed for various indications, ranging from ruling out metastases, follow-up of nodules, follow-up of other pulmonary abnormalities, and other miscellaneous indications.

From the acquired dataset, examinations without nodules (97 scans) were used to determine the negative predictive power of the CAD system (both contrast-enhanced and unenhanced CT examinations were included). The negative predictive value of the CAD system is an important clinical parameter as it might bias the radiologist’s decision to conclude that no nodules are present. In addition, the NPV must be very high for use in a screening setting, in order not to miss any cancers.

Acquisition

The CT systems used to acquire the dataset were a 128-section DSCT system Siemens Somatom Flash (95 patients) and a 128-section DSCT system Siemens Somatom Definition Drive (50 patients; Siemens Medical Solutions, Erlangen, Germany). The acquisition and reconstruction settings have been summarised in Table 1. The radiation dose was reported as volume CT dose index (CTDlvol), dose–length product (DLP), and effective dose (ED). For the included patients, the average DLP was 181 mGy cm.

CAD system

The DL-CAD system used was the Veye Chest software by Aidence (Amsterdam, the Netherlands). The technical

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Overview of applied acquisition and reconstruction settings.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acquisition</strong></td>
<td></td>
</tr>
<tr>
<td>Tube voltage (kV)</td>
<td>Variable 80–140</td>
</tr>
<tr>
<td>Section thickness, mm</td>
<td>1 or 3</td>
</tr>
<tr>
<td>Dose modulation</td>
<td>XYZ_EC</td>
</tr>
<tr>
<td>Reconstruction</td>
<td></td>
</tr>
<tr>
<td>Iterative reconstruction algorithm</td>
<td>SAFIRE</td>
</tr>
<tr>
<td>Iterative reconstruction strength</td>
<td>3</td>
</tr>
<tr>
<td>Post-processing filtering</td>
<td>26,40,50</td>
</tr>
</tbody>
</table>
The details of the DL-CAD system have been described previously. The system is a cloud-based application that analyses chest CT images automatically as part of the routine clinical workflow for selected protocols. The images were sent automatically to the cloud-based application based on the name of the imaging protocol. The processed results are sent back and stored in the PACS system (Sectra, Linköping, Sweden) as a separate image stack where they are presented alongside the reconstructed CT images as an overlay that can be switched on or off at will. There is also a separate report, which included nodule size and, if available, growing speed. Image analysis takes approximately 5–10 minutes, but was performed before the images were available to report. When the radiologists analyse the study, the CAD system is used alongside the standard reading as a concurrent reading.

The initial version of the Veye software that was used during this study was model (25-05-2018), which was upgraded to a newer version (18-03-2019) during the study. The system threshold settings for both software model versions were chosen such that they give a similar sensitivity and false-positive rate as specified by the vendor of 87% respectively 1.06 false positives/scan. See Fig 1 for the detection performance of the 18-03-2019 model version. The initial version of the Veye software that was used during this study was model (25-05-2018), which was upgraded to a newer version (18-03-2019) during the study. The system threshold settings for both software model versions were chosen such that they give a similar sensitivity and false-positive rate as specified by the vendor of 87% respectively 1.06 false positives/scan. See Fig 1 for the detection performance of the 18-03-2019 model version.

Radiologist reading

A pulmonary nodule was defined as a focal, rounded intraparenchymal opacity, with the greatest in-plane dimension of ≤3 cm diameter, mostly surrounded by aerated lung, including contact with pleura. Nodules <4 mm were excluded because of the certification requirements of the software and the clinical insignificance for follow-up in low-risk patients. As part of the study, all included CT examinations were subsequently read by a radiologist with 11 years of experience and a resident radiologist, specialising in chest radiology with 4 years of experience. The radiologists were blinded to each other's results, but not to the CAD system nor previously applied annotations. Following a consensus reading, any discrepancies were resolved by a third chest radiologist with 20 years of experience. The resulting set of reported nodules with their compositions and dimensions functions as the reference standard. For each scan, it was recorded which nodules were found by both the Veye software and the radiologists, the false nodules found, and the nodules missed by the software. The maximum diameter in the axial plane (both manually and with Veye), and the composition (solid, subsolid, or a mixture of both) of the nodules were also measured for each nodule. For the false-positive cases, the clinical deviations were described and reported. Further data analysis was done using descriptive statistics. For discrepancies in diameter, the potential reason for this was also reported.

Ethical considerations

This retrospective study was approved by the local institutional review board, which waived the need for informed consent.
Results

Nodule classification

The results and comparison for pulmonary lung nodule detection of the CAD system and the reference standard are summarised in Table 2.

In total, the DL-CAD system found 130 nodules. Of the 130 nodules detected by the software, 80 (62%) were also found to be true positive by the reference standard. Fifty nodules were considered to be false positive by the radiologists. Table 3 shows a breakdown of the false-positive findings. The DL-CAD system missed 11 nodules.

Size

The nodule size was rounded to a measurement in millimetres by the individual radiologists. The average size of the nodules found by the reference standard was 7 mm (SD ± 4.1 mm) and by the DL-CAD system 9 mm (SD ± 7.1 mm). The mean size of the false-positive nodules was 11.8 mm (SD ± 10 mm) and of the false negatives was 6.7 mm (SD ± 6.1 mm).

Comparing the nodule size measurements, the agreement between CAD results and the reference standard was 67% (54/80). For 18 nodules (23%), the discrepancy was 1 mm of which 16 (20%) were measured smaller by the radiologists and two larger, for four nodules it was 2 mm of which two were measured smaller and two larger by the radiologists, and for one nodule, the difference was 4 mm of which the radiologists measured a smaller diameter. For three nodules, the DL-CAD system reported an unknown diameter. Of the size discrepancies, 83% (19/23) were measured smaller by the radiologists. For most of the size discrepancies, the reason is not clear. These were all discrepancies of 1 mm. For three nodules (1, 2, and 4 mm discrepancy) an adjacent artery was also measured by the DL-CAD system, for one (2 mm discrepancy) the measurement was performed on the wrong section, for one (2 mm discrepancy) a subsolid part of the nodule was not measured; for one (1 mm discrepancy) there were surrounding spiculae; and another (2 mm discrepancy) was a cavitating nodule (see Fig 2).

Table 3 Reasons for false-positive results.

<table>
<thead>
<tr>
<th>Reason for false-positive result</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not rounded (flat or triangular)</td>
<td>10 (20%)</td>
</tr>
<tr>
<td>Fibrosis</td>
<td>12 (24%)</td>
</tr>
<tr>
<td>Atelectasis</td>
<td>19 (38%)</td>
</tr>
<tr>
<td>Other *</td>
<td>9 (18%)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>50</td>
</tr>
</tbody>
</table>

* Bronchiectasis, consolidation, pleural not mostly surrounded by lung.

Solid/subsolid

Regarding solid/subsolid classification, 73 nodules in the reference standard were solid, 16 were subsolid, and two were a mixture of both. The agreement on classification between the CAD results and the reference standard was 95%. Two cases were determined solid by the DL-CAD system and subsolid by the radiologists, another two were determined solid by the software and mixed solid/subsolid by the radiologists.

False positives and negatives

As summarised in Table 3, the false-positive cases can be divided into three main categories. Nineteen cases showed considerable atelectasis, 12 cases were found to be fibrosis, and 10 cases were not rounded. The atelectasis and fibrosis also had a non-round shape. The other nine cases were found to be false positive for various reasons, e.g., a gland, bronchiectasis, or a large consolidation. Illustrative nodules are shown in Fig 3. The average size of the false-positive findings was 10 mm, which was larger than the true positives. This is explained by the fact that fibrosis and atelectasis are usually larger aetiologies than most nodules.

Out of the 11 false-negative cases, three were calcified nodules. The radiologists reported nine nodules with a diameter of 4 mm (eight solid including three calcified and one subsolid), two with a diameter of 18 and 20 mm, respectively (both subsolid). Illustrative nodules are shown in Fig 4.

Negative predictive value

The analysis of the 97 cases without any found nodules used the same clinical workflow sensitivity setting of 87% and showed a negative predictive value of 95%. There were five scans with a total of six false-negative nodules. In two of the nodules, there were prominent spiculae in the outskirts of the lesions. One lesion showed extensive surrounding fibrotic tissue. For the other three lesions, there were no specific findings. All lesions were solid.

Discussion

In the present study, the performance of a commercially available DL-CAD system was evaluated in the clinical workflow of a large teaching hospital. The main findings of the study are as follows: the DL-CAD system was found to
have a sensitivity of 88% and a rate of 1.04 false positives/scan (40% of the nodules). Most false positives were due to a non-spherical shape, atelectasis, and fibrosis.

There was a very good agreement of 95% of the composition of the nodules (solid, non-solid, or a mixture). Out of the 80 true positive nodules evaluated, the DL-CAD system disagreed with the reference size measurement in 26 cases. In 19 of these cases, the DL-CAD system overestimated the nodule size by 1 mm or more.

The measured sensitivity in the patient population of 88% with an false-positive rate of 1.04 false positives/scan matches the set sensitivity (87%) and false-positive rate (1.06 false positives/scan) of the DL-CAD system very well. The single system threshold setting of the DL-CAD system for various different uses (e.g., follow-up versus screening) has been found to be a limitation of the software. It would be a useful improvement to be able to set different sensitivity and false-positive rate thresholds for different clinical uses.

Low false-positive rates are often named as an important boundary condition in perfecting pulmonary nodule CAD systems. The rate of false positives in the present study was as high as 40%. False positives were due to atelectasis, fibrosis, and non-rounded shapes. Because the cases with atelectasis and fibrosis also had non-rounded shapes, most of the false positives could be resolved by applying stricter selection criteria to the nodule shape. This is a known issue when addressing false-positive reduction. With a measured false-positive rate of 1.04 false positives/scan the pros and cons seem to be well balanced. Still, further improvements in reducing the false positives are necessary, including the abovementioned stricter shape criteria. In addition, the DL-CAD system will improve due to the combination of continued input and deep learning.

![Figure 2](image-url)  
Figure 2: Size discrepancies with in (a) an adjacent vascular structure is implemented in the measurement leading to an overestimation of the size, in (b) a partly subsolid nodule, the subsolid part was not measured by the DL-CAD system resulting in an underestimation of the size. (c) Nodule size overestimation by 1 mm by the CAD system with no obvious reason.
Eleven false-negative nodules were found in the initial dataset of which three were calcified. Calcified nodules, however, have a very low risk of being malignant, so one could argue that these cases are not relevant when missed by the software.

The reference standard, which comprised the consensus of three radiologists, was not blinded to the CAD system, to approach the best reflection of the true nodules. This could have had a bias for nodules that were pointed out by the CAD system, which in a blinded situation would not have

Figure 3 False-positive cases. (a,b) Fibrosis in the top of the lung as well as a non-rounded shape; (c,d) the DL-CAD system highlighted atelectasis; (d) a sagittal reconstruction to illustrate the typical shape of atelectasis; and (e,f) the DL-CAD system marked a subsolid nodule.
been pointed out as nodules; however, this influence was of minimal significance, as there was not a lot of doubt in deciding whether a nodule was a true nodule or not. Furthermore, because of this setting, a true comparison of the CAD system with radiologists in contemporary practice was not possible.

An important finding was the discrepancy of the measured nodule size in which the program often over-estimated the measurement of the radiologist for no obvious reason. In some cases, the overestimation had a clear reason, such as the measurement of an adjacent artery of measurement on a wrong section. Some cases could possibly be explained by rounding errors, but this could have an important impact on patient management and follow-up. Particularly in incidentally found nodules, because the minimum diameter for follow-up in the latest Fleischner criteria is 5 mm (average diameter) and 4 mm (maximum diameter) for the British Thoracic society.\textsuperscript{5,6} One could argue that the size manually measured by the radiologist is less accurate than that measured by a software program; however, most of the trials on which the recommendations for follow-up are made, are based on manual measurements.\textsuperscript{3,21,22} So even if the manual measurement is not accurate, follow-up regimens are based on this method.

Nodules of <4 mm were excluded from the study, because of the relative clinical insignificance in this setting and the minimal size threshold specified by the manufacturer; however, in a clinical non-screening setting these nodules could be relevant, especially in patients with a high risk of developing metastases.

Five false-negative cases were found in the subgroup without found nodules leading to a negative predictive value of 95%, which is probably not high enough for use in a screening setting. It is to be expected, however, that the negative predictive value will increase (at the cost of a higher false-positive rate) when higher sensitivity settings are used.

No changes were made to the clinical reconstruction protocol for the evaluation of the CT examinations by the DL-CAD system, and as a result, it cannot be confirmed that the software is performing in an optimal way. In addition, data were collected on different reconstruction parameters, such as section thickness, post-reconstruction filtering, and iterative reconstruction strength settings, which have been

---

**Figure 4** False-negative cases. (a) A rounded nodule of 4 mm; (b) a subsolid nodule of 20 mm; (c) a calcified nodule of 4 mm; and (d) a rounded nodule of 4 mm.
The reproducibility of measurements between different CT dose levels has been investigated previously. In conclusion, good sensitivity (88%) and a high negative predictive value of 95% was found when using the DL-CAD system in clinical practice. There was a high rate of false positives (40%), which has room for improvement, especially if flat nodules could be filtered out. There was a surprising overestimation of nodule size by the CAD system. Further studies should focus on the benefit of nodule detection of the CAD system in contemporary radiological practice. In addition, a repeat study with a larger sample size would be of value.

**Conflict of interest**

The authors declare no conflict of interest.

**References**

1. KNL. Integraal kankercentrum Nederland/longkanker. Available at: https://www.knl.nl/kankersoorten/longkanker. [Accessed 8 December 2019].