CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

| | |
|---|---|
| Title | Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database |
| Authors | Jacobs, Colin;van Rikxoort, Eva M.;Murphy, Keelin;Prokop, Mathias;Schaefer-Prokop, Cornelia M.;van Ginneken, Bram |
| Publication date | 2015-10-06 |
| Original Citation | Jacobs, C., van Rikxoort, E. M., Murphy, K., Prokop, M., Schaefer-Prokop, C. M. and van Ginneken, B. (2016) 'Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database', European Radiology, 26(7), pp. 2139-2147. doi: 10.1007/s00330-015-4030-7 |
| Type of publication | Article (peer-reviewed) |
| Link to publisher's version | 10.1007/s00330-015-4030-7 |
| Rights | © 2915, The Author(s). This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. - http://creativecommons.org/licenses/by-nc/4.0/ |
| Download date | 2026-09-09 16:47:26 |
| Item downloaded from | https://hdl.handle.net/10468/9248 |

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Eur Radiol (2016) 26:2139–2147
DOI 10.1007/s00330-015-4030-7

COMPUTER APPLICATIONS

# Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database

**Colin Jacobs**[1] · **Eva M. van Rikxoort**[1,2] · **Keelin Murphy**[4] · **Mathias Prokop**[1] · **Cornelia M. Schaefer-Prokop**[1,3] · **Bram van Ginneken**[1,2]

Received: 12 June 2015 / Revised: 20 July 2015 / Accepted: 14 September 2015 / Published online: 6 October 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

**Abstract**

*Objectives* To benchmark the performance of state-of-the-art computer-aided detection (CAD) of pulmonary nodules using the largest publicly available annotated CT database (LIDC/IDRI), and to show that CAD finds lesions not identified by the LIDC's four-fold double reading process.

*Methods* The LIDC/IDRI database contains 888 thoracic CT scans with a section thickness of 2.5 mm or lower. We report performance of two commercial and one academic CAD system. The influence of presence of contrast, section thickness, and reconstruction kernel on CAD performance was assessed. Four radiologists independently analyzed the false positive CAD marks of the best CAD system.

*Results* The updated commercial CAD system showed the best performance with a sensitivity of 82 % at an average of 3.1 false positive detections per scan. Forty-five false positive CAD marks were scored as nodules by all four radiologists in our study.

*Conclusions* On the largest publicly available reference database for lung nodule detection in chest CT, the updated commercial CAD system locates the vast majority of pulmonary nodules at a low false positive rate. Potential for CAD is substantiated by the fact that it identifies pulmonary nodules that were not marked during the extensive four-fold LIDC annotation process.

*Key Points*

- *CAD systems should be validated on public, heterogeneous databases.*
- *The LIDC/IDRI database is an excellent database for benchmarking nodule CAD.*
- *CAD can identify the majority of pulmonary nodules at a low false positive rate.*
- *CAD can identify nodules missed by an extensive two-stage annotation process.*

**Keywords** Computer-assisted diagnosis · Image interpretation, computer-assisted · Lung cancer · Solitary pulmonary nodule · Lung

**Abbreviations**

| | |
|---|---|
| CAD | Computer-aided detection |
| CT | Computed tomography |
| LIDC | Lung image database consortium |
| IDRI | Image database resource initiative |

## Introduction

The last two decades have shown substantial research into computer-aided detection (CAD) of pulmonary nodules in thoracic computed tomography (CT) scans [1, 2]. Although many academic and several commercial CAD algorithms have been developed, CAD for lung nodules is still not commonly used in daily clinical practice. Possible explanations for this are a lack of reimbursement, technical impediments to integration into PACS systems, but also low sensitivity and high false positive rates. The recent positive results of the

✉ Colin Jacobs
colin.jacobs@radboudumc.nl

[1] Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands

[2] Fraunhofer MEVIS, Bremen, Germany

[3] Department of Radiology, Meander Medical Center, Amersfoort, The Netherlands

[4] Irish Centre for Fetal and Neonatal Translational Research, University College Cork, Cork, Ireland

NLST lung cancer screening trial [3] and the subsequent developments towards implementation of lung cancer screening in the United States [4, 5] have renewed the interest into CAD for pulmonary nodules. If lung cancer screening will be implemented on a large scale, the burden on radiologists will be substantial and CAD could play an important role in reducing reading time and thereby improving cost-effectiveness [6, 7].

Following the general demand for open and reproducible science, public databases have been established to facilitate objective measures of CAD performance, and to move CAD development to a next level [8–10]. In 2011, the complete LIDC/IDRI (Lung Image Database Consortium / Image Database Resource Initiative) database was released [10]. This dataset provides by far the largest public resource to assess the performance of algorithms for the detection of pulmonary nodules in thoracic CT scans. A large effort has gone into the collection of annotations on these cases, but CAD was not used to assist the readers [10].

In this paper, we apply two commercial and one state-of-the-art academic nodule detection systems on the LIDC/IDRI database with the aim to set a first benchmark performance on the full database. To our knowledge, this is the first paper, which reports the performance of CAD systems on the full LIDC/IDRI database. We performed an extensive analysis of the performance of the applied CAD systems and make our evaluation publicly available so that other CAD developers can compare with this benchmark. Furthermore, we hypothesize that CAD can find lesions, which were not detected in the extensive LIDC annotation process consisting of a blinded and unblinded review by four radiologists. To investigate the latter, we evaluated the false positives of the best CAD system using a similar reading protocol as had been used in LIDC.

## Materials and methods

### Data

This study used the LIDC/IDRI data set [10], consisting of 1,018 helical thoracic CT scans collected retrospectively from seven academic centres. Nine cases with inconsistent slice spacing or missing slices were excluded. In addition, 121 CT scans, which had a section thickness of 3 mm and higher, were excluded since thick section data is not optimal for CAD analysis. This resulted in 888 CT cases available for evaluation. In Tables 1, 2, and 3, the characteristics of the input data are shown.

### LIDC/IDRI image annotation

The LIDC/IDRI employed a two-phase image annotation process [10]. In the first phase (the blind phase), four radiologists independently reviewed all cases. In the second phase (the unblinded phase), all annotations of the other three radiologists were made available and each radiologist independently reviewed their marks along with the anonymized marks of their colleagues. Findings were annotated and categorized into *nodule≥3 mm*, *nodule<3 mm*, or *non-nodule. Non-nodule* marks were used to indicate abnormalities in the scan, which were not considered a nodule. Using this two-phase process, the LIDC investigators aimed to identify as completely as possible all lung nodules, without forcing consensus among the readers. More details about the annotation process can be found in [10]. An XML file with the annotations is publicly available for every case.

**Table 1** Manufacturer and scanner model distribution of the 888 CT scans in our dataset

| Manufacturer | Model name | Number |
|---|---|---|
| GE MEDICAL SYSTEMS | LightSpeed16 | 197 |
| GE MEDICAL SYSTEMS | LightSpeed Ultra | 162 |
| GE MEDICAL SYSTEMS | LightSpeed QX/i | 97 |
| GE MEDICAL SYSTEMS | LightSpeed Pro 16 | 79 |
| GE MEDICAL SYSTEMS | LightSpeed VCT | 61 |
| GE MEDICAL SYSTEMS | LightSpeed Plus | 56 |
| GE MEDICAL SYSTEMS | LightSpeed Power | 10 |
| Philips | Brilliance 16P | 54 |
| Philips | Brilliance 64 | 49 |
| Philips | Brilliance 40 | 9 |
| Philips | Brilliance16 | 5 |
| SIEMENS | Sensation 16 | 95 |
| SIEMENS | Sensation 64 | 5 |
| SIEMENS | Definition | 3 |
| SIEMENS | Emotion 6 | 1 |
| TOSHIBA | Aquilion | 5 |
| Total | | 888 |

### Nodule selection and purpose

In this study, we included all annotations available in the XML files for the 888 scans. The focus of this study was on the

**Table 2** Section thickness distribution of the 888 CT scans in our dataset

| Section thickness | Number |
|---|---|
| 0.6 | 7 |
| 0.75 | 30 |
| 0.9 | 2 |
| 1 | 58 |
| 1.25 | 343 |
| 1.5 | 5 |
| 2 | 123 |
| 2.5 | 320 |
| Total | 888 |

**Table 3** Distribution of the reconstruction kernels used for the 888 CT scans in our dataset

| Manufacturer and reconstruction kernel | Type | Number |
|---|---|---|
| GE MEDICAL SYSTEMS - BONE | Enhancing | 220 |
| GE MEDICAL SYSTEMS - LUNG | Overenhancing | 70 |
| GE MEDICAL SYSTEMS - STANDARD | Standard | 372 |
| Philips - B | Standard | 21 |
| Philips - C | Enhancing | 7 |
| Philips - D | Overenhancing | 45 |
| SIEMENS - B20s | Soft | 1 |
| SIEMENS - B30f | Standard | 102 |
| SIEMENS - B31f | Standard | 1 |
| SIEMENS - B45f | Enhancing | 30 |
| SIEMENS - B50f | Enhancing | 2 |
| SIEMENS - B70f | Overenhancing | 12 |
| TOSHIBA - FC03 | Standard | 2 |
| TOSHIBA - FC10 | Soft | 3 |
| Total | | 888 |

*nodule≥3 mm* group. As a result of the LIDC/IDRI image annotation process, each *nodule≥3 mm* had been annotated by one, two, three, or four radiologists. In total, the data set of this study included 777 locations, which were marked as *nodule≥3 mm* by all four radiologists. The 777 *nodule≥3 mm* annotations marked by all four radiologists can be categorized by size as follows: 22 nodules <4 mm, 228 nodules 4–6 mm, 199 nodules 6–8 mm, and 328 nodules >8 mm. The number of nodules per scan ranged between 1 and 8.

The purpose of this study was twofold. First, we aimed to assess the performance of three state-of-the-art nodule CAD systems. Secondly, we performed an observer experiment to investigate whether CAD can find additional lesions, missed during the extensive LIDC annotation process.

## CAD systems

Three CAD systems were used: a commercial CAD system *Visia* (MeVis Medical Solutions AG, Bremen, Germany), a commercial prototype CAD system *Herakles* (MeVis Medical Solutions AG, Bremen, Germany), and an academic nodule CAD system *ISICAD* (Utrecht Medical Center, Utrecht, the Netherlands) [11]. *ISICAD* was the leading academic CAD system in the ANODE09 nodule detection challenge [9]. For all three CAD systems, a list of candidate marks per CT scan was obtained. Each CAD candidate is described by a 3D location. Additionally, *Herakles* and *ISICAD* also provide a CAD score per CAD candidate. The CAD score is the output of the internal classification scheme of the CAD system and is a measure of the likelihood that a candidate is a pulmonary nodule. An internal threshold on the CAD scores determines which candidates are active CAD marks and, hence, will be shown to the user, and which candidates are not shown. Since different thresholds can be applied on the CAD score, a CAD system can have multiple operating points. A low threshold generates more CAD marks, thereby typically increasing sensitivity at the cost of more false positive detections. A high threshold will generate less false positives but may reduce the sensitivity of a CAD system. For all three CAD systems, one fixed operating point is internally set which we will refer to as the system operating point.

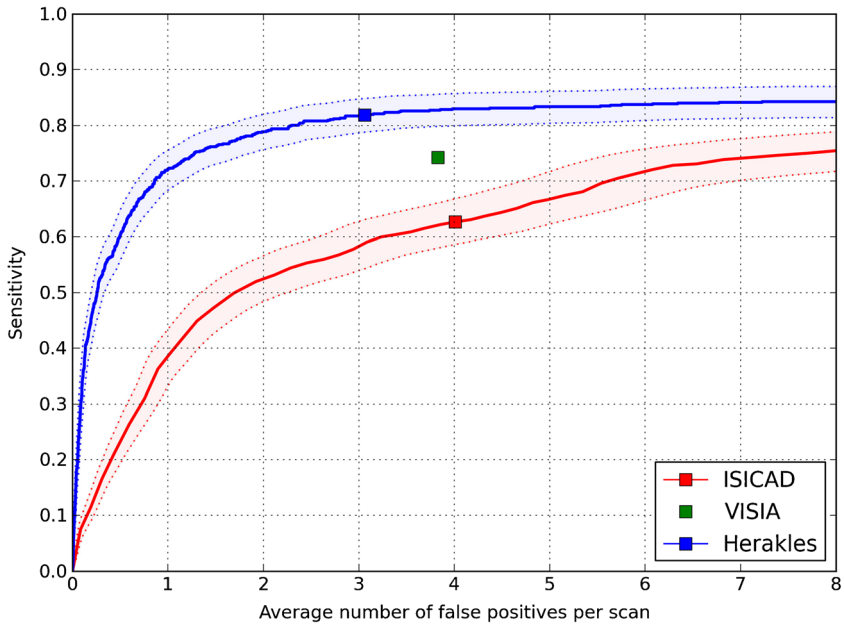


**Fig. 1** FROC curves for all three CAD systems on the full database of 888 CT scans containing 777 nodules for which all four radiologists classified it as *nodule≥3 mm*. The points on the curves indicate the system operating points of the three CAD systems. For *Visia,* no continuous FROC curve but only a single operating point can be provided since the CAD scores of the CAD marks are not available. Shaded areas around the curve indicate 95 % confidence intervals

## Evaluation

The performance of the CAD systems was analyzed on the set of 777 nodules annotated by 4/4 radiologists as a *nodule≥3 mm.* We employed free-response operating characteristic (FROC) analysis [12] where detection sensitivity is plotted against the average number of false positive detections per scan. Confidence intervals were estimated using bootstrapping with 5,000 iterations [13]. If a CAD system marked locations which were annotated by three or fewer radiologists as *nodule≥3 mm*, as *nodule< 3 mm*, and as *non-nodules*, these CAD marks were counted as false positives. For *Visia*, no CAD scores were available for the CAD candidates. Consequently, only one operating point and not a full FROC curve could be generated for *Visia*.

To gain more insight into which type of nodules were missed by CAD, we looked at the characteristics, as scored by the LIDC readers, for all *nodule≥3 mm* findings, of the false negatives. We defined subsolid nodules as nodules for which the majority of the radiologists gave a texture score smaller than 5 (1=ground-glass/non-solid, 3=part-solid, 5=solid). Subtle nodules were defined as nodules for which the majority of the radiologists gave a subtlety score smaller or equal than 3 (1=extremely subtle, 5=obvious).

To assess the robustness of the CAD algorithms, we also evaluated the CAD results on different subsets of the data. The LIDC-IDRI data set is a heterogeneous set of CT scans and CAD algorithms that could conceivably exhibit a different performance on different types of data. We analyzed the following factors: (1) presence of contrast material, i.e., non-

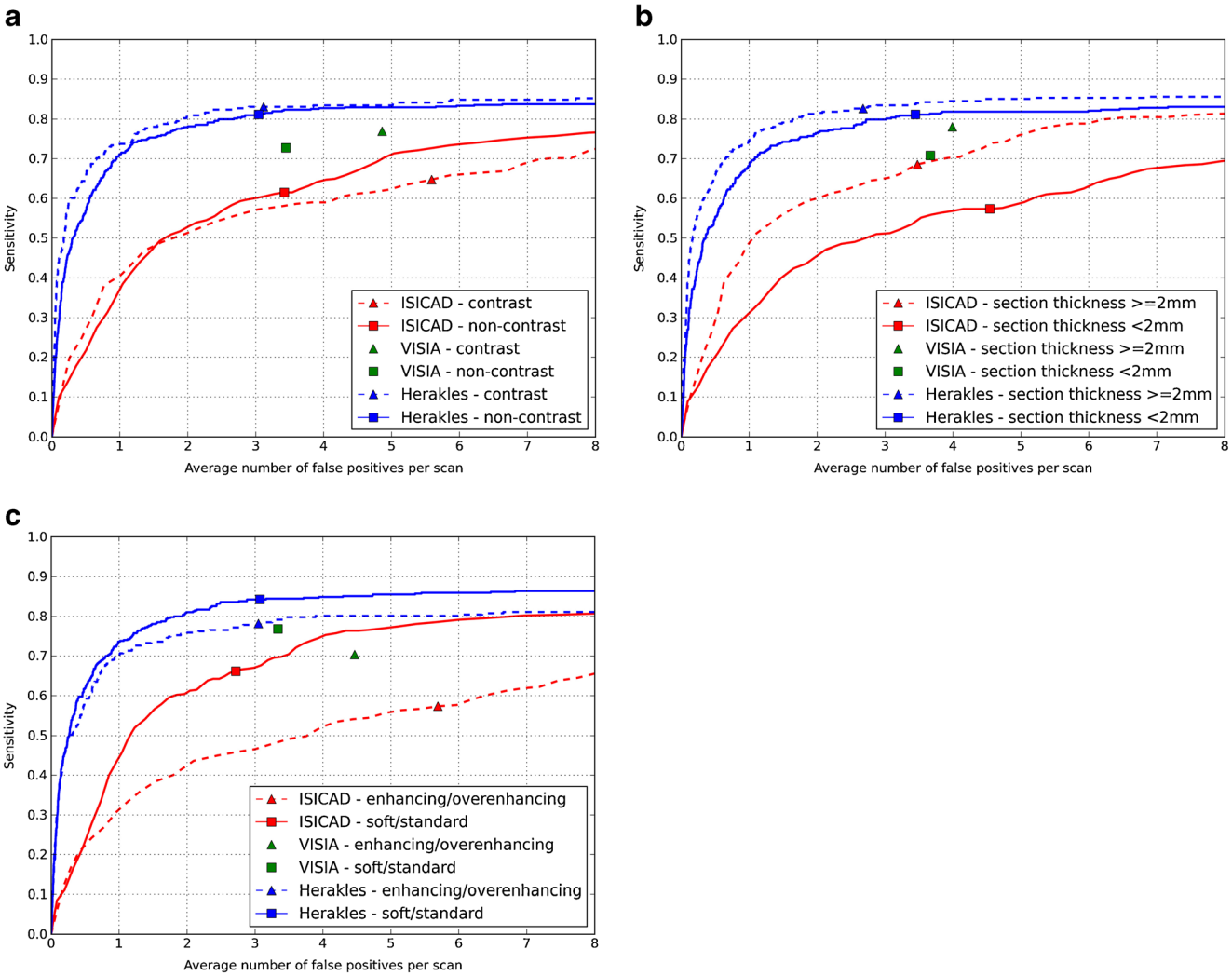


**Fig. 2** FROC curves for all three CAD systems on (**a**) contrast scans (*n*=242) versus non-contrast scans (*n*=646), (**b**) scans with a section thickness <2 mm (*n*=445) versus scans with a section thickness ≥2 mm (*n*=443), and (**c**) scans with a soft or standard reconstruction kernel (*n*=502) versus scans with an enhancing or overenhancing reconstruction kernel (*n*=386). The reference set of nodules consists of nodules for which all four radiologists classified it as *nodule≥3 mm*. The points on the curves indicate the system operating points of the three systems. For *Visia,* no continuous FROC curve but only a single operating point can be provided since the CAD scores of the CAD marks are not available

contrast versus contrast enhanced scans, (2) section thickness, i.e., cases with section thickness <2 mm versus section thickness ≥2 mm, and (3) reconstruction kernel, i.e., soft/standard versus enhancing/overenhancing kernels.

### Observer study

In order to evaluate whether CAD can find lesions missed during the extensive annotation process of the LIDC/IDRI database, we considered the CAD marks of the best CAD algorithm , which were counted as false positives at its system operating point. Two conditions were differentiated: the location of the CAD mark had in fact been marked in the LIDC annotation process, but not by all four readers as *nodule≥3 mm* as warranted for being counted as a true positive. The second condition comprised those CAD marks that had no corresponding LIDC marks at all. The CAD marks corresponding to the first condition can be subdivided according to the LIDC readings. The latter CAD marks were independently inspected by four chest radiologists, since these are potentially nodules overlooked by all four LIDC readers. Thus, we mimic the original LIDC annotation process as though CAD had been included as another independent reader in the first phase of the image annotation process. CAD marks were categorized as *nodule≥3 mm*, *nodule<3 mm*, *non-nodules*, or false positive. Electronic measurement tools were available to measure size. To reduce the workload for the radiologists, a research scientist (5 years experience in nodule CAD research) first removed the marks which were obviously not a nodule. CAD marks which were marked as nodule>3 mm by all four radiologists in our study were independently evaluated by an experienced radiologist that scored subtlety, location, type, and attachment to other structures. Subtlety was scored on a five-point scale (1=extremely subtle, 5=obvious).

## Results

### Comparative CAD performance

The performance of the three CAD systems is depicted in Fig. 1. From the FROC curves it is evident that *Herakles* performed best. The system performances were significantly different ($p<0.001$). At its system operating point, *Herakles* reached a sensitivity of 82 % at an average of 3.1 false positives per scan for nodules all four LIDC readers had agreed on.

We evaluated the characteristics of the 141 false negative nodules. 42 (30 %) false negatives were subsolid nodules. The size distribution of the missed nodules was as follows: five nodules <4 mm, 53 nodules 4–6 mm, 31 nodules 6–8 mm, and 52 nodules >8 mm. Thus, a large portion of the missed nodules were smaller than 6 mm, but still a substantial number of missed nodules, 52 (37 %), were larger than 8 mm. Finally, we found that 33 (23 %) of the missed nodules were subtle. Figure 3 shows eight randomly chosen missed nodules.

The performance of the three CAD systems on the different subsets is depicted in Fig. 2. This figure shows that the performance of *ISICAD* and *Visia* was influenced by different data sources. *ISICAD* shows the largest performance difference between soft/standard versus enhancing/overenhancing reconstruction kernels. *Herakles* showed the most stable and

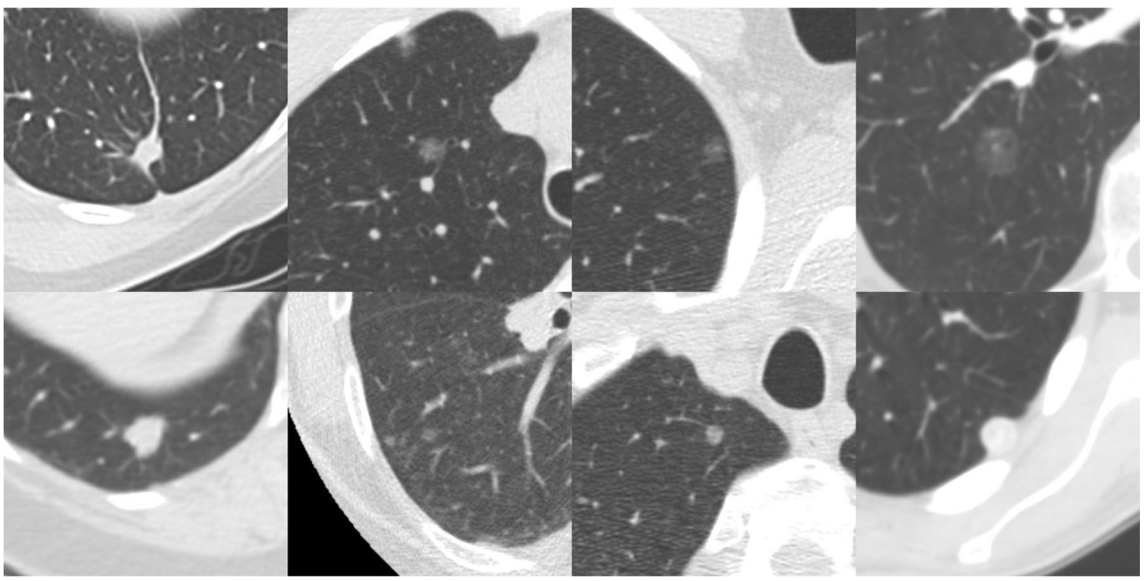

**Fig. 3** Eight randomly chosen examples of false negatives of *Herakles*. Each image shows a transverse field of view of 60 x 60 mm in which the nodule is centred. Note that many missed nodules are subsolid

robust performance for all different data sources and consistently outperformed the other two CAD systems.

We categorized the CAD marks of *Herakles,* which were counted as false positives at its system operating point. In total, there were 2,720 false positive CAD marks in the 888 cases (Table 4). The majority of the CAD marks, 1,612 out of 2,720 (59 %), had at least one corresponding mark from the LIDC readers. These CAD marks can be further categorized into marks on annotations marked as *nodule≥3 mm* by three out of four radiologists, two out of four radiologists, one out of four radiologists; and annotations marked as *nodule<3 mm* by at least one radiologist (and, hence, no *nodule≥3 mm* annotations); and finally annotations marked as *non-nodule* by at least one out of four radiologists (and, hence, no *nodule≥3 mm* or *nodule<3 mm* annotations). Table 4 shows how the CAD marks were further split out into these categories. The remaining 1,108 false positive CAD marks had no corresponding mark from the LIDC readers.

**Table 4** Overview of the categories in which the false positives of *Herakles* at the system operating point can be divided. In this analysis, we first check for corresponding *nodule≥3 mm* annotations, then we check for corresponding *nodule<3 mm* annotations, and finally we check for corresponding *non-nodule* annotations. This means that in the top row where three out of four radiologists annotated the location as *nodule≥3 mm*, the fourth radiologist may have marked the location as *nodule<3 mm, non-nodule,* or did not mark it at all. In the *nodule<3 mm* category, all false positives whose location was marked as *nodule<3 mm* by at least one radiologist were placed (and, hence, no radiologist marked it as *nodule≥3 mm*). The non-nodule category contains all false positives whose location was marked as *non-nodule* by at least one radiologist (and, hence, no radiologist marked the location as *nodule≥3 mm* or *nodule<3 mm*). False positives for which no corresponding annotation was found were assigned to the last category

| Category | Number |
|---|---|
| Nodule≥3 mm - 3/4 | 254 |
| Nodule≥3 mm - 2/4 | 208 |
| Nodule≥3 mm - 1/4 | 219 |
| Nodule<3 mm | 423 |
| Non-nodule | 508 |
| No corresponding annotation | 1108 |
| Total | 2720 |

## Observer study results

In our observer experiment, we focused on these 1,108 false positive CAD marks of *Herakles,* which had no corresponding mark from any of the LIDC readers. These are locations, which were potentially overlooked by all four LIDC readers. After CAD marks, which were obviously not a nodule had been removed by the research scientist, 269 CAD marks were left for analysis by the four radiologists. Common sources of false positive detections removed by the research scientist included fissure thickening at the chest wall, vessel bifurcations and (micro-)atelectasis. Table 5 depicts how each of the observers scored these 269 CAD marks. In total, 45 CAD marks were considered to be a *nodule≥3 mm* by all four radiologists; 177 CAD marks were considered to be a *nodule≥3 mm* by at least one of the radiologists. The size distribution of the 45 CAD marks was as follows: nine nodules <4 mm, 27 nodules 4-6 mm, seven nodules 6-8 mm, and two nodules >8 mm. Subtlety was scored lower or equal than 3 for 32 (71 %) nodules. Location was scored as central for 11 nodules, peripheral for 11 nodules, and in-between for 23 nodules. Nodule type was scored as follows: 32 solid, 2 ground-glass, 1 part-solid, and 10 calcified. Vascular, pleural or fissural attachment was found for 18 (40 %) nodules. Figure 4 shows eight randomly chosen examples of CAD marks, which were considered a *nodule≥3 mm* by all four radiologists and were scored as solid. In addition, 169 marks were considered a *nodule≥ 3 mm* or a *nodule<3 mm* by all four radiologists; 250 marks were considered a *nodule≥3 mm* or a *nodule<3 mm* by at least one of the radiologists. Thus, following the reference of the 4-reader agreement and adding these 45 CAD marks to the set of nodules, the updated performance of *Herakles* at its system operating point would reach a sensitivity of 83 % at an average of 3.0 false positive detections per scan. In this FROC analysis, CAD marks on locations marked as *nodule≥3 mm* by three out of four radiologists, two out of four radiologist, one out of four radiologists, or as *nodule<3 mm* by at least one radiologist were counted as false positives. Evidently, one could argue whether CAD marks on these locations should be counted as false positives or not. If CAD marks on these locations were not to be counted as false positives but ignored in the FROC analysis, a performance of 83 % sensitivity at an average of only 1.0 false positives per scan would be reached.

## Discussion

Though clear definitions are available for what represents a pulmonary nodule (Fleischner Glossary [14]), the literature lists a number of publications demonstrating the lack of observer agreement of what indeed represents a pulmonary nodule [15–17]. Not surprisingly this effect is larger for small

**Table 5** Results of the observer experiment. The distribution of the scores of all observers is tabulated

| Type | Observer 1 | Observer 2 | Observer 3 | Observer 4 |
|---|---|---|---|---|
| Nodule≥3 mm | 119 | 97 | 84 | 153 |
| Nodule<3 mm | 125 | 141 | 136 | 50 |
| Non-nodule | 20 | 20 | 46 | 41 |
| False positive | 5 | 11 | 3 | 25 |
| Total | 269 | 269 | 269 | 269 |

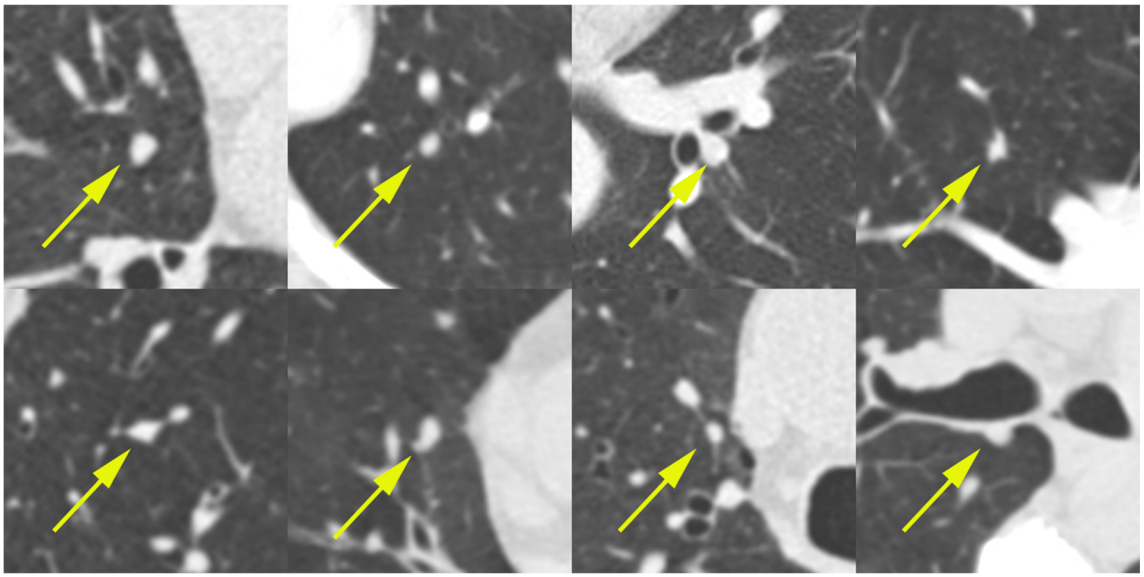

**Fig. 4** Eight randomly chosen examples of solid nodule annotations marked as *nodule≥3 mm* by all four readers in our observer experiment. These nodules were not annotated by any of the original LIDC readers. Each image shows a transverse field of view of 60 x 60 mm in which the nodule is centred

lesions [15]. This lack of an absolute standard of truth makes benchmarking of CAD systems very difficult. Therefore, we decided to use the largest publicly available database of CT annotated pulmonary nodules. An elaborate double reading process involving four radiologists had been undertaken to define various levels of evidence for the presence of nodules to avoid the need for a consensus statement. In our study we used the extensive annotation information of the LIDC/IDRI database to benchmark the performance of state-of-the-art nodule CAD systems. To our knowledge, this is the first study that uses the full LIDC database and secondly accepts the fact that there is no absolute standard of truth for the presence of pulmonary nodules in the absence of pathological correlation.

Our study showed substantial performance differences between the three CAD systems, with the commercial prototype *Herakles* demonstrating the best performance. At its system operating point, *Herakles* detected 82 % of all *nodule≥3 mm* findings marked by all four LIDC readers at an average of 3.1 false positives per scan. If marks on the other LIDC annotations were ignored in the analysis, a sensitivity of 83 % at an average of only 1.0 false positives was reached.

The best CAD system still misses a subset of the nodules (18 % of the 777 nodules). We observed that a substantial part of the missed nodules (30 %) were subsolid nodules, which are more rare and have a markedly different internal character than solid nodules. Therefore, integrating a dedicated subsolid nodule detection scheme [18] in a complete CAD solution for pulmonary nodules may prove helpful to improve overall CAD performance.

Both *Visia* and *ISICAD* showed substantial performance differences on different subsets of the data, but *Herakles* achieved a more robust performance. The performance of *ISICAD* dropped substantially on data with enhancing or overenhancing reconstruction kernels. This may be attributed to the fact that *ISICAD* was developed and trained exclusively with data from the Dutch-Belgian NELSON lung cancer screening trial, which consists of homogeneous thin-slice data reconstructed with a soft/standard reconstruction kernel [19]. This indicates that although *ISICAD* was the leading CAD system for the data used in the ANODE challenge [9], which consisted only of data obtained from the NELSON trial, its performance drops when applied to data of other sources. Therefore, the heterogeneity of a reference database is an important aspect for a reliable CAD evaluation and an advantage of the LIDC/IDRI database.

Although a blinded and unblinded review of all images had been performed by the LIDC investigators, we showed that CAD can find lesions missed by the original LIDC readers. We found 45 nodules, which were accepted as a *nodule≥3 mm* by all four radiologists involved in our observer study. Previous studies have already shown that CAD can find lesions missed by multiple readers [18, 20]. One possible reason why the LIDC readers missed nodules may be that the LIDC readers only inspected transverse sections [10]. Characteristic features of the 45 nodules not included in the LIDC/IDRI database but seen by CAD were subtle conspicuity, small size (<6 mm), and attachment to pleura or vasculature.

Since an extensive evaluation on a large reference database is essential to move CAD to the next level, we have published our results on a public website (http://luna.grand-challenge.org/) which allows other CAD researchers to upload results of their CAD systems for which the same FROC curves as presented in Figs. 1 and 2 will be computed and published on the website. The annotation files of the reference standard and

the extra annotations by the human readers in our observer study are available for download. By making the extra annotations available to other researchers, this study contributes to an improved reference standard for the LIDC/IDRI database, and we hope future CAD studies will use the improved reference standard.

We primarily evaluated the performance of CAD on nodules for which all four radiologists agreed that it was a *nodule≥3 mm*. Previous publications have also focused on the nodules detected by three, two, or one out of four radiologists [21, 22]. For using CAD in a screening setting, a high sensitivity even at the expense of specificity is desirable to find all potential cancerous nodules. High false positive rates, on the other hand, increase the workload to radiologists and potentially increase unnecessary follow-up. We, therefore, report the sensitivity using the highest level of evidence (four out of four readers) and considered the lower levels of agreement for quantifying the false positive rates. For future CAD reference databases, a large database of CT images including follow-up CT and histopathological correlation would be helpful to remove subjectivity from the reference standard, and to verify whether CAD detects the clinically relevant nodules.

In conclusion, we found that, on the largest publicly available database of annotated chest CT scans for lung nodule detection, *Herakles* detects the vast majority of pulmonary nodules at a low false positive rate. The results show that the new prototype outperforms the other two CAD systems and is robust to different acquisition factors, such as presence of contrast, section thickness, and reconstruction kernel. Our observer experiment showed that *Herakles* was able show to pulmonary nodules that had been missed by the extensive LIDC annotation process. Given the growing interest and need for CAD in the context of screening, it can be expected that new CAD algorithms will be presented in the near future. Our results are publicly available and other CAD researchers may compare the performance of their CAD algorithm to the results reported here, utilizing the LIDC/IDRI database for benchmarking of available CAD systems.

**Acknowledgements** The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

The scientific guarantor of this publication is Bram van Ginneken. The authors of this manuscript declare relationships with the following companies: MeVis Medical Solutions AG, Bremen, Germany

This study has received funding by a research grant from MeVis Medical Solutions AG, Bremen, Germany and by a research grant from the Netherlands Organisation for Scientific Research (NWO), project number 639.023.207. No complex statistical methods were necessary for this paper. Institutional Review Board approval was obtained. Written informed consent was waived by the Institutional Review Board.

Not applicable since no animals were involved in this study. Some study subjects or cohorts have been previously reported in previous studies involving the LIDC/IDRI database. The following publication describes the complete LIDC/IDRI database:

Armato SG, McLennan G, Bidaut L, et al. (2011) The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38: 915–931

Methodology: retrospective, experimental, multicenter study.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Li Q (2007) Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Comput Med Imaging Graph 31:248–257
2. Lee SLA, Kouzani AZ, Hu EJ (2012) Automated detection of lung nodules in computed tomography images: a review. Mach Vis Appl 23:151–163
3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365: 395–409
4. de Koning HJ, Meza R, Plevritis SK, Ten Haaf K, Munshi VN, Jeon J et al (2014) Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the U.S. Preventive Services Task Force. Ann Intern Med 160:311–320
5. Black WC, Gareen IF, Soneji SS, Sicks JD, Keeler EB, Aberle DR et al (2014) Cost-effectiveness of CT screening in the National Lung Screening Trial. N Engl J Med 371:1793–1802
6. Aberle DR, Henschke CI, McLoud TC, Boiselle PM (2012) Expert opinion: barriers to CT screening for lung cancer. J Thorac Imaging 27:208
7. Prokop M (2014) Lung cancer screening: the radiologist's perspective. Semin Respir Crit Care Med 35:91–98
8. Armato SG, McLennan G, McNitt-Gray MF, Meyer CR, Yankelevitz D, Aberle DR et al (2004) Lung image database consortium: developing a resource for the medical imaging research community. Radiology 232:739–748
9. van Ginneken B, Armato SG, de Hoop B, van de Vorst S, Duindam T, Niemeijer M et al (2010) Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. Med Image Anal 14: 707–722
10. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP et al (2011) The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38: 915–931
11. Murphy K, van Ginneken B, Schilham AMR, de Hoop BJ, Gietema HA, Prokop M (2009) A large scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification. Med Image Anal 13:757–770
12. Chakraborty DP, Berbaum KS (2004) Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 31:2313–2330
13. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–29

14. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J (2008) Fleischner society: glossary of terms for thoracic imaging. Radiology 246:697–722
15. Armato SG, Roberts RY, Kocherginsky M, Aberle DR, Kazerooni EA, Macmahon H et al (2009) Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". Acad Radiol 16:28–38
16. Armato SG, McNitt-Gray MF, Reeves AP, Meyer CR, McLennan G, Aberle DR et al (2007) The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. Acad Radiol 14:1409–1421
17. Rubin GD, Lyo JK, Paik DS, Sherbondy AJ, Chow LC, Leung AN et al (2005) Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. Radiology 234:274–283
18. Jacobs C, van Rikxoort EM, Twellmann T, Scholten ET, de Jong PA, Kuhnigk JM et al (2014) Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. Med Image Anal 18:374–384
19. van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K et al (2009) Management of lung nodules detected by volume CT scanning. N Engl J Med 361:2221–2222
20. Drew T, Vo ML, Olwal A, Jacobson F, Seltzer SE, Wolfe JM (2013) Scanners and drillers: characterizing expert visual search through volumetric images. J Vis 13:1–13
21. Camarlinghi N (2013) Automatic detection of lung nodules in computed tomography images: training and validation of algorithms using public research databases. Eur Phys J Plus 128:1–21
22. Tan M, Deklerck R, Cornelis J, Jansen B (2013) Phased searching with NEAT in a time-scaled framework: experiments on a computer-aided detection system for lung nodules. Artif Intell Med 59:157–167