Peer review file

Article information: http://dx.doi.org/10.21037/jtd-20-3522

Reviewer A

Authors evaluated the performance of AI based CAD system for the detection of lung nodules, by comparing to that of radiologists in training. They have shown that diagnostic accuracy was significantly higher in radiologists than that of AI system, and concluded that AI based CAD system should be a second reader.
The result of this study seems to be useful, but I have some comments.

Comment 1. In background, I could not grasp why radiologists in training were chosen for comparison.

Reply 1:
First, we would like to thank the reviewer very much for the appreciation of our work and the insightful comments.

Concerning the rational for choosing readers that are still in training. We decided to include exclusively radiologists in training because we expected this group the most likely to benefit from using an AI CAD system. The results of Nam et al. implied that the use of a deep learning based automatic detection (DLAD) system enhanced the performance of all groups (non-radiology physicians, residents, board-certified radiologists and thoracic experts), but had the greatest impact on the non-radiology physicians and residents [*Nam JG, Park SG, et al. Development and Validation of Deep Learning-Based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology 2019; 290:218–228.].*

Furthermore, we believe that in reality only in a minority of institutions chest radiographs are actually read by subspecialized, thoracic radiologists rather than by radiologists in training or board-certified radiologists, who are not further subspecialized thoracic imaging experts.
In addition, the chosen title might be misleading. Our readers included five resident

readers as well as two board-certified radiologists. The title refers to the fact that the board-certified readers are currently in subspecialization training. We apologize for the indistinct phrasing. If required, we would suggest rephrasing the title as follows:

"Title: Performance of an AI based CAD system in solid lung nodule detection on chest phantom radiographs compared to radiology residents and fellow radiologists."

Details on the readers experience can be found on page 7, lines 9 - 11: "Human readout".

Comment 2.
As authors mentioned in Discussion, the main drawback of this study was usage of phantoms, resulting in improbable findings (such as pleural effusion and rib fracture) by AI system. If radiologists know that they are phantom, the comparison seems not to be fair.

Reply 2:
This statement is correct; we do appreciate the feedback. This study was not primarily designed to compare false positive rates between software and radiologists. The substantial false positive findings of the software were rather a coincidental observation, which we thought merit a closer look. The false-positive detections revealed a software' problem with ancillary findings and may be useful in the future regarding optimization of the algorithm. However, if required, we would suggest deleting the regarding paragraph on page 9, line 29 - page 10, line 2.

Reviewer B

Comment 1: Why did the authors just investigate the performance of radiologists in training?

Reply 1:

Thank you very much for your positive assessment of the submitted manuscript.

This point has also been raised by reviewer 1. First, we apologize for the misleading phrasing in the title. Actually, the readers included five radiology residents as well as two board-certified radiologists. The title refers to the fact that the board-certified readers are currently in sub- specialization training. Furthermore, the rational to include radiologists in (sub specialization) training is that presumably the most beneficial implementation of an AI-CAD system can be expected among less experienced readers. Results previously published by Nam and colleagues also showed that the positive impact of an AI-supported CAD system is highest among resident readers. In order to clarify, we would suggest rephrasing the title as follows:

"Title: Performance of an AI based CAD system in solid lung nodule detection on chest phantom radiographs compared to radiology residents and fellow radiologists."

Comment 2: How could the radiologists evaluate the images? (on a FDA approved setup? Were they allowed to change the window settings?)

Reply 2:

Thank you for this comment, please excuse the inaccuracy. All examinations were read and analyzed on a dedicated PACS-workstation (Sectra PACS IDS7, Sectra). The readers were allowed to change the window settings at will. We modified the methods section by including the following paragraph:

Changes in the text: page 8, lines 23-26:

"The readout was performed on a dedicated PACS-workstation (Sectra PACS IDS7, Sectra) with dedicated monitors (BARCO Coronis Fusion 6MP LED). The readers were

allowed to adjust window settings to allow for natural reading conditions."

Comment 3: How dense were the nodules (all soft tissue density?)

Reply 3:

For this study, only nodules with a soft tissue density were used. Since the detectability of sub-solid nodules on chest radiography is rather limited, we decided not to include those into this study. However, it would be interesting to evaluate the performance of the software regarding sub-solid nodule detection, once the algorithm has been further optimized. In order to clarify, we suggest the nodule density values to the methods section.

Changes in the text: page 7, line 8:

"Artificial solid nodules (density: 100 HU; diameters of 5, 8, 10, and 12 mm) were randomly placed inside the phantoms."

Reviewer C

This study addressed the diagnostic performance of AI CXR classifier and radiologists Through chest phantom with pulmonary nodules. This study found that human readers show superior accuracy as compared to an AI 12 based CAD system. However, the AI-CAD system used did detect different lesions than the 13 radiologists; emphasizing the role of such a system as a second reader device.

Comment 1: Introduction

Please try to address different vendor AI assisted CXR reading for pulmonary nodules performance such as Quibim (Spain), Lunit (Korea), etc. As far as we know, previous study had demonstrated that different AI algorithen diagnostic performance, showing these algorithms could support triage workflow via double reading to improve sensitivity and specificity during the diagnostic process.

References:

1. Identifying pulmonary nodules or masses on chest radiography using deep learning:

external validation and strategies to improve clinical practice. Clinical radiology 2020, Volume 75, Issue 1, January 2020, Pages 38-45

2. Nam JG, Park SG, et al. Development and Validation of Deep Learning-Based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs, Radiology 2018

Reply 1:

Thank you very much for pointing at those excellent references. As you proposed, we have included them into the background section and the text was modified accordingly. In addition, we updated the references as suggested.

Changes in the text:

- Page 5, lines 20 – 26:

"Liang et al. stated that not only the sensitivity of nodule detection increased by the use of an AI algorithm as a second reader device, but that the algorithm could make the daily workflow more efficient (Liang at al.). The QUIBIM Chest X-ray classifier achieved rapid processing times of 94.07±16.54 seconds per case (Liang et al.). Another deep learning based automatic detection algorithm applied in the study of Nam et al. enhanced the performance of both, unexperienced and expert readers, regarding nodule detection and therefore resulted in an optimized workflow as well (Nam et al.)."

Changes in the references:

-Page 16, lines 14 – 21:

23. Liang CH, Liu YC, Wu MT, et al. Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. Clin Radiol 2020 Jan;75(1):38-45.

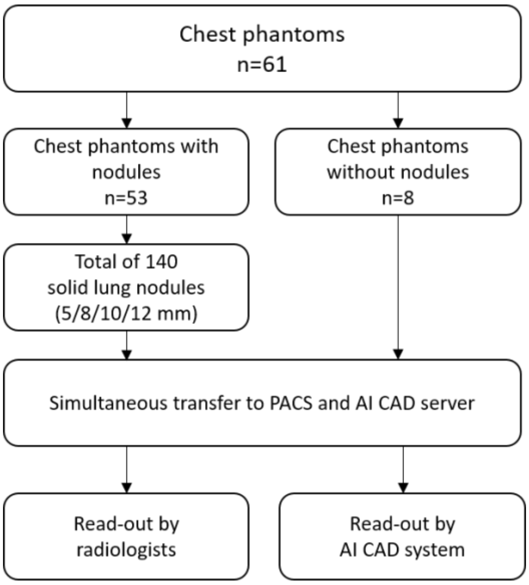
24. Nam JG, Park S, Hwang EJ, et al. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology 2019 Jan;290(1):218-228.

Comment 2: Method

Please try to add the study design flowchart to clearly address these study design points.

Reply 2:

Please find attached a study design flowchart. We hope this flowchart clearly depicts the study design. In case more information is required, we would be happy to expand the chart. We suggest to include the study design flowchart as figure 1 (page 7, line 14). The figure numbering has been rearranged accordingly.


Chest phantoms
n=61
Chest phantoms with nodules
n=53
Chest phantoms without nodules
n=8
Total of 140 solid lung nodules (5/8/10/12 mm)
Simultaneous transfer to PACS and AI CAD server
Read-out by radiologists
Read-out by AI CAD system

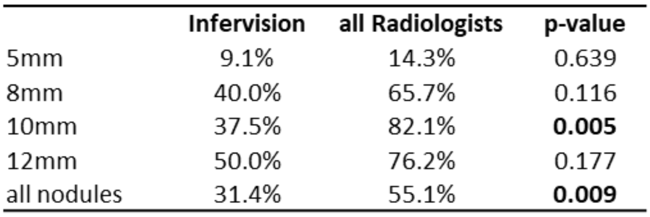
Comment 3: Result and discussion

1. Please try to analysis the diagnostic performance of different nodule size (AI CXR vs. radiologists, 5/8/10/12mm).

Reply 3.1:

1. Thank you for the valuable input. We now performed a comparison of AI vs. radiologists regarding the different nodule sizes (please see table below).

Size dependent nodule sensitivity of AI versus radiologists



	Infervision	all Radiologists	p-value
5mm	9.1%	14.3%	0.639
8mm	40.0%	65.7%	0.116
10mm	37.5%	82.1%	**0.005**
12mm	50.0%	76.2%	0.177
all nodules	31.4%	55.1%	**0.009**

In the table, we compared the size-dependent per-nodule sensitivity of the AI-CAD and the radiologists. Although the sensitivity of radiologists was in general superior as compared to the algorithm (p=.009), significant results were only found when comparing the true-positive rate in lesions measuring 10mm. This is an interesting finding since the systems sensitivity was better in lesions measuring 8mm.

We would suggest including the size-dependent performance analysis in the results section. The new table number is table 4, entitled "Size dependent nodule sensitivity of AI versus radiologists".

Changes in the text: page 11, lines 15 - 18:

"A size dependent analysis of nodule detection sensitivity showed, that radiologists in general detected significantly more nodules as compared to the AI-CAD (p=.009). However, only in lesions measuring 10mm in diameter, results were found to be significant (p=.005; table 4)."

Comment 3.2. The text in Figure 2 is very unclear, please revise it

Reply 3.2.

We apologize for the poor quality of figure 2. We revised figure 2 accordingly.

Comment 4.

Others: Grammar and spelling

A few minor typos, grammar hiccups should also be corrected, e.g., "All data analysis were" in line 36, page 4.

Page 6 -Corrected grammatical mistake in the "for example" to "for example," (line 27)

Reply 4:

Thank you for pointing at those typos and grammatical flaws. We have corrected the manuscript accordingly.