TECHNOLOGY REVIEW (MINI-HTA)

ARTIFICIAL INTELLIGENCE-BASED CHEST X-RAY FOR LUNG CANCER SCREENING

Malaysian Health Technology Assessment Section (MaHTAS)
Medical Development Division
Ministry of Health Malaysia
009/2022

IPT Mark
INAHTA
Mini-HTA

DISCLAIMER

This technology review (mini-HTA) is prepared to assist health care decision-makers and health care professionals in making well-informed decisions related to the use of health technology in health care system, which draws on restricted review from analysis of best pertinent literature available at the time of development. This technology review has been subjected to an external review process. While effort has been made to do so, this document may not fully reflect all scientific research available. Other relevant scientific findings may have been reported since the completion of this technology review. MaHTAS is not responsible for any errors, injury, loss or damage arising or relating to the use (or misuse) of any information, statement or content of this document or any of the source materials.

Please contact htamalaysia@moh.gov.my if further information is required.

Malaysian Health Technology Assessment Section (MaHTAS)
Medical Development Division
Ministry of Health Malaysia
Level 4, Block E1, Precinct 1
Government Office Complex
62590, Putrajaya
Tel: 603 8883 1229

Available online via the official Ministry of Health Malaysia website: http://www.moh.gov.my

e ISBN 978-967-2887-52-2

SUGGESTED CITATION: Syful Azlie MF and Izzuna MMG. Artificial intelligence-based chest x-ray for lung cancer screening. Technology Review. Ministry of Health Malaysia: Malaysian Health Technology Assessment Section (MaHTAS); 2022. 48 p. Report No.: 009/2022

DISCLOSURE: The author of this report has no competing interest in this subject and the preparation of this report is entirely funded by the Ministry of Health Malaysia.

Prepared by

Syful Azlie Md Fuzi
Biochemist
Principal Assistant Director
Health Technology Assessment Section (MaHTAS)
Medical Development Division
Ministry of Health Malaysia

Reviewed by

Dr. Izzuna Mudla Mohamed Ghazali
Public Health Physician
Deputy Director
Malaysian Health Technology Assessment Section (MaHTAS)
Medical Development Division
Ministry of Health Malaysia

External reviewer(s)

Dr. Irfhan Ali Hyder Ali
Head of Department and Senior Consultant Pulmonologist
Hospital Pulau Pinang
(Head of Respiratory Services, Ministry of Health)

Dr. Norzaini Rose Mohd Zain
Consultant Radiologist
Hospital Kuala Lumpur
(Head of Radiology Services, Ministry of Health)

EXECUTIVE SUMMARY

Background

Because of its accessibility, low cost, low radiation dose exposure, and reasonable diagnostic accuracy, chest radiographs are frequently used for early detection of pulmonary nodules despite their inferiority to low-dose computed tomography (LDCT). Although LDCT has demonstrated a clear benefit for reducing all-cause mortality among high-risk group in lung cancer screening, the high rate of false-positives, low uptake, and the cost of unnecessary diagnostic procedures are important limitations of this approach. The introduction of targeted therapies and immunotherapeutic agents have resulted in a longer duration of overall survival compared with standard chemotherapy. However, these novel therapies are not effective in all patients; thus, early detection remains the most important intervention window for improving patient survival. The emergence of artificial intelligence (AI) as a new tool for assessing medical data implies new opportunities for improving the diagnosis and treatment of various human diseases. In the case of lung cancer diagnosis, coupling AI algorithms with available clinical and biomedical data seems to have the potential to improve lung cancer screening methods. These points lead to a question - Can AI help to provide a "second pair of eyes" for detecting nodules more accurately and earlier in their progression? Hence, this technology review was requested by the Head of Cancer Unit, Disease Control Division, Ministry of Health Malaysia to evaluate the role that AI-based chest x-ray could play in assisting radiologists and to increase the accuracy and efficiency of lung cancer diagnosis.

Objective/ aim

The objective of this technology review was to evaluate the diagnostic accuracy, efficacy, safety, and economic implication of artificial intelligence-based chest x-ray for lung cancer screening, and to compare its performance with radiologists.

Results and conclusions:

Search results

A total of 2,288 records were identified through the Ovid interface and PubMed while four were identified from references of retrieved articles. No duplicates references were found; 2,292 potentially relevant titles were screened using the inclusion and exclusion criteria. Of these, 13 relevant abstracts were retrieved in full text. After reading, appraising and applying the inclusion and exclusion criteria to the 13 full text articles, 10 were included while three were excluded since the studies included irrelevant population (cases with combined lung disease) and irrelevant outcome. All full text articles finally selected for this review were retrospective diagnostic study. The studies were conducted mainly in the United States, United Kingdom, Germany, South Korea, and Japan.

Diagnostic accuracy/ efficacy

There was substantial fair level of retrievable evidence to suggest that radiologists had better diagnostic performance interpretation with artificial intelligence (AI) algorithm than without for the detection of lung cancer on chest radiographs. Findings in general reported that:

i. The average sensitivity improved from 66.4% (ranged 45.0%-87.7%) to 74.7% (ranged 55.5%-93.9%) and the number of false-positive findings per radiograph declined from 0.25 (ranged 0.20-0.30) to 0.18 when the radiologists re-reviewed radiographs with AI algorithm. However, specificity was similar with AI-aided (ranged 86.0%-97.0%) and without AI-aided interpretation (ranged 79.0%-96.0%).
ii. Junior radiologists showed greater improvement in sensitivity with AI-aided interpretation as compared with their senior counterparts (12.0%, 95% confidence interval [CI]: 4.0% to 19.0% versus 9.0%, 95% CI: 1.0% to 17.0%) while senior radiologists experienced larger improvement in specificity (mean improvement 4.0%, 95% CI: -2.0% to 9.0%) compared with the junior group (1.0%, 95% CI: -3.0% to 5.0%).
iii. General physicians benefited more from the use of the AI algorithm than radiologists. The performance of general physicians was improved from 47.0% to 60.0% for sensitivity, from 96.0% to 97.0% for specificity, from 75.0% to 82.0% for positive predicted value (PPV), and from 89.0% to 91.0% for negative predicted value (NPV) while the performance of radiologists was improved from 51.0% to 60.0% for sensitivity, from 96.0% to 96.0% for specificity, from 76.0% to 80.0% for PPV, and from 89.0% to 91.0% for NPV.
iv. Artificial intelligence algorithm enhanced the performance of readers for the detection of lung cancers on chest radiographs when used as second reader. Compared to that without AI, the average sensitivity increased significantly for radiology residents (61.0% [95% CI: 55.0% to 67.0%] versus 72.0% [95% CI: 66.0% to 77.0%]; p=0.016), but specificity was similar with AI (p=0.89). For radiologists, average sensitivity was similar (p=1.00) but specificity increased with AI (79.0% [95% CI: 77.0% to 81.0%] versus 86.0% [95% CI: 84.0% to 87.0%]; p<0.001).
v. With AI, radiology residents were able to recommend more chest CT examinations (54.7% versus 70.2%, p<0.001) for patients with visible lung cancer whereas radiologists recommended significantly less proportion of unnecessary chest CT examinations (16.4% versus 11.7%, p<0.001) in cancer-negative patients.
vi. For detection of visible lung cancers on the chest radiography in healthy population, the stand-alone AI algorithm performance was comparable to that radiologist with sensitivity, specificity, PPV, NPV, and false-positive rate of 83.0%, 97.0%, 1.3%, 100%, and 3.0%, respectively.

Summary of studies related to diagnostic accuracy/ efficacy of artificial intelligence-based chest x-ray for lung cancer screening are shown in **Table 1**.

Table 1: Diagnostic accuracy/ efficacy of artificial intelligence-based chest x-ray for lung cancer screening reported by the included studies

Study/ criteria	Sensitivity (%)			Specificity (%)			PPV (%)			NPV (%)			FPPI			Accuracy			AUC			JAFROC FOM		
	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI
Sim Y 2019																								
-MPN	67.3	65.1	70.3										0.20	0.20	0.18									
Cha MJ 2019																								
-Lung cancer	86.5	78.0											0.1	0.1					0.899	0.819				
	92.0	85.0											0.3	0.3										
Nam JG 2019																								
-MPN	80.7	70.4		95.2	85.8								0.3	0.25					0.91	0.86	0.95	0.885	0.794	0.928
Lee JH 2020																								
-General population	83.0	40.0		97.0	97.0		1.3	1.3		100	100					97.0	97.0		0.97					
Yoo H 2020																								
1-Chest radiograph																			0.93					
-Nodule detection	86.2	87.7		85.0	88.0																			
-Cancer detection	75.0			83.3			3.8			99.8														
-MPN	94.1			83.3			3.4			100.0														
2-Digital radiographs																			0.99					
-Nodule detection	96.0	88.0		93.2	82.8																			
-Cancer detection	76.0	80.0		90.0	91.1		9.1	9.8		99.7	99.9													
-MPN	100.0	94.1		90.9	91.0		8.2	7.8		100.0	99.9													

Study/ criteria	Sensitivity (%)			Specificity (%)			PPV (%)			NPV (%)			FPPI			Accuracy			AUC			JAFROC FOM		
	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI	AI	Rad	Rad + AI
3-CT radiograph																			0.86					
-Nodule detection	77.8	86.1		78.8	90.4																			
-Cancer detection	68.4	89.5		76.7	91.4		1.9	6.3		99.7	99.9													
-MPN	85.7	92.9		76.7	91.3		1.8	5.0		99.9	100.0													
Homayounieh F 2021																								
-Pulmonary nodule		45.0	55.0		93.0	95.0											69.0	75.0						
Koo YH 2021																								
-Pulmonary nodule		88.6	93.9																0.87	0.93	0.96	0.926	0.929	0.964
Tam D 2021																								
-Lung cancer	0.80	0.78	0.91	0.93	0.96	0.90										0.87	0.87	0.91						
Ueda D 2021																								
-Nodule detection	0.66	0.49	0.60	0.96	0.96	0.97	0.78	0.75	0.81	0.92	0.89	0.91				0.90	0.87	0.90						
Yoo H 2021																								
Lung cancer		Without AI	With AI		Without AI	With AI		Without AI	With AI		Without AI	With AI		Without AI	With AI		Without AI	With AI		Without AI	With AI		Without AI	With AI
Radiology residents		0.61	0.72		0.88	0.88								0.15	0.12		0.76	0.82						
Radiologist		0.76	0.76		0.79	0.86								0.24	0.17		0.82	0.84						

PPV, positive predictive value; NPV, negative predictive value; FPPI, false-positives per image; AUC, area under the ROC curve; JAFROC FOM, jackknife alternative free-response receiver-operating characteristic figure of merit; AI, artificial intelligence; Rad, radiologist; MPN, malignant pulmonary nodule

Safety

There was no retrievable evidence on the adverse events or complications related to the use of artificial intelligence-based chest x-ray for lung cancer screening. Currently, there have been several AI algorithms approved by the United States Food and Drug Administration (FDA) for specific clinical indications, specifically to thoracic radiology. The majority of these algorithms are approved for detection and segmentation of pulmonary nodules. Several algorithms were registered as CE-mark (Class IIa) medical device.

Organisational

There was no retrievable evidence in the context of procedural time points and training or learning curve related to artificial intelligence-based chest x-ray for lung cancer screening. Nevertheless, AI-aided seems to help radiologists to read images effectively and providing complementary interpretation, highlighting areas of the scan requiring particularly close inspection. This information may reduce time to diagnosis, help avoid fatigue- or workload-induced missed diagnoses, and can be invaluable in helping address shortage of trained radiologists.

Economic implication

The cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening has not yet been formally evaluated. However, cost associated to use the technologies generally consists of implementation or integration fee (ranged from £6,000 to £10,000 per centre), annual licence and maintenance fee (£60,000), and fixed cost per scan processed (between £0.90 and £1.66 per image) which would be in addition to standard care (£32.73) and it depends on the total patient throughput.

Conclusion

A substantial body of retrievable evidence suggests that radiologists assisted with artificial intelligence algorithm was associated with improved detection of lung cancer on chest radiographs with a higher sensitivity and fewer false-positive findings per image compared with radiologists alone, irrespective of radiologist experience and nodule characteristics. Indeed, it can enhance the performance when used as second reader by improving the quality of reading for various reader groups. Since AI-based chest x-ray can help reduce the false-positive rate, it will enable cost-effective lung cancer screening by reducing over-diagnosis and the follow-up costs for additional scans and biopsies of benign nodules. Several factors (robust interoperability, data sharing infrastructure, real-world validation studies, and training) need to be taken into consideration.

Methods

The following electronic databases were searched through the Ovid interface: MEDLINE (R) ALL 1946 to Jul 29, 2022. Parallel searches were run in PubMed, US FDA and INAHTA database while additional articles were retrieved from reviewing the bibliographies of retrieved articles. The search was limited to articles on human. There was no language limitation in the search. The last search was conducted on 2nd August 2022.

TABLE OF CONTENTS

ABBREVIATION

AEs	Adverse events or adverse effects
AI	Artificial intelligence
AUC	Area under the ROC curve
CASP	Critical Appraisal Skills Programme
CAD	Computer-aided detection
CI	Confidence interval
CXR	Chest radiography
DCNN	Deep convolutional neural network
DICOM	Digital imaging and communications in medicine
DLAD	Deep learning–based automatic detection algorithm
FROC	Free response receiver-operating characteristics curve
INAHTA	International Network of Agencies for Health Technology Assessment
JAFROC FOM	Jackknife alternative free-response receiver-operating characteristic figure of merit
LDCT	Low-dose computed tomography
MaHTAS	Malaysian Health Technology Assessment Section
NPV	Negative predicted value
MOH	Ministry of Health
PACS	Picture archiving and communication system
PPV	Positive predicted value
QALY	Quality adjusted life year
RIS	Radiology information system
RCT	Randomised controlled trial
US FDA	United States Food and Drug Administration

1.0 BACKGROUND

Chest radiography are one of the most basic imaging tests in medicine and the most common examination in routine clinical work such as screening for chest disease, diagnostic workup, and observation. One of the features physicians look for in these chest radiographs is nodules - an indicator of lung cancer, which has the highest mortality rate in the world. The majority of the patients diagnosed with lung cancer are in the late-stage, and therefore have a poor prognosis (5-year survival rates decreasing from 62% for stage I diagnosis to 3% for stage IV diagnosis). In addition to the late stage at diagnosis, the heterogeneity of imaging features and histopathology of lung cancer also makes it a challenge for clinicians to choose the best treatment option.[1]

In practice, low-dose computed tomography (LDCT) is recommended for lung cancer screening for at-risk individuals rather than chest radiography despite a false positive rate of approximately 27%.[2-3] Several studies concluded that LDCT was superior to radiographs which had a sensitivity of 36% to 84%,[4-7] varying widely according to tumour size, study population, and reader performance. Other studies showed that 19% to 26% of lung cancers visible on chest radiographs were actually missed at the time of initial reading.[6, 8] However, chest radiography remains the primary diagnostic imaging test for chest conditions because of its advantages over chest CT, including ease of access, lower cost, and lower radiation exposure. Notably, the higher number of chest radiographs per capita than chest CT indicates that chest radiography has more opportunities to detect lung abnormalities in individuals who are not considered at risk, leading to a diagnostic chest CT.[8]

To overcome the inherent limitations of human perception and time pressure in the clinical setting and to improve the efficacy of chest radiography for nodule detection, computer-aided detection (CAD) techniques were introduced in the early 2000s. Many prior studies have demonstrated promising results of CAD systems in improving nodule detection rates on chest radiographs.[9-11] On the contrary, others have raised concerns over the CAD that it produces too many false positive cases, hampering the utilisation of CAD as stand-alone diagnostic tool in routine practice.[12-14] With these conflicting results, it is expected that recent advances in artificial intelligence (AI) may have a potential to achieve parity or even surpass radiologist performances in certain interpretations of chest radiographs. Recently, the application of deep learning, a field of AI,[15-16] has led to dramatic, state-of the-art improvements in visual object recognition and detection. Automated feature extraction, a critical component of deep learning, has great potential for application in the medical field,[17] especially in radiology.[18]

Hence, this technology review was requested by the Head of Cancer Unit, Disease Control Division, Ministry of Health Malaysia to evaluate the role that AI could play in assisting radiologists and to increase the accuracy and efficiency of lung cancer diagnosis.

2.0 OBJECTIVE / AIM

The objective of this technology review was to evaluate the diagnostic accuracy, efficacy, safety, and economic implication of artificial intelligence-based chest x-ray for lung cancer screening, and to compare its performance with radiologists.

3.0 TECHNICAL FEATURE

Artificial intelligence is a general term that does not have a strict definition. It is an algorithm driven by existing data to predict or classify objects. The main components include the dataset used for training, pre-treatment method, an algorithm used to generate the prediction model, and the pre-trained model to accelerate the speed of building models and inherit previous experience. Machine learning (ML) is a subclass of AI, and is the science of obtaining algorithms to solve problems without being explicitly programmed, including decision trees (DTs), support vector machines (SVMs), and Bayesian networks (BNs). Deep learning is a further subclass of ML, featured with multiple layered ML, achieving feature selection and model fitting at the same time. The hierarchical relationship between those definitions is displayed in **Figure 1**. Moreover, AI is a general term for a program that predicts an answer to a certain problem, where one of the conventional methods is logistic regression. Machine learning learns the algorithm through input data without explicit programming. It includes algorithms such as DTs, SVMs, and BNs. By using each ML algorithm as a neuron with multiple inputs and a single output, a neural network is a structure that mimics the human brain. Deep learning is formed with multiple layers of neural networks, and convolutional neural network (CNN) is one of the elements of the famous architecture.[19-20]

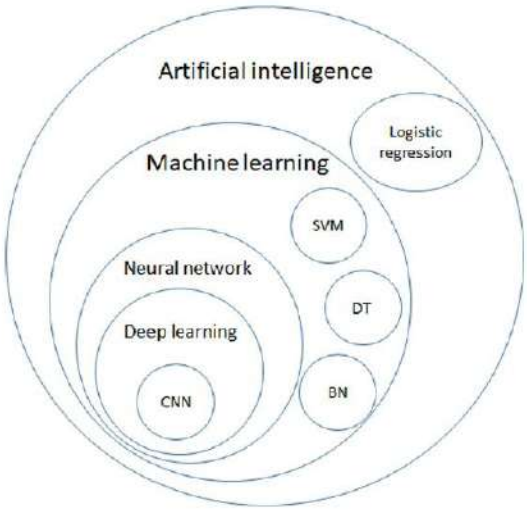

Figure 1: Venn diagram of AI, ML, neural network, deep learning, and further algorithms in each category

The chest x-ray imaging AI technologies in this review are standalone software platforms that use deep learning algorithms to interpret radiology images. Some technologies allow images to be transferred from the hospital to the software platform. The software analyses the chest DICOM (digital imaging and communications in medicine) image using proprietary algorithms. The image analysis may be sent directly back to the hospital to be viewed with hospital systems such as PACS (picture archiving and communication system) and some radiology information systems (RIS), using protocols such as DICOM and HL7. Some technologies may also allow uploading and viewing of images and analysis using a web interface.[21]

The AI technology may help identify images as normal or abnormal, highlight suspected abnormalities and provide results as heat maps or clinically relevant labels. It may also provide support for prioritising x-rays for specialist review. The AI analyses are intended to be used with radiology images to support radiologist review and decision making to improve diagnostic accuracy. They are not intended to be used as medical advice.[21]

Figure 2: Artificial intelligence software for analysing chest x-ray images

4.0 METHODS

A systematic review was conducted. Search strategy was developed by the main author and an *Information Specialist*.

4.1 SEARCHING

The following electronic databases were searched through the Ovid interface: **MEDLINE (R) ALL 1946 to Jul 29, 2022.**

Other databases:

- PubMed
- Other websites: US FDA, INAHTA.

General databases such as Google and Yahoo were used to search for additional web-based materials and information. Additional articles retrieved from reviewing the bibliographies of retrieved articles. The search was limited to articles on human. There was no language limitation in the search. **Appendix 1** showed the detailed search strategies. The last search was conducted on 2nd August 2022.

4.2 SELECTION

A reviewer screened the titles and abstracts against the inclusion and exclusion criteria. Relevant articles were then critically appraised using *Critical Appraisal Skills Programme (CASP) checklist*. Studies were graded according to *US/ Canadian Preventive Services Task Force* (**Appendix 2**) or NHS Centre for Reviews and Dissemination (CRD) University of York, Report Number 4 (2nd Edition) for diagnostic accuracy studies (**Appendix 3**). All data were extracted and summarised in evidence table as in **Appendix 4**.

The inclusion and exclusion criteria were:

Inclusion criteria:

a.	**Population**	Adults who are at risk of having lung cancer or referred for radiological imaging
b.	**Intervention**	Artificial intelligence-based chest x-ray, chest imaging AI, machine learning, computer-aided diagnosis systems

c.	**Comparator**	Chest x-ray, radiologist review, LDCT, test assay, lung biopsy
d.	**Outcomes**	**Diagnostic accuracy:** sensitivity, specificity, area under the curve (AUC), number of false-positive findings per image **Safety:** Mortality, adverse events (AEs), postoperative complications **Organisational issues:** Hospital utilisation (readmission, length of stay), procedural time points and training or learning curve **Economic implications:** Cost, cost-effectiveness, cost-utility analysis
e.	**Study design**	HTA reports, systematic review with/out meta-analysis, randomised controlled trial (RCT), cohort, diagnostic, case-control, case series, economic evaluation studies
f.	Full text articles published in English	

Exclusion criteria:

a.	**Study design**	Case report, animal study, laboratory study, narrative review
b.	Non-English full text articles	

5.0 RESULTS

Search results

An overview of the search is illustrated in **Figure 3**. A total of **2,288** records were identified through the Ovid interface and PubMed while **four** were identified from references of retrieved articles. No duplicates references were found; **2,292** potentially relevant titles were screened using the inclusion and exclusion criteria. Of these, **13** relevant abstracts were retrieved in full text. After reading, appraising and applying the inclusion and exclusion criteria to the **13** full text articles, **10** were included while **three** were excluded since the studies included irrelevant population (cases with combined lung disease) and irrelevant outcome. All full text articles finally selected for this review were retrospective diagnostic studies. The studies were conducted mainly in the United States, United Kingdom, Germany, South Korea, and Japan.

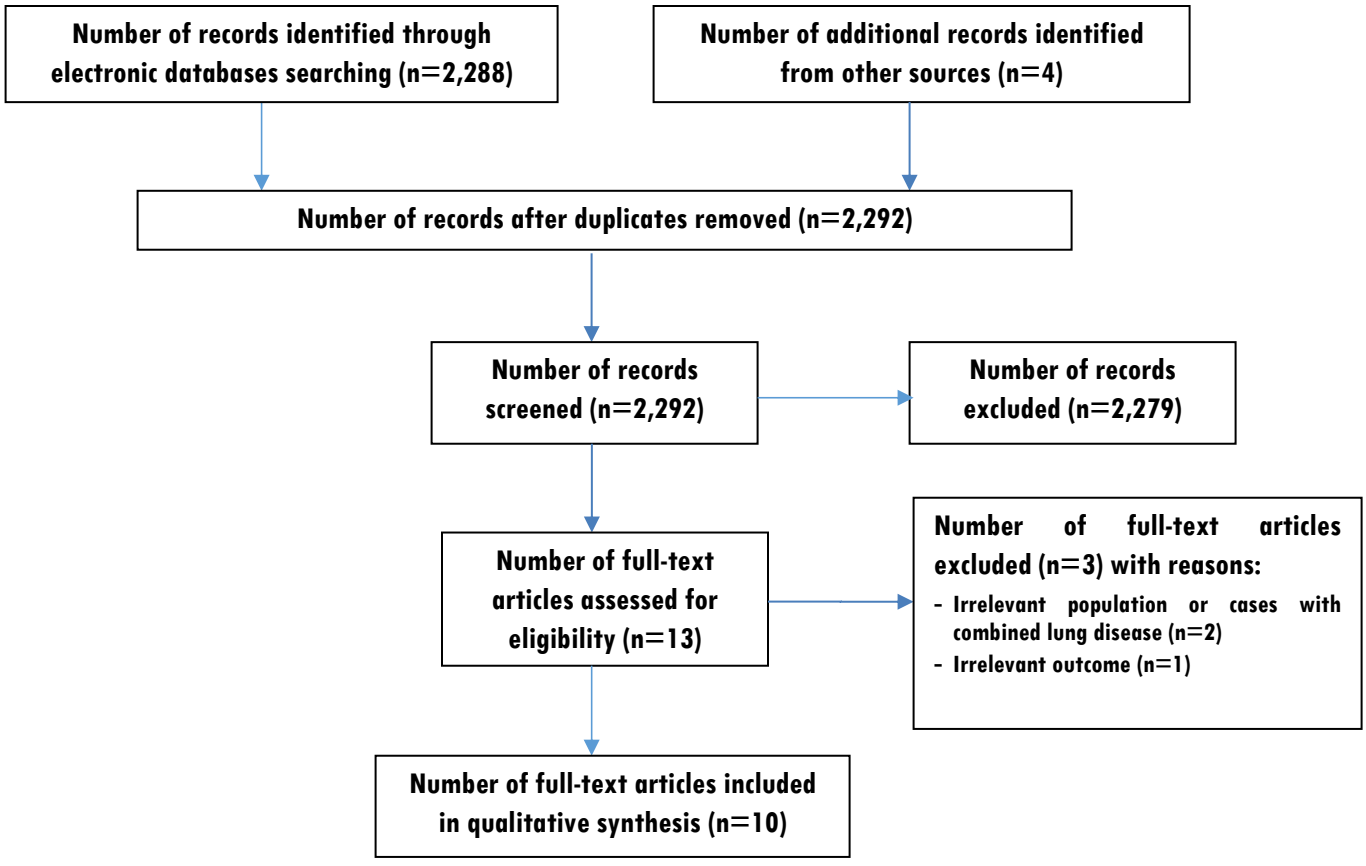

Figure 3: Flow chart of retrieval of articles used in the results

Quality assessment of the studies

Risk of bias was assessed using Critical Appraisal Skill Programme (CASP) checklist for all diagnostic studies in this review. These assessments involved answering a pre-specified question of those criteria assessed and assigning a judgement relating to the risk of bias.

Risk of bias assessment for included diagnostic study

Ten studies were included in this assessment. All were judged to have overall low risk of bias although these was a retrospective in nature and hence did not exactly represent the real-world setting, primarily the prevalence of lung cancers in the dataset (**Figure 4**).

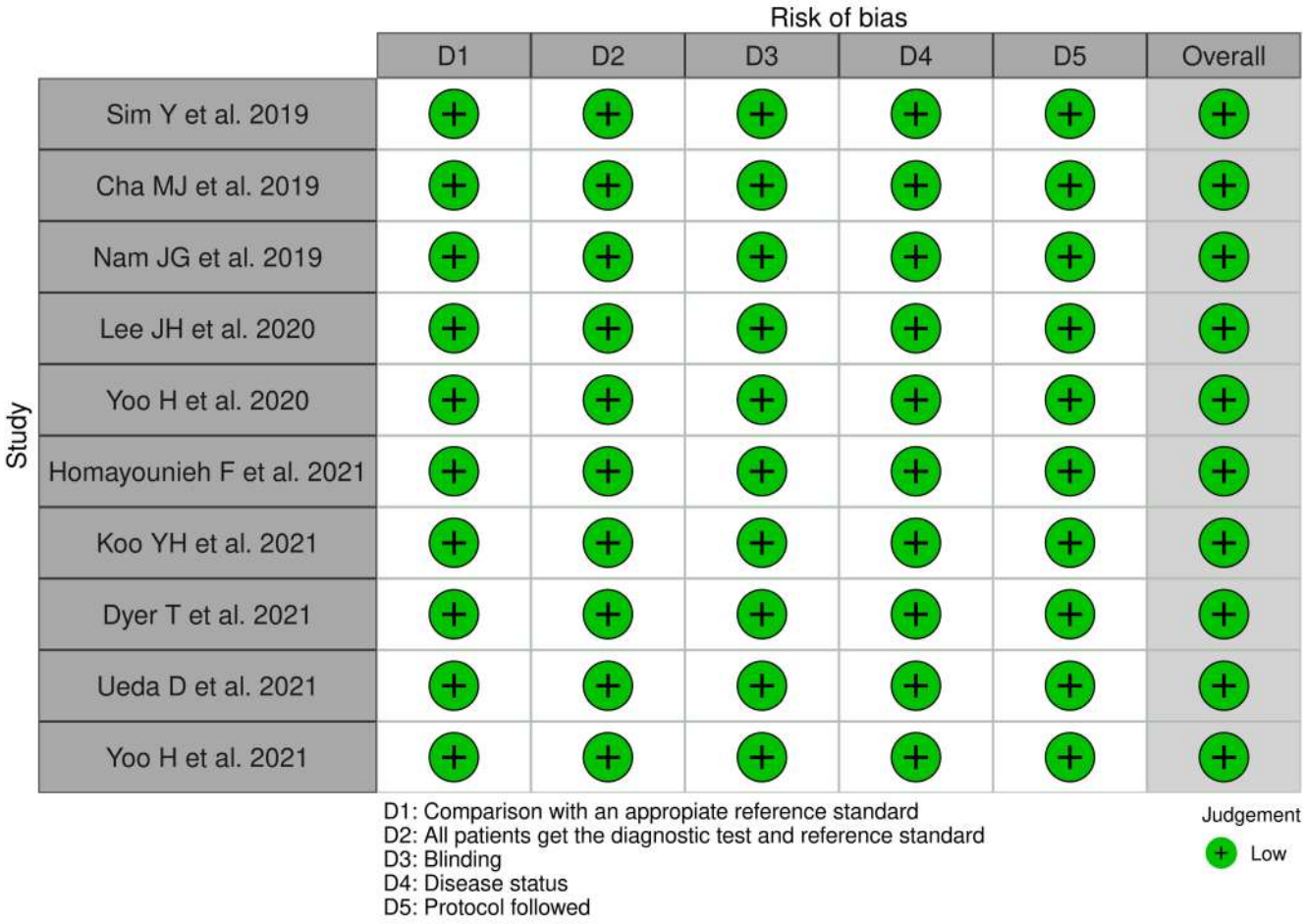

Figure 4: Assessment of risk of bias of diagnostic study (CASP)

5.1 DIAGNOSTIC ACCURACY/ EFFICACY

Recent evidence has demonstrated that deep learning algorithms can predict the risk of lung cancer with model performances rated on par with radiologists. In 2019, Sim Y et al. evaluated the clinical efficacy of software using a deep convolutional neural network (DCNN) algorithm for the detection of malignant pulmonary nodules on chest radiographs. A total of 800 radiographs were included in the study; 200 normal radiographs (mean age 52.4 ± 12.2 years) and 600 lung cancer-containing radiographs (mean age 61.5 ± 12.3 years) from four different centres with zero to three malignant nodules and analysed by 12 radiologists with varying levels of experience. The standard of reference for malignant nodules was CT with pathologic confirmation. The sensitivity and number of false-positive findings per image of DCNN software, radiologists alone, and radiologists with the use of DCNN software were analysed by using logistic regression and Poisson regression. The average sensitivity of radiologists improved from 65.1% (1,375 of 2,112; 95% confidence interval {CI]: 62.0% to 68.1%) to 70.3% (1,484 of 2,112; 95% CI: 67.2% to 73.1%; p<0.001) and the number of false-positive findings per radiograph declined from 0.20 (488 of 2,400; 95% CI: 0.18 to 0.22] to 0.18 (422 of 2,400; 95% CI: 0.16 to 0.2]; p<0.001) when the radiologists re-reviewed radiographs with the DCNN software (**Table 2**). The stand-alone sensitivities of the DCNN and the number of false-positive findings per image were similar to those of the readers when analysed by centre. For the 12 radiologists in this study, 104 of 2,400 radiographs were positively changed (from false-negative to true-positive or from false-positive to true-negative) using the DCNN, while 56 of 2,400 radiographs were changed negatively. Small nodules were less sensitively detected both by radiologists alone and by radiologists with AI algorithm (1 cm $\leqslant$ x $\leqslant$ 2 cm; 57.6% versus 61.2%) when compared with larger nodule (2 cm $\leqslant$ x $\leqslant$ 3 cm; 75.5% versus 81.9%).[22, level II-3]

Table 2: Pooled results of radiologist reader and DCNN performance

Overall (12 radiologists, 704 nodules)	Sensitivity (per nodule) (%)	False-positive per image (FPPI)	Radiologists with DCNN vs. Radiologists without DCNN	Radiologists with DCNN vs. DCNN	Radiologists without DCNN vs. DCNN
Radiologists without DCNN	65.1	0.20			
			Sensitivity p<0.001	Sensitivity p=0.006	Sensitivity p=0.12
Radiologists with DCNN	70.3	0.18			
			FPPI p<0.001	FPPI p=0.13	FPPI p=0.76
DCNN alone	67.3	0.20			

In another study by Cha MJ et al. (2019), a deep learning model based on DCNN was trained with 17,211 chest radiographs (5,700 CT-confirmed lung nodules in 3,500 chest radiographs and 13,711 normal chest radiographs), finally augmented to 600,000 images. For validation, a trained deep learning model was tested with 1,483 chest radiographs with surgically resected lung cancer, marked and scored by two radiologists. The reference standard for the location and margin of each lesion was based on the review of CT. Diagnostic performances of deep learning model and six human observers (radiologists with experience for chest radiology from four to 21 years) were then compared with 500 cases (200 visible T1 lung cancer on chest radiographs and 300 normal image) and analysed using free response receiver-operating characteristics curve (FROC) analysis. They found that deep learning model showed high diagnostic performance in detecting operable lung cancer in chest radiography with a sensitivity of 76.8% (1,139/1,483) at a false-positive per image (FPPI) of 0.3 and area under the FROC curve (AUC) of 0.732 (**Table 3**). In the comparison with human readers, deep learning model demonstrated a sensitivity of 86.5% at 0.1 FPPI and a sensitivity of 92.0% at 0.3 FPPI with AUC of 0.899 at an FPPI range of 0.03 to 0.44 for detecting visible T1 lung cancers, which were superior to the average of six human readers (mean sensitivity; 78.0% [range 71.6% to 82.6%] at an FPPI of 0.1% and 85% [range 80.2% to 89.2%] at an FPPI of 0.3, AUC of 0.819 [range, 0.754 to 0.862] at an FPPI of 0.03 to 0.44) (**Table 4**).[23, level II-3]

Table 3: Performance of deep learning model for detecting operable lung cancer (test set)

Subtlety Score	1 (n = 55)		2 (n = 230)		3 (n = 609)		4 (n = 589)		
Location	Visible Lung	Opaque Area	Visible Lung	Opaque Area	Visible Lung	Opaque Area	Visible Lung	Opaque Area	Total
Number	44	11	194	36	561	48	559	30	1483
True positive	1	0	58	9	467	30	551	23	1139
False positive	18	3	98	26	166	24	98	10	443
Sensitivity (%)	2.3	0.0	30.1	24.3	83.5	60.0	98.9	71.9	76.8
FPPI	0.42	0.25	0.51	0.70	0.30	0.48	0.18	0.31	0.30
FNPI	0.98	1.00	0.70	0.75	0.17	0.38	0.02	0.23	0.23

Table 4: Comparison of diagnostic performance for detecting T1 lung cancer between human observers and deep learning model using a total of 500 chest radiographs (200 cases of visible T1 lung cancer on chest radiograph and 300 cases of normal image)

			Sensitivity (%)	Sensitivity (%)	AUC
	TP	FP	FPPI 0.1	FPPI 0.3	FPPI 0.03-0.44
Human					
Observer 1	172/200	83/500	82.0	89.2	0.862
Observer 2	149/200	16/500	80.4	84.5	0.828
Observer 3	168/200	216/500	74.8	82.0	0.792
Observer 4	167/200	52/500	82.6	88.2	0.860
Observer 5	162/200	136/500	71.6	80.2	0.754
Observer 6	173/200	169/500	76.5	85.5	0.818
Human average	165.17/200	112/500	78	85	0.819
Deep learning model	175/200	69/500	86.5	92	0.899

Nam JG et al. (2019) developed and validated a deep learning–based automatic detection algorithm (DLAD) for malignant pulmonary nodules on chest radiographs and then compared its performance with physicians including thoracic radiologists. For this retrospective study, DLAD was developed by using 43,292 chest radiographs (normal radiograph-to-nodule radiograph ratio 34,067:9,225) in 34,676 patients (healthy-to-nodule ratio 30,784:3,892; 19,230 men [mean age 52.8 years]; 15 446 women [mean age 52.3 years]) obtained between 2010 and 2015, which were labelled and partially annotated by 13 board-certified radiologists, in a convolutional neural network. Radiograph classification and nodule detection performances of DLAD were validated by using one internal and four external data sets from three South Korean hospitals and one U.S. hospital. For internal and external validation, radiograph classification and nodule detection performances of DLAD were evaluated by using the area under the receiver operating characteristic curve (AUROC) and jackknife alternative free-response receiver-operating characteristic (JAFROC) figure of merit (FOM), respectively. Performances of DLAD, physicians, and physicians assisted with DLAD were evaluated and compared. According to one internal and four external validation data sets, radiograph classification and nodule detection performances of DLAD were a range of 0.92 to 0.99 (AUROC) and 0.831 to 0.924 (JAFROC FOM), respectively. Regarding radiograph classification performance comparison, DLAD showed an excellent AUROC of 0.91, which was higher than 16 of 18 physicians (p value range <0.001 to 0.52), with statistical significance demonstrated in 11 physicians. Regarding nodule detection performance comparison, DLAD exhibited a JAFROC FOM of 0.885, higher than all physicians and significantly higher than those of 15 of 18 physicians ($p<0.05$). Regarding the added value of DLAD as a second reader, the radiograph classification performance of physicians improved with DLAD (17 of 18 physicians; mean AUROC improvement of 0.04 [range -0.0001 to 0.14]) and the changes were statistically significant in 15 of 18 physicians. Regarding nodule detection performance, all physicians showed improved detection performances by using DLAD (mean JAFROC FOM improvement 0.043 [range 0.006-0.190]) with significant differences in 14 physicians. A JAFROC FOM of non-radiology physicians, radiology residents, board-certified radiologists, and thoracic radiologists' subgroups were 0.691, 0.796, 0.821, and 0.833, respectively, and in all four subgroups, their nodule detection performances improved with DLAD (JAFROC FOM for non-radiology physicians, radiology residents, board-certified radiologists, and thoracic radiologists, respectively: 0.828, 0.829, 0.840, and 0.854; corrected $p<0.05$).[24, level II-3]

To test whether a deep learning algorithm could evaluate chest radiographs for lung cancer in a large-scale health screening population, Lee JH et al. (2020) applied a commercially available deep learning model based on DCNN and compared its performance to that of radiologists. Out-of-sample testing of a deep learning algorithm was retrospectively performed (validation test) using chest radiographs from individuals undergoing a comprehensive medical check-up between July and December 2008 (10,206 individuals [5,859 men and 4,347 women; mean age 54 ± 11 years] with 10,289 chest radiographs were included in the task of detecting cancer-positive radiographs while for detecting visible lung

cancer, 10,285 radiographs from 10,202 individuals were used). Additionally, the deep learning algorithm was applied to a screening cohort undergoing chest radiography (50,098 individuals [28,105 men and 21,993 women; mean age 53 ± 11 years] with 100,576 radiographs were included in the task of detecting cancer-positive radiographs while 50,070 radiographs from 100,525 individuals were used for detecting visible lung cancer) between January 2008 and December 2012, and its performances were calculated. For the detection performances of the deep learning algorithm and pooled radiologists for visible lung cancers on chest radiographs, the algorithm's AUC was 0.99 (95% CI: 0.97 to 1.00), and it detected three more lung cancers (9/10 chest radiographs; sensitivity 90%) than the radiologists (6/10 chest radiographs; sensitivity 60%). However, this difference was not statistically significant (p=0.25). The algorithm had a negative predicted value (NPV) equivalent to that of the radiologists (100% versus 100%; p=0.09), but it had a lower specificity and positive predictive value (PPV) and a higher false-positive rate (specificity 97% versus 100%; p<0.001; PPV 2.7% versus 19%; p<0.001; false-positive rate 3.1% versus 0.3%; p<0.001) (**Table 5**). For detection of visible lung cancers on the chest radiography in the entire screening cohort, the algorithm had an AUC of 0.97 (95% CI: 0.95 to 0.99). Visible lung cancers were correctly detected on 39 of 47 radiographs (sensitivity 83%). The specificity, NPV, PPV, and false-positive rate of the algorithm for detecting visible lung cancers were 97%, 100%, 1.3%, and 3.0%, respectively (**Table 6**).[25, level II-3]

Table 5: Comparison between diagnostic performance of deep learning algorithm and that of three board-certified radiologists for detection of visible lung cancers on chest radiographs in validation test cohort

Variable	Sensitivity (%)	*P* Value	Specificity (%)	*P* Value	FPR (%)	*P* Value	NPV (%)	*P* Value	PPV (%)	*P* Value	Accuracy (%)
Pooled performance of three radiologists	60 (6/10) [26, 88]	…	100 (10 249/10 275) [100, 100]	…	0.3 (26/10 275) [0.2, 0.4]	…	100 (10 249/10 253) [100, 100]	…	19 (6/32) [7, 36]	…	100 (10 255/10 285) [100, 100]
Deep learning algorithm*	90 (9/10) [55, 100]	.25	97 (9956/10 275), [97, 97]	<.001	3.1 (319/10 275) [2.8, 3.5]	<.001	100 (9956/9957), [100, 100]	.09	2.7 (9/328) [1.3, 5.1]	<.001	97 (9965/10 285) [97, 97]
Matched threshold, 0.847†	70 (7/10) [35, 93]	>.99	100 (10 249/10 275) [100, 100]	NA	0.3 (26/10 275) [0.2, 0.4]	NA	100 (10 249/10 252) [100, 100]	.56	21 (7/33) [9, 39]	.26	100 (10 256/10 285) [100, 100]

FPR=false-positive rate, NA=not applicable, NPV=negative predictive value, PPV=positive predictive value

Table 6: Diagnostic Performance of deep learning algorithm for detection of lung cancers on health screening cohort chest radiographs

Variable	Sensitivity (%)	Specificity (%)	FPR (%)	NPV (%)	PPV (%)	Accuracy (%)
Cancer-positive chest radiographs	40 (39/98) [30, 50]	97 (97 479/100 478) [97, 97]	3 (2999/100 478) [2.9, 3.1]	100 (97 479/97 538) [100, 100]	1.3 (39/3038) [0.9, 1.8]	97 (97 518/100 576) [97, 97]
Visible cancers on chest radiographs	83 (39/47) [69, 92]	97 (97 479/100 478) [97, 97]	3 (2999/100 478) [2.9, 3.1]	100 (97 479/97 487) [100, 100]	1.3 (39/3038) [0.9, 1.8]	97 (97 518/100 525) [97, 97]
Clearly visible cancers on chest radiographs	100 (28/28) [88, 100]	97 (97 479/100 478) [97, 97]	3 (2999/100 478) [2.9, 3.1]	100 (97 479/97 479) [100, 100]	0.9 (28/3027) [0.7, 1.3]	97 (97 507/100 506) [97, 97]

FPR=false-positive rate, NA=not applicable, NPV=negative predictive value, PPV=positive predictive value

The improvement of pulmonary nodule detection, which is a challenging task when using chest radiographs, may help to elevate the role of chest radiographs for the diagnosis of lung cancer. Following this, Yoo H et al. (2020) assessed the performance of a deep learning–based AI algorithm for the detection of pulmonary nodules and lung cancer on chest radiographs using separate training (in-house) and validation (National Lung Screening Trial, NLST) data sets. In this diagnostic study, baseline (T0) posteroanterior chest radiographs from 5,485 participants (full T0 data set) were used to assess lung cancer detection and a subset of 577 of these images (nodule data set) were used to assess nodule detection. The change in the performance of the AI algorithm compared with the performance of the NLST radiologists for the detection of **nodules**, **lung cancer**, and **malignant pulmonary nodules** was analysed.[26, level II-3]

The overall area under the ROC curve (AUROC) of the AI algorithm was 0.93 (95% CI: 0.90 to 0.96) for all **chest radiographs**, 0.99 (95% CI: 0.97 to 1.00) for **digital radiographs**, and 0.86 (95% CI: 0.79 to 0.93) for **computed radiographs**. For **nodule detection** performance, the differences between the AI algorithm and the NLST radiologists in both sensitivity (86.2% [95% CI: 77.8% to 94.6%] versus 87.7% [95% CI: 79.7% to 95.7%]; p=0.80) and specificity (85.0% [95% CI: 81.9% to 88.1%] versus 86.7% [95% CI: 83.8% to 89.7%]; p=0.42) were statistically nonsignificant in all **chest radiographs**. However, the sensitivity and specificity of the AI algorithm were higher compared with those of the NLST radiologists for **digital radiographs** (sensitivity 96.0% [95% CI: 88.3% to 100.0%] versus 88.0% [95% CI: 75.3% to 100.0%], respectively; p=0.32; specificity 93.2% [95% CI: 89.9% to 96.5%] versus 82.8% [95% CI: 77.8% to 87.8%]; p=0.001) but were lower for **computed radiographs** (sensitivity 77.8% [95% CI: 64.2% to 91.4%] versus 86.1% [95% CI: 74.8% to 97.4%]; p=0.37; specificity 78.8% [95% CI: 73.9% to 83.8%] versus 90.4% [95% CI: 86.8% to 94.0%]; p<0.001).[26, level II-3]

For **cancer detection** performance, the sensitivity, specificity, PPV, and NPV of the AI algorithm were 75.0% (95% CI: 62.8% to 87.2%), 83.3% (95% CI: 82.3% to 84.3%), 3.8% (95% CI: 2.6% to 5.0%), and 99.8% (95% CI: 99.6% to 99.9%), respectively, in all **chest radiographs**. **In digital radiographs**, the AI algorithm and the NLST radiologists had similar sensitivity (76.0% [95% CI: 59.3% to 92.7%] versus 80.0% [95% CI: 64.3% to 95.7%], respectively; p=0.65), similar specificity (90.0% [95% CI: 89.7% to 92.2%] versus 91.1% [95% CI: 89.9% to 92.3%]; p=0.82), similar PPV (9.1% [95% CI: 5.2% to 13.0%] versus 9.8% [95% CI: 5.7% to 13.8%]; p=0.62), and similar NPV (99.7% [95% CI: 99.4% to 99.9%] versus 99.7% [95% CI: 99.5% to 100.0%]; p=0.65). In **computed radiographs**, the AI algorithm had lower sensitivity (68.4% [95% CI: 47.5% to 89.3%] versus 89.5% [95% CI: 75.7% to 100.0%]; p=0.10), lower specificity (76.7% [95% CI: 75.2% to 78.3%] versus 91.4% [95% CI: 90.3% to 92.4%]; p<0.001), lower PPV (1.9% [95% CI: 0.9% to 3.0%] versus 6.3% [95% CI: 3.5% to 9.5%]; p<0.001), and similar NPV (99.7% [95% CI: 99.5% to 99.9%] versus 99.9% [95% CI: 99.8% to 100.0%]; p=0.07) compared with the NLST radiologists.[26, level II-3]

For **malignant pulmonary nodules** performance, the sensitivity, specificity, PPV, and NPV of the AI algorithm were 94.1% (95% CI: 86.2% to 100.0%), 83.3% (95% CI: 82.3% to 84.3%), 3.4% (95% CI: 2.2% to 4.5%), and 100.0% (95% CI: 99.9% to 100.0%), respectively, in all **chest radiographs**. In **digital radiographs**, the AI algorithm had higher sensitivity (100.0% [95% CI: 100.0% to 100.0%] versus 94.1% [95% CI: 82.9% to 100.0%]; p=0.32), similar specificity (90.9% [95% CI: 89.6% to 92.1%] versus 91.0% [95% CI: 89.7% to 92.2%]; p=0.91), similar PPV (8.2% [95% CI: 4.4% to 11.9%] versus 7.8% [95% CI: 4.1% to 11.5%]; p=0.65), and similar NPV (100.0% [95% CI: 100.0% to 100.0%] versus 99.9% [95% CI: 99.8% to 100.0%]; p=0.32) as compared with the NLST radiologists. In **computed radiographs**, the AI algorithm had lower sensitivity (85.7% [95% CI: 67.4% to 100.0%] versus 92.9% [95% CI: 79.4% to 100.0%]; p= 0.56), lower specificity (76.7% [95% CI: 75.2% to 78.3%] versus 91.3% [95% CI: 90.2% to 92.3%]; p<0.001), lower PPV (1.8% [95% CI: 0.8% to 2.8%] versus 5.0% [95% CI: 2.3% to 7.6%]; p<0.001), and similar NPV (99.9% [95% CI: 99.8% to 100.0%] versus 100.0% [95% CI: 99.9% to 100.0%]; p=0.48) as compared with the NLST radiologists.[26, level II-3]

The performance of the AI algorithm compared with the NLST radiologists for the detection of all cancers and malignant pulmonary nodules in the nodule data set and the full T0 data set is shown in **Table 7**.

Table 7: Comparison of performance of artificial intelligence algorithm versus National Lung Screening Trial Radiologists

	All images					Digital radiographic images					Computed radiographic images				
				P value					P value					P value	
Variable	AI	NLST nodule	NLST cancer	AI vs NLST nodule	AI vs NLST cancer	AI	NLST nodule	NLST cancer	AI vs NLST nodule	AI vs NLST cancer	AI	NLST nodule	NLST cancer	AI vs NLST nodule	AI vs NLST cancer
Sensitivity (all cancer detection)															
Nodule data set	75.0 (62.8-87.2)	77.1 (65.2-89.0)	85.4 (75.4-95.4)	.78	.13	76.0 (59.3-92.7)	68.0 (49.7-86.3)	80.0 (64.3-95.7)	.41	.65	68.4 (47.5-89.3)	84.2 (67.8-100.0)	89.5 (75.7-100.0)	.26	.10
Full T0 data set	75.0 (62.8-87.2)	77.1 (65.2-89.0)	85.4 (75.4-95.4)	.78	.13	76.0 (59.3-92.7)	68.0 (49.7-86.3)	80.0 (64.3-95.7)	.41	.65	68.4 (47.5-89.3)	84.2 (47.5-89.3)	89.5 (75.7-100.0)	.26	.10
Specificity (all cancer detection)															
Nodule data set	81.7 (78.4-85.0)	83.4 (80.2-86.5)	83.9 (80.8-87.1)	.43	.30	91.0 (87.2-94.7)	80.5 (75.3-85.8)	82.4 (77.3-87.4)	.001	.009	74.7 (69.6-79.8)	85.6 (81.4-89.7)	85.2 (81.0-89.4)	<.001	<.001
Full T0 data set	83.3 (82.3-84.3)	91.2 (90.4-91.9)	91.5 (90.7-92.2)	<.001	<.001	90.0 (89.7-92.2)	90.4 (89.1-91.7)	91.1 (89.9-92.3)	.52	.82	76.7 (75.2-78.3)	91.6 (90.5-92.6)	91.4 (90.3-92.4)	<.001	<.001
Sensitivity (malignant pulmonary nodule detection)															
Nodule data set	94.1 (86.2-100.0)	91.2 (81.6-100.0)	94.1 (86.2-100.0)	.65	>.99	100.0 (100.0-100.0)	88.2 (72.9-100.0)	94.1 (82.9-100.0)	.16	.32	85.7 (67.4-100.0)	92.9 (79.4-100.0)	92.9 (79.4-100.0)	.56	.56
Full T0 data set	94.1 (86.2-100.0)	91.2 (81.6-100.0)	94.1 (86.2-100.0)	.65	>.99	100.0 (100.0-100.0)	88.2 (72.9-100.0)	94.1 (82.0-100.0)	.16	.32	85.7 (67.4-100.0)	92.9 (79.4-100.0)	92.9 (79.4-100.0)	.56	.56
Specificity (malignant pulmonary nodule detection)															
Nodule data set	81.4 (78.1-84.7)	82.7 (79.5-85.9)	82.7 (79.5-85.9)	.56	.56	90.4 (86.6-94.2)	80.3 (75.2-85.5)	81.2 (76.2-86.3)	.002	.005	74.8 (69.8-79.9)	84.8 (80.6-88.9)	84.0 (79.8-88.3)	.002	.003
Full T0 data set	83.3 (82.3-84.3)	91.1 (90.3-91.8)	91.3 (90.6-92.1)	<.001	<.001	90.9 (89.6-92.1)	90.3 (89.1-91.6)	91.0 (89.7-92.2)	.53	.91	76.7 (75.2-78.3)	91.5 (90.4-92.5)	91.3 (90.2-92.3)	<.001	<.001

Abbreviations: AI, artificial intelligence; NLST, National Lung Screening Trial; NLST cancer, National Lung Screening Trial radiologists using cancer label; NLST nodule, National Lung Screening Trial radiologists using nodule label; T0, baseline.

Most early lung cancers present as pulmonary nodules on imaging, but these can be easily missed on chest radiographs. Recently, Homayounieh F et al. (2021) try to assess if AI algorithm can help detect pulmonary nodules on radiographs at different levels of detection difficulty. They included 100 posterior-anterior chest radiographs belonging to 100 adult patients (mean age 55 ± 20 years) from two different sources in 2019 - an ambulatory health care centre in Germany and the Lung Image Database Consortium in the US. Two thoracic radiologists established the ground truth and nine test radiologists from Germany and the US independently reviewed all images in two sessions (unaided and AI-aided mode) with at least a 1-month washout period. The study demonstrated that AI-aided interpretation was associated with significantly improved detection of pulmonary nodules on chest radiographs as compared with unaided interpretation of chest radiographs. The mean sensitivity, specificity, and accuracy for AI-aided interpretation increased by 11% (95% CI: 4% to 17%), 3% (95% CI: −2% to 5%), and 7% (95% CI: 2% to 11%) compared with unaided interpretation (**Table 8**). There were significant differences in performance of senior and junior radiologists. Junior radiologists had a 12% (95% CI: 4% to 19%) improvement for sensitivity for nodule detection with AI-aided interpretation over unaided interpretation. The corresponding improvement for senior radiologists was 9% (95% CI: 0.5% to 17%). On the other hand, senior radiologists (for 3 of 4 senior radiologists) witnessed a larger improvement in specificity with AI-aided interpretation (mean improvement 4%; 95% CI: -2% to 9%) compared with the junior group (for 1 of 5 junior radiologists; mean improvement 1%; 95% CI: -3% to 5%). Both groups had similar improvements in accuracy (6%) with AI-aided interpretation.[27, level II-3]

Table 8: Case-level sensitivity, specificity, and accuracy of unaided and AI-aided interpretation modes for pulmonary nodule detection

	Sensitivity, mean (SE), %			Specificity, mean (SE), %[a]			Accuracy, mean (SE), %		
Readers	Unaided	AI-aided	Change[b]	Unaided	AI-aided	Change[b]	Unaided	AI-aided	Change[b]
J1	46 (7)	62 (7)	16 (7)	98 (2)	96 (3)	−2 (2)	72 (4)	79 (4)	7 (4)
S1	74 (6)	60 (7)	−14 (8)	86 (5)	92 (3)	6 (6)	80 (3)	76 (4)	−4 (4)
J2	30 (7)	72 (6)	42 (7)	96 (3)	96 (3)	0 (4)	63 (5)	84 (4)	21 (4)
S2	58 (7)	62 (7)	4 (8)	88 (4)	92 (4)	4 (4)	73 (5)	77 (4)	4 (4)
J4	54 (7)	58 (7)	4 (6)	82 (6)	96 (3)	14 (6)	68 (4)	77 (5)	9 (4)
S3	30 (7)	30 (7)	0 (6)	94 (3)	98 (2)	4 (4)	62 (5)	64 (5)	2 (4)
J5	46 (7)	54 (7)	8 (7)	94 (4)	94 (3)	0 (4)	70 (5)	74 (4)	4 (4)
J6	36 (7)	32 (7)	−4 (6)	98 (2)	98 (2)	0)	67 (5)	65 (5)	−2 (3)
S4	30 (7)	68 (7)	38 (6)	98 (2)	94 (3)	−4 (3)	64 (4)	81 (4)	17 (4)
Total, mean (95% CI), %	45 (38 to 53)	55 (48 to 63)	11 (4 to 17)	93 (89 to 96)	95 (91 to 99)	3 (−2 to 5)	69 (62 to 77)	75 (70 to 81)	7 (2 to 11)

Koo YH et al. (2021) validated and evaluated the reproducibility of a commercial DCNN algorithm for pulmonary nodules on chest radiographs and compared its performance with two radiology residents and two thoracic radiologists. This retrospective study enrolled 434 chest radiographs (normal to abnormal ratio 246:188) from 378 patients that visited a tertiary hospital. Abnormality assessment (using AUROC) and nodule detection (using JAFROC FOM) were compared among three groups: DCNN only, radiologist without DCNN, and

radiologist with DCNN). A subset of 56 paired cases, having two chest radiographs taken within a 7-day period, were assessed for intraobserver reproducibility using the intraclass correlation coefficient. The AUROC for DCNN chest radiography classification was 0.87 (95% CI: 0.84 to 0.90), the sensitivity was 88.3%, and the specificity was 86.2%. In terms of nodule detection, the JAFROC FOM was 0.926, the sensitivity was 89.1, and the false-positive rate was 0.138. When compared DCNN to radiologists without DCNN (test 2), all radiologists exhibited higher AUROCs than the DCNN (0.93 versus 0.87; p<0.001). When compared radiologists working with and without DCNN data (test 2 versus test 3), all exhibited significant improvements in terms of classifying abnormal chest radiography images (mean AUROC improvement 0.03 [0.96 versus 0.93]) (**Table 9**). For nodule detection rates, JAFORC analyses revealed that all performance improved significantly (mean JAFROC FOM 0.964 versus 0.929; p<0.05). The overall false-positive rate also fell. Sensitivity improved from 88.6 to 93.9 (**Table 10**). The DCNN evaluations of the paired chest radiographs were in good agreement (0.80 95% CI: 0.66 to 0.88); the software is stable and thus engenders user confidence. Radiologist reproducibility also increased when the DCNN data were available (as shown by the intraclass correlation coefficient; p<0.05); the radiologists detected the same nodules in paired chest radiographs.[28, level II-3]

Table 9: Results of radiograph classification at observer performance test

Observer	Test 2			Test 3			Test 2 vs. 3 (*p*)
	AUROC	Sensitivity	Specificity	AUROC	Sensitivity	Specificity	AUROC
Radiology residents							
R1	0.89	87.8	90.2	0.94	90.9	97.9	0.007
R2	0.95	93.6	95.5	0.96	95.7	97.2	0.039
Group	0.92	90.7	92.8	0.95	93.3	97.5	0.001
Thoracic radiologists							
S1	0.93	97.9	89.0	0.97	98.9	95.1	0.0002
S2	0.94	90.4	97.5	0.97	94.7	98.7	0.0024
Group	0.94	94.2	93.3	0.97	96.8	96.9	<0.0001
Overall	0.93	92.4	93.1	0.96	95.1	97.2	<0.0001

AUROC, area under the receiver operating characteristics.

Table 10: Results of nodule detection using JAFROC figures of merit at observer performance test

Observer	Test 2			Test 3			Test 2 vs. 3 (*p*)
	JAFROC	Sensitivity	FP Rate	JAFROC	Sensitivity	FP Rate	JAFROC
R1	0.886	81.6	0.089	0.943	90.1	0.029	0.0001
R2	0.934	90.6	0.041	0.959	95.0	0.034	0.0111
S1	0.959	93.6	0.092	0.982	96.0	0.043	0.0076
S2	0.937	88.6	0.032	0.978	94.6	0.023	0.0003
Total	0.929	88.6	0.063	0.964	93.9	0.032	0.0405

FP, false positive; JAFROC, jackknife alternative free-response receiver operating characteristics.

The present study by Tam D et al. (2021) evaluated the role of an AI algorithm could play in assisting radiologists as the first reader of chest radiographs. The demographic breakdown of the tumour dataset comprised 92 male and 108 female patients (mean age 72.6 ± 10.4 years) and the control set comprised 87 male and 113 female patients (mean age 61.8 ± 15.6 years). Examinations were reviewed by three FRCR radiologists and an AI algorithm to establish performance in tumour identification. Artificial intelligent and radiologist labels were combined retrospectively to simulate the proposed AI triage workflow. Classification performance of the AI algorithm and radiologists was assessed using overall accuracy, sensitivity, specificity, and precision. The performance of three independent FRCR consultant radiologists in identifying lung cancer on this dataset is shown in **Table 11**. The mean accuracy of cancer detection is 87.0% (84.0% to 90.0%) and the overall mean sensitivity to cancer is 78% (69% to 86%). Notably, all radiologists have a low rate of false positives, between one and nine examinations in total (average precision 95.7%). When deployed as a standalone algorithm, the AI algorithm achieved an overall accuracy of 87.0% on this tumour dataset, equivalent to the mean performance of the three reviewing radiologists. The algorithm sensitivity was superior to two of three radiologists at 80.0% whilst specificity was marginally lower than radiologists at 93.0%. There was an increase in false-positive examinations, with an overall precision of 92.0%. The best overall performances were achieved when AI was combined with radiologists, improving average overall scores accuracy and sensitivity for tumour identification by +3.67% and +13.33%, respectively (**Table 12**). False-negative cases, where cancer findings were missed, were reduced by between 15 and 40 cases. Combined performance did show an increase in false positive examinations in all cases, with an average precision change of -5.3% and specificity change of -6.0%.[29, level II-3]

Table 11: Standalone tumour classification performance for radiologists and the AI algorithm

	Accuracy	Sensitivity	Specificity	Precision	True positives	False positives	False negatives
Rad 1	**0.90**	**0.86**	0.94	0.93	**171**	12	27
Rad 2	0.87	0.79	0.95	0.95	157	9	41
Rad 3	0.84	0.69	**0.99**	**0.99**	136	**1**	62
Ai	0.87	0.8	0.93	0.92	159	14	39

The top performance for each metric is highlighted in bold.
Rad, radiologist; AI, artificial intelligence.

Table 12: Tumour classification performance when radiologists are aided by AI

	Accuracy	Sensitivity	Specificity	Precision	True positives	False positives	False negatives
Rad 1 + AI	0.91 (+0.01)	0.94 (+0.08)	0.88 (-0.06)	0.89 (-0.04)	186 (+15)	23 (+11)	12 (-15)
Rad 2 + AI	0.90 (+0.04)	0.91 (+0.12)	0.90 (-0.05)	0.90 (-0.05)	180 (+23)	20 (+11)	18 (-23)
Rad 3 + AI	0.91 (+0.07)	0.89 (+0.20)	0.92 (-0.07)	0.92 (-0.07)	176 (+40)	15 (+14)	22 (-40)

For each metric, the change from standalone radiologist performance is given in brackets.
Rad, radiologist; AI, artificial intelligence.

Ueda D et al. (2021) investigated the performance improvement of physicians with varying levels of chest radiology experience when using a commercially available AI-based CAD software to detect lung cancer nodules on chest radiographs from multiple vendors. This is the first study to evaluate the performance not only of radiologists but also general physicians. Chest radiographs and their corresponding chest CT were retrospectively collected from one institution between July 2017 and June 2018. A total of 312 chest radiographs (59 malignant radiographs from 59 patients and 253 non-malignant radiographs from 253 patients; mean age 59 ± 13 years) were used for the test dataset to examine reader performance. Two author radiologists annotated pathologically proven lung cancer nodules on the chest radiographs while referencing CT. Eighteen readers (nine general physicians and nine radiologists) from nine institutions interpreted the radiographs alone and then reinterpreted them referencing the AI-based CAD output. Given this information, the standalone AI-based CAD sensitivity, specificity, accuracy, PPV, and NPV were 0.66 (0.53-0.78), 0.96 (0.92-0.98), 0.90 (0.86-0.93), 0.78 (0.64-0.88), and 0.92 (0.88-0.95) with mean false positive indications per image of 0.05, respectively. All readers improved their overall performance by referring to the AI-based CAD output. The overall increases for sensitivity, specificity, accuracy, PPV, and NPV were 1.22 (1.14-1.30), 1.00 (1.00-1.01), 1.03 (1.02-1.04), 1.07 (1.03-1.11), and 1.02 (1.01-1.03), respectively (**Table 13**). It can be seen that general physicians benefited more from the use of the AI-based CAD than radiologists did. The performance of general physicians was improved from 0.47 to 0.60 for sensitivity, from 0.96 to 0.97 for specificity, from 0.87 to 0.90 for accuracy, from 0.75 to 0.82 for PPV, and from 0.89 to 0.91 for NPV while the performance of radiologists was improved from 0.51 to 0.60 for sensitivity, from 0.96 to 0.96 for specificity, from 0.87 to 0.90 for accuracy, from 0.76 to 0.80 for PPV, and from 0.89 to 0.91 for NPV.[30, level II-3]

Table 13: Results of readers with and without CAD

	General physicians		Radiologists		Overall		Ratio	95% Confidence Interval		
	Non-CAD	CAD	Non-CAD	CAD	Non-CAD	CAD	(CAD/Non-CAD)	Lower	Upper	*P* value
Sensitivity	0.47	0.60	0.51	0.60	0.49	0.60	1.22	1.14	1.30	< 0.001
Specificity	0.96	0.97	0.96	0.96	0.96	0.97	1.00	1.00	1.01	0.221
Accuracy	0.87	0.90	0.87	0.90	0.87	0.90	1.03	1.02	1.04	< 0.001
Positive Predictive Value	0.75	0.82	0.76	0.80	0.75	0.81	1.07	1.03	1.11	0.002
Negative Predictive Value	0.89	0.91	0.89	0.91	0.89	0.91	1.02	1.01	1.03	< 0.001

CAD: computer-assisted detection

Previously, Yoo H et al. (2020)[26, level II-3] validated the performance of an AI algorithm for the detection of malignant pulmonary nodules in the NLST data set and suggested that AI can help to improve lung cancer detection on chest radiographs, but did not assess the performance improvement of blinded readers with AI-aided interpretation. In this study, they presented the results of eight readers including three radiology residents and five board-certified radiologists. The goal of this study was to assess whether an AI algorithm improves the reader performance for lung cancer detection without increasing unnecessary false-

positive findings on chest radiographs. This reader study included 173 images from 98 cancer-positive patients and 346 images from 196 cancer-negative patients selected from NLST. Artificial intelligence algorithm provided image-level probability of pulmonary nodule or mass on chest radiographs and a heatmap of detected lesions. Each reader reviewed each chest radiography twice: first, without AI and then with AI at least four weeks of wash-out within the two reviews. Reader performance was compared with AUC, sensitivity, specificity, false-positives per image, and rates of chest CT recommendations. The study demonstrated that the average AUC for the detection of visible lung cancer increased significantly for radiology residents with AI compared to that without AI (0.76 [95% CI: 0.67 to 0.86] versus 0.82 [95% CI: 0.75 to 0.89]; p=0.003), but was similar for radiologists (0.82 [95% CI: 0.74 to 0.91] versus 0.84 [95% CI: 0.79 to 0.89]; p=0.24). Compared to that without AI, the average sensitivity increased significantly for radiology residents (0.61 [95% CI: 0.55 to 0.67] versus 0.72 [95% CI: 0.66 to 0.77]; p=0.016), but specificity was similar with AI (0.88 [95% CI: 0.86 to 0.90] versus 0.88 [95% CI: 0.86 to 0.90]; p=0.89). For radiologists, average sensitivity was similar (0.76 [95% CI: 0.72 to 0.81] versus 0.76 [95% CI: 0.72 to 0.81]; p=1.00), but specificity increased with AI (0.79 [95% CI: 0.77 to 0.81] versus 0.86 [95% CI: 0.84 to 0.87]; p< 0.001). Average false-positives per image without and with AI was similar for radiology residents (0.15 [95% CI: 0.11 to 0.18] versus 0.12 [95% CI: 0.09 to 0.16]; p=0.13), but was significantly lower with AI for radiologists (0.24 [95% CI: 0.20 to 0.29] versus 0.17 [95% CI: 0.13 to 0.20]; p<0.001) (**Table 14**). In patients positive for visible lung cancer on chest radiography, the average chest CT recommendation rate increased significantly for residents, but was similar for radiologists without and with AI (54.7% [95% CI: 48.2 to 61.2%] versus 70.2% [95% CI: 64.2 to 76.2%]; p<0.001 for residents and 72.5% [95% CI: 68.0 to 77.1%] versus 73.9% [95% CI: 69.4 to 78.3%]; p=0.68 for radiologists). Conversely, in patients without visible lung cancer, the average chest CT recommendation rate was similar without and with AI for residents, but decreased for radiologists (11.2% [95% CI: 9.6 to 13.1%] versus 9.8% [95% CI: 8.0 to 11.6%]; p=0.32 for residents and 16.4% [95% CI: 14.7 to 18.2%] versus. 11.7% [95% CI: 10.2 to 13.3%]; p<0.001 for radiologists) (**Table 15**). When AI was used as a second reader, residents detected more missed lung cancer present in prior chest radiographs (39% versus 71%, p=0.021) (**Table 16**).[31, level II-3]

Table 14: The performance of readers for the detection of visible lung cancer

Group	AUC			Sensitivity			Specificity			FPPI		
	Without AI	With AI	*p* value	Without AI	With AI	*p* value	Without AI	With AI	*p* value	Without AI	With AI	*p* value
Radiology residents												
1	0.75 (0.69–0.81)	0.85 (0.79–0.90)	< 0.001	0.56 (0.45–0.67)	0.73 (0.63–0.83)	0.024	0.90 (0.87–0.94)	0.92 (0.89–0.95)	0.42	0.09 (0.06–0.12)	0.08 (0.05–0.11)	0.56
2	0.74 (0.67–0.80)	0.79 (0.73–0.85)	0.06	0.57 (0.46–0.69)	0.67 (0.56–0.77)	0.24	0.86 (0.83–0.90)	0.86 (0.83–0.90)	1.00	0.18 (0.14–0.22)	0.15 (0.11–0.19)	0.31
3	0.81 (0.75–0.87)	0.83 (0.78–0.89)	0.38	0.69 (0.59–0.80)	0.75 (0.65–0.85)	0.47	0.86 (0.82–0.90)	0.85 (0.81–0.89)	0.67	0.17 (0.13–0.21)	0.14 (0.10–0.18)	0.33
Average	0.76 (0.67–0.86)	0.82 (0.75–0.89)	0.003	0.61 (0.55–0.67)	0.72 (0.66–0.77)	0.016	0.88 (0.86–0.90)	0.88 (0.86–0.90)	0.89	0.15 (0.11–0.18)	0.12 (0.09–0.16)	0.13
Radiologists												
1	0.90 (0.85–0.94)	0.89 (0.85–0.94)	0.79	0.89 (0.82–0.96)	0.84 (0.76–0.92)	0.34	0.74 (0.70–0.79)	0.85 (0.81–0.88)	<0.001	0.27 (0.23–0.33)	0.18 (0.14–0.22)	0.003
2	0.72 (0.65–0.78)	0.78 (0.72–0.84)	0.02	0.67 (0.56–0.77)	0.71 (0.60–0.81)	0.6	0.66 (0.61–0.71)	0.82 (0.78–0.86)	<0.001	0.46 (0.39–0.52)	0.22 (0.18–0.27)	< 0.001
3	0.82 (0.77–0.88)	0.82 (0.76–0.88)	0.91	0.79 (0.69–0.88)	0.80 (0.71–0.89)	0.84	0.73 (0.69–0.78)	0.73 (0.68–0.77)	0.80	0.29 (0.23–0.34)	0.31 (0.26–0.36)	0.53
4	0.82 (0.77–0.88)	0.86 (0.81–0.91)	0.09	0.68 (0.57–0.79)	0.75 (0.65–0.85)	0.37	0.94 (0.91–0.96)	0.96 (0.94–0.98)	0.24	0.06 (0.04–0.09)	0.06 (0.04–0.08)	0.89
5	0.86 (0.81–0.91)	0.84 (0.79–0.89)	0.48	0.79 (0.69–0.88)	0.72 (0.62–0.82)	0.34	0.87 (0.84–0.91)	0.93 (0.90–0.96)	0.010	0.14 (0.11–0.18)	0.06 (0.04–0.09)	< 0.001
Average	0.82 (0.74–0.91)	0.84 (0.79–0.89)	0.24	0.76 (0.72–0.81)	0.76 (0.72–0.81)	1.00	0.79 (0.77–0.81)	0.86 (0.84–0.87)	<0.001	0.24 (0.20–0.29)	0.17 (0.13–0.20)	< 0.001

Table 15: The percentages of chest CT recommendation

Group	75 CXRs from patients positive for visible lung cancer (n = 68)			346 CXRs from cancer-negative patients (n = 196)		
	Without AI	With AI	*p* value	Without AI	With AI	*p* value
Radiology residents						
1	45.3 (34.1–56.6)	62.7 (51.7–73.6)	0.031	8.7 (5.7–11.6)	5.5 (3.1–7.9)	0.10
2	57.3 (46.1–68.5)	74.7 (64.8–84.5)	0.022	16.5 (12.6–20.4)	14.2 (10.5–17.8)	0.40
3	61.3 (50.3–72.4)	73.3 (63.3–83.3)	0.11	8.4 (5.5–11.3)	9.8 (6.7–13.0)	0.51
Average	54.7 (48.2–61.2)	70.2 (64.2–76.2)	< 0.001	11.2 (9.3–13.1)	9.8 (8.0–11.6)	0.32
Radiologists						
1	82.7 (74.1–91.2)	81.3 (72.5–90.2)	0.83	16.8 (12.8–20.7)	8.4 (5.5–11.3)	< 0.001
2	65.3 (54.6–76.1)	69.3 (58.9–79.8)	0.60	33.5 (28.6–38.5)	17.9 (13.9–22.0)	< 0.001
3	68.0 (57.4–78.6)	73.3 (63.3–83.3)	0.47	13.9 (10.2–17.5)	21.1 (16.8–25.4)	0.012
4	68.0 (57.4–78.6)	74.7 (64.8–84.5)	0.37	6.4 (3.8–8.9)	4.3 (2.2–6.5)	0.24
5	78.7 (69.4–87.9)	70.7 (60.4–81.0)	0.26	11.6 (8.2–14.9)	6.9 (4.3–9.6)	0.04
Average	72.5 (68.0–77.1)	73.9 (69.4–78.3)	0.68	16.4 (14.7–18.2)	11.7 (10.2–13.3)	< 0.001

Table 16: The detection rate and chest CT recommendation rate for lung cancers visible in previous chest radiographs

Group	Detection rate (image-level)			Detection rate (lesion-level)			Chest CT recommendation rate		
	Without AI	With AI	*p* value	Without AI	With AI	*p* value	Without AI	With AI	*p* value
Radiology residents									
1	57 [4/7]	71 [5/7]	0.57	57 [4/7]	71 [5/7]	0.57	43 [3/7]	57 [4/7]	0.59
2	29 [2/7]	71 [5/7]	0.08	29 [2/7]	71 [5/7]	0.08	29 [2/7]	86 [6/7]	0.008
3	29 [2/7]	71 [5/7]	0.08	29 [2/7]	71 [5/7]	0.08	29 [2/7]	71 [5/7]	0.076
Average	39 [2.7/7]	71 [5.0/7]	0.021	39 [2.7/7]	71 [5.0/7]	0.021	33 [2.3/7]	71 [5/7]	0.008
Radiologists									
1	100 [7/7]	57 [4/7]	0.025	100 [7/7]	57 [4/7]	0.025	71 [5/7]	57 [4/7]	0.57
2	29 [2/7]	57 [4/7]	0.26	29 [2/7]	43 [3/7]	0.57	29 [2/7]	57 [4/7]	0.26
3	57 [4/7]	57 [4/7]	1.00	57 [4/7]	57 [4/7]	1.00	57 [4/7]	43 [3/7]	0.59
4	29 [2/7]	43 [3/7]	0.57	29 [2/7]	43 [3/7]	0.57	29 [2/7]	43 [3/7]	0.57
5	71 [5/7]	43 [3/7]	0.26	71 [5/7]	43 [3/7]	0.26	71 [5/7]	43 [3/7]	0.26
Average	57 [4.0/7]	51 [3.6/7]	0.63	57 [4.0/7]	49 [3.4/7]	0.47	51[3.6/7]	49 [3.4/7]	0.81

5.2 SAFETY

There was no retrievable evidence on the adverse events or complications related to the use of artificial intelligence-based chest x-ray for lung cancer screening. No manufacturer field safety notices or medical device alerts for these technologies have been identified. Currently, there are 190 FDA-approved radiology AI-based software devices, specifically to thoracic radiology. The majority of these algorithms are approved for detection and segmentation of pulmonary nodules.[32] In addition, several AI-based chest x-rays were registered as CE-mark (Class IIa) medical device.[21]

5.3 ORGANISATIONAL ISSUES

There was no retrievable evidence in the context of procedural time points and training or learning curve related to artificial intelligence-based chest x-ray for lung cancer screening. Nevertheless, AI-aided seems to help radiologists to read images effectively and providing complementary interpretation, highlighting areas of the scan requiring particularly close inspection. This information may reduce time to diagnosis, help avoid fatigue- or workload-induced missed diagnoses, and can be invaluable in helping address shortage of trained radiologists. An AI-aided diagnosis can improve consistency and reduce interpreter variability in assessing and reporting lung cancer risk among radiologists and pulmonary medicine physicians. Besides, AI can be vital for less experienced or nonspecialized clinicians to help classify the malignancy risk of lung nodules, allowing a quantifiable and reproducible interpretation.

One of the biggest challenges for AI- based chest x-ray is the need for robust interoperability and data sharing infrastructure, which included lack of IT capacity for integration (and potential issues of incompatibility between the AI technology and existing IT infrastructure). At installation, software hosting and bandwidth may need capital investment, and the AI software would need to be effectively integrated into clinical systems (usually PACS or electronic health record). Specific training may be needed to understand limitations and scope of the technology function. Training needs may be fairly modest and depend on product design and radiologist or radiographer experience levels. Other factors or barriers to AI in radiology need to be taken into consideration: real-world validation studies before clinical implementation (validity of training data and its application to the local population), privacy and data protection, safety and regulatory issues.

5.4 ECONOMIC IMPLICATION

There was no retrievable evidence on the cost-effectiveness or other economic analysis related to artificial intelligence-based chest x-ray for lung cancer screening. According to the NICE guideline on artificial intelligence for analysing chest x-ray images, the cost associated to use the technologies generally consists of implementation or integration fee (to connect to the hospital PACS and RIS systems) which ranged from £6,000 to £10,000 per centre, annual licence and maintenance fee (£60,000), and fixed cost per scan processed (between £0.90 and £1.66 per image) which would be in addition to standard care and it depends on the total patient throughput. The unit cost of standard care (direct access x-ray imaging) is £32.73 and includes the cost of reporting (National schedule of NHS costs for 2019 to 2020).[21]

5.5 LIMITATIONS

We acknowledge some limitations in our review and these should be considered when interpreting the results. The selection of the studies and appraisal was done by one reviewer. Although there was no restriction in language during the search, only the full text articles in English published in peer-reviewed journals were included in the report, which may have excluded some relevant articles and further limited our study numbers. Above all, most studies were retrospective using historic data to train algorithms and did not exactly represent the real-world setting. However, the lack of real-world validation is rapidly being addressed with many studies integrating AI validation in their study design. Besides, most of the trials focused on short-term outcomes while the small number of participants with cancer make it difficult to achieve statistical significance for the differences in sensitivity between the AI algorithm and radiologists. The non-randomised nature of such study also limits its validity. Indeed, heterogeneity in algorithms, predictive models, and training datasets affect reproducibility, generalisability, and therefore limit the degree to which results of these technologies can be accurately compared. Standardisation of AI elements including software for analysing chest x-ray images and training, pursuit of randomised multicentre trials comparing directly to radiologist interpretation is all urgently needed.

6.0 CONCLUSION

A substantial body of retrievable evidence suggests that radiologists assisted with artificial intelligence algorithm was associated with improved detection of lung cancer on chest radiographs with a higher sensitivity and fewer false-positive findings per image compared with radiologists alone, irrespective of radiologist experience and nodule characteristics. Indeed, AI can enhance the performance when used as second reader by improving the quality of reading for various reader groups. Since AI-based chest x-ray can help reduce the false-positive rate, it will enable cost-effective lung cancer screening by reducing over-diagnosis and the follow-up costs for additional scans and biopsies of benign nodules. Several factors (robust interoperability, data sharing infrastructure, real-world validation studies, and training) need to be taken into consideration.

7.0 REFERENCES

1. Bray F, Ferlay J, Soerjomataram I et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68(6): 394-424
2. Manser R, Lethaby A, Irving LB et al. Screening for lung cancer. Cochrane Database of Systematic Reviews. 2013; Cd001991
3. Aberle DR, Adams AM, Berg CD et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011; 365(5): 395-409
4. Aberle DR, DeMello S, Berg CD et al. Results of the two incidence screenings in the National Lung Screening Trial. N Engl J Med. 2013; 369(10): 920-931
5. de Hoop B, Schaefer-Prokop C, Gietema HA et al. Screening for lung cancer with digital chest radiography: sensitivity and number of secondary work-up CT examinations. Radiology. 2010; 255(2): 629-637
6. Gavelli G, Giampalma E. Sensitivity and specificity of chest X-ray screening for lung cancer: review article. Cancer. 2000; 89(S11): 2453-2456
7. Potchen EJ, Cooper TG, Sierra AE et al. Measuring performance in chest radiography. Radiology. 2000; 217(2): 456-459
8. Quekel LG, Kessels AG, Goei R et al. Miss rate of lung cancer on the chest radiograph in clinical practice. Chest. 1999; 115(3): 720-724
9. Kakeda S, Moriya J, Sato H et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. AJR Am J Roentgenol. 2004; 182: 505-510
10. Bley TA, Baumann T, Saueressig U et al. Comparison of radiologist and CAD performance in the detection of CT confirmed subtle pulmonary nodules on digital chest radiographs. Invest Radiol. 2008; 43: 343-348
11. Li F, Engelmann R, Metz CE et al. Lung cancers missed on chest radiographs: results obtained with a commercial computer-aided detection program. Radiology. 2008; 246: 273-280
12. de Hoop B, De Boo DW, Gietema HA et al. Computer-aided detection of lung cancer on chest radiographs: effect on observer performance. Radiology. 2010; 257: 532-540
13. Dellios N, Teichgraeber U, Chelaru R et al. Computer-aided detection fidelity of pulmonary nodules in chest radiograph. J Clin Imaging Sci. 2017; 7: 8
14. Schalekamp S, van Ginneken B, Koedam E et al. Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images. Radiology. 2014; 272: 252-261
15. Yang Y, Feng X, Chi W et al. Deep learning aided decision support for pulmonary nodules diagnosing: a review. J Thorac. 2018; 2018: S867-S875
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436-444

17. Hinton G. Deep learning - a technology with the potential to transform health care. JAMA. 2018; 320(11): 1101-1102
18. Ueda D, Shimazaki A, Miki Y. Technical and clinical overview of deep learning in radiology. Jpn J Radiol. 2019; 37(1): 15-33
19. Klang E. Deep learning and medical imaging. J. Thorac. 2018; 10: 1325-1328
20. Lawson CE, Marti JM, Radivojevic T et al. Machine learning for metabolic engineering: A review. Metab. Eng. 2021; 63: 34-60
21. Artificial intelligence for analysing chest X-ray images. Available at www.nice.org.uk/guidance/mib292 Accessed on 22.09.2022
22. Sim Y, Chung MJ, Kotter E et al. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology. 2019; 00: 1-11
23. Cha MJ, Chung MJ, Lee JH et al. Performance of deep learning model in detecting operable lung cancer with chest radiographs. J Thorac Imaging 2019; 34: 86-91
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; 290: 218-228
25. Lee JH, Sun HY, Park S et al. Performance of a deep learning algorithm compared with radiologic interpretation for lung cancer detection on chest radiographs in a health screening population. Radiology. 2020; 00: 1-11
26. Yoo H, Kim KH, Singh R et al. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. JAMA Network Open. 2020; 3(9): e2017135
27. Homayounieh F, Digumarthy S, Ebrahimian S et al. An artificial intelligence-based chest x-ray model on human nodule detection accuracy from a multicenter study. JAMA Network Open. 2021; 4(12): e2141096
28. Koo YH, Shin KE, Park JS et al. Extravalidation and reproducibility results of a commercial deep learning-based automatic detection algorithm for pulmonary nodules on chest radiographs at tertiary hospital. J Med Imaging Radiat Oncol. 2021; 65(1): 15-22
29. Tam D, Dissez G, Morgan T et al. Augmenting lung cancer diagnosis on chest radiographs: positioning artificial intelligence to improve radiologist performance. Clin Radiol. 2021; 76(8): 607-614
30. Ueda D, Yamamoto A, Shimazaki A et al. Artificial intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: a retrospective clinical validation study. BMC Cancer. 2021; 21: 1120
31. Yoo H, Lee SH, Arru CD et al. AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset. Eur Radiol. 2021; 31(12): 9664-9674
32. Milam ME and Koo CW. The current status and future of FDA-approved artificial intelligence tools in chest radiology in the United States. Available at https://doi.org/10.1016/j.crad.2022.08.135 Accessed on 06.10.2022

8.0 APPENDIX

APPENDIX 1: LITERATURE SEARCH STRATEGY

OVID MEDLINE® ALL 1946 to Jul 29, 2022.

1. ADULT/
2. Adult*.tw.
3. LUNG NEOPLASMS/
4. Cancer of lung.tw.
5. Cancer of the lung.tw.
6. (lung adj3 (cancer* or neoplasm* or carcinoma or tumo?r*)).tw.
7. (pulmonary adj1 (cancer* or neoplasm*)).tw.
8. RADIOGRAPHY, THORACIC/
9. (thoracic adj1 radiograph*).tw.
10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9
11. ARTIFICIAL INTELLIGENCE/
12. ai.tw.
13. (artificial adj1 intelligence).tw.
14. (knowledge adj1 acquisition).tw.
15. (computational adj1 intelligence).tw.
16. (computer adj1 reasoning).tw.
17. (computer adj2 vision system).tw.
18. (intelligence adj1 machine).tw.
19. (knowledge adj1 representation*).tw.
20. DIAGNOSIS, COMPUTER-ASSISTED/
21. (computer assisted adj1 diagnos*).tw.
22. ALGORITHMS/
23. algorithm*.tw.
24. MACHINE LEARNING/
25. (machine adj1 learning).tw.
26. (transfer adj1 learning).tw.
27. IMAGE INTERPRETATION, COMPUTER-ASSISTED/
28. (computer-assisted adj2 image interpretation*).tw.
29. 1 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28
30. RADIOLOGISTS/
31. radiologist*.tw.
32. 30 or 31
33. 10 and 29 and 32
34. limit 33 to (English language and humans and yr="2000 -Current")

Other Databases

PubMed
INAHTA
US FDA

} Same MeSH and keywords as per MEDLINE search

APPENDIX 2: HIERARCHY OF EVIDENCE FOR EFFECTIVENESS

DESIGNATION OF LEVELS OF EVIDENCE

I	Evidence obtained from at least one properly designed randomized controlled trial.
II-I	Evidence obtained from well-designed controlled trials without randomization.
II-2	Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one centre or research group.
II-3	Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.
III	Opinions or respected authorities, based on clinical experience; descriptive studies and case reports; or reports of expert committees.

SOURCE: US/CANADIAN PREVENTIVE SERVICES TASK FORCE (Harris 2001)

APPENDIX 3: HIERARCHY OF EVIDENCE FOR TEST ACCURACY STUDIES

Level	Description	
1.	A blind comparison with reference standard among an appropriate sample of consecutive patients	
2.	Any one of the following	Narrow population spectrums
3.	Any two of the following	Differential use of reference standard
4.	Any three or more of the following	Reference standard not blind
		Case control study
5.	Expert opinion with no explicit critical appraisal, based on physiology, bench research or first principles.	

SOURCE: NHS Centre for Reviews and Dissemination (CRD) University of York, Report Number 4 (2nd Edition)

APPENDIX 4: EVIDENCE TABLE

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
1. Sim Y, Chung MJ, Kotter E et al. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology. 2019; 00: 1-11	**Diagnostic test study** To evaluate the clinical efficacy of software using a deep convolutional neural network (DCNN) algorithm for the detection of malignant pulmonary nodules on chest radiographs (CXRs). Twelve radiologists from the four centres independently analysed the CXRs and marked regions of interest. Commercially available deep learning-based computer-aided detection software separately trained, tested, and validated with 19,330 radiographs was used to find suspicious nodules. The radiologists then reviewed the images with the assistance of DCNN software. The standard of reference for malignant nodules was CT with pathologic confirmation. The sensitivity and number of false-positive findings per image of DCNN software, radiologists alone, and radiologists with the use of DCNN software were analysed by using logistic regression and Poisson regression.	II-3	A total of 800 radiographs were included in the study; 200 normal radiographs (mean age 52.4 ± 12.2 years) and 600 lung cancer-containing radiographs (mean age 61.5 ± 12.3 years) from the four centres.	Readers without DCNN Readers with DCNN DCNN alone	Between intervention	-	The average sensitivity of radiologists improved from 65.1% (1,375 of 2,112; 95% confidence interval {CI]: 62.0% to 68.1%) to 70.3% (1,484 of 2,112; 95% CI: 67.2% to 73.1%; p<0.001) and the number of false-positive findings per radiograph declined from 0.2 (488 of 2,400; 95% CI: 0.18 to 0.22] to 0.18 (422 of 2,400; 95% CI: 0.16 to 0.2]; p<0.001) when the radiologists re-reviewed radiographs with the DCNN software. The stand-alone sensitivities of the DCNN and the number of false-positive findings per image were similar to those of the readers when analysed by centre. For the 12 radiologists in this study, 104 of 2,400 radiographs were positively changed (from false-negative to true-positive or from false-positive to true-negative) using the DCNN, while 56 of 2,400 radiographs were changed negatively.	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
2. Cha MJ, Chung MJ, Lee JH et al. Performance of deep learning model in detecting operable lung cancer with chest radiographs. J Thorac Imaging 2019; 34: 86-91	**Diagnostic test study** To evaluate the diagnostic performance of a trained deep convolutional neural network (DCNN) model for detecting operable lung cancer with chest radiographs (CXRs). A deep learning model (DLM) based on DCNN was trained with 17,211 CXRs (5700 CT-confirmed lung nodules in 3,500 CXRs and 13,711 normal CXRs), finally augmented to 600,000 images. For validation, a trained DLM was tested with 1,483 CXRs with surgically resected lung cancer, marked and scored by 2 radiologists. The reference standard for the location and margin of each lesion was based on the review of CT. Furthermore, diagnostic performances of DLM and 6 human observers were compared with 500 cases (200 visible T1 lung cancer and 300 normal on CXRs) and analysed using free response receiver-operating characteristics curve (FROC) analysis.	II-3	For the test set, they identified 2,509 consecutive patients with a surgically proven lung cancer (Male: Female 1,167:1,342; mean age, 60 years) whereby 1,483 cases were finally included. The mean diameter of 1,483 resected lung cancers was 27.2 $\pm$ 14.6 mm with a range of 10 to 109 mm.	Deep convolutional neural network (DCNN)	Six radiologists (with experience for chest radiology from 4 to 21 years) participated as human observers.	-	The overall detection rate of DLM for all operable lung cancers was a sensitivity of 76.8% (1,139/1,483) with a false positive per image (FPPI) of 0.3 and area under the FROC curve (AUC) of 0.732. In the comparison with human readers, DLM demonstrated a sensitivity of 86.5% at 0.1 FPPI and a sensitivity of 92.0% at 0.3 FPPI with AUC of 0.899 at an FPPI range of 0.03 to 0.44 for detecting visible T1 lung cancers, which were superior to the average of 6 human readers (mean sensitivity; 78.0% [range 71.6% to 82.6%] at an FPPI of 0.1% and 85% [range 80.2% to 89.2%] at an FPPI of 0.3, AUC of 0.819 [range, 0.754 to 0.862] at an FPPI of 0.03 to 0.44).	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
3. 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; 290: 218-228	**Diagnostic test study** To develop and validate a deep learning–based automatic detection algorithm (DLAD) for malignant pulmonary nodules on CXRs and to compare its performance with physicians including thoracic radiologists. Radiograph classification and nodule detection performances of DLAD were validated by using one internal and four external data sets from three South Korean hospitals and one U.S. hospital. For internal and external validation, radiograph classification and nodule detection performances of DLAD were evaluated by using the area under the receiver operating characteristic curve (AUROC) and jackknife alternative free-response receiver-operating characteristic (JAFROC) figure of merit (FOM), respectively. An observer performance test involving 18 physicians, including nine board-certified radiologists, was conducted by using one of the four external validation data sets. Performances of DLAD, physicians, and physicians assisted with DLAD were evaluated and compared.	II-3	DLAD was developed by using 43,292 chest radiographs (normal radiograph-to-nodule radiograph ratio, 34,067:9,225) in 34,676 patients (healthy-to-nodule ratio, 30,784:3,892; 19,230 men [mean age, 52.8 years; age range, 18-99 years]; 15 446 women [mean age, 52.3 years; age range, 18–98 years]) obtained between 2010 and 2015, which were labelled and partially annotated by 13 board-certified radiologists, in a convolutional neural network.	DLAD (DCNN) Physicians Physicians assisted with DLAD (DCNN)	Between intervention	-	According to one internal and four external validation data sets, radiograph classification and nodule detection performances of DLAD were a range of 0.92-0.99 (AUROC) and 0.831-0.924 (JAFROC FOM), respectively. **Comparison of radiograph classification and nodule detection performances of DLAD** Regarding radiograph classification performance comparison, DLAD showed an excellent AUROC of 0.91, which was higher than 16 of 18 physicians (p value range <0.001 to 0.52), with statistical significance demonstrated in 11 physicians. Regarding nodule detection performance comparison, DLAD exhibited a JAFROC FOM of 0.885, higher than all physicians and significantly higher than those of 15 of 18 physicians ($p<0.05$). Regarding the added value of DLAD as a second reader, the radiograph classification performance of physicians improved with DLAD (17 of 18 physicians; mean AUROC improvement of 0.04 [range -0.0001 to 0.14]) and the changes were statistically significant in 15 of 18 physicians. Regarding nodule detection performance, all physicians showed improved detection performances by using DLAD (mean JAFROC FOM improvement 0.043 [range 0.006-0.190]) with significant differences in 14 physicians. JAFROC FOM of non-radiology physicians, radiology residents, board-certified radiologists, and thoracic radiologists' subgroups were 0.691, 0.796, 0.821, and 0.833, respectively, and in all four subgroups, their nodule detection performances improved with DLAD (JAFROC FOM for non-radiology physicians, radiology residents, board-certified radiologists, and thoracic radiologists, respectively: 0.828, 0.829, 0.840, and 0.854; corrected $p<0.05$).	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
4. Lee JH, Sun HY, Park S et al. Performance of a deep learning algorithm compared with radiologic interpretation for lung cancer detection on chest radiographs in a health screening population. Radiology. 2020; 00: 1-11	**Diagnostic test study** To evaluate the performance of a deep learning algorithm for lung cancer nodule detection on CXRs, in a healthy screening population with average risk for lung cancer, compared with radiologic interpretation. Out-of-sample testing of a deep learning algorithm was retrospectively performed using CXRs from individuals undergoing a comprehensive medical check-up between July 2008 and December 2008 (validation test). To evaluate the algorithm performance for visible lung cancer detection, the area under the receiver operating characteristic curve (AUC) and diagnostic measures, including sensitivity and false-positive rate (FPR), were calculated. The algorithm performance was compared with that of radiologists using the McNemar test and the Moskowitz method. Additionally, the deep learning algorithm was applied to a screening cohort undergoing CXR between January 2008 and December 2012, and its performances were calculated.	II-3	For the validation test, 10,206 individuals (5,859 men and 4,347 women; mean age 54 ± 11 years) with 10,289 CXRs were included in the task of detecting cancer-positive radiographs. For detecting visible lung cancer, 10,285 radiographs from 10,202 individuals were used. For the screening cohort, 50,098 individuals (28,105 men and 21,993 women; mean age 53 ± 11 years) with 100,576 radiographs were included in the task of detecting cancer-positive radiographs. For detecting visible lung cancer, 50,070 radiographs from 100,525 individuals were used.	A deep learning algorithm (Lunit INSIGHT CXR) (DCNN)	Radiologist interpretation	-	**Lung cancer detection performance in the validation test** For the detection performances of the deep learning algorithm and pooled radiologists for visible lung cancers on CXRs, the algorithm's AUC was 0.99 (95% CI: 0.97 to 1), and it detected three more lung cancers (9/10 chest radiographs; sensitivity 90%) than the radiologists (6/10 chest radiographs; sensitivity 60%). However, this difference was not statistically significant ($p=0.25$). The algorithm had an NPV equivalent to that of the radiologists (100% vs. 100%; $p=0.09$), but it had a lower specificity and PPV and a higher FPR (specificity 97% vs. 100%; $p<0.001$; PPV 2.7% vs. 19%; $p<0.001$; FPR 3.1% vs. 0.3%; $p<0.001$). **Lung cancer detection performance of the deep learning algorithm in the entire screening cohort** In the detection of visible lung cancers on the CXR, the algorithm had an AUC of 0.97 (95% CI: 0.95 to 0.99). Visible lung cancers were correctly detected on 39 of 47 radiographs (sensitivity 83%). The specificity, NPV, PPV, and FPR of the algorithm for detecting visible lung cancers were 97%, 100%, 1.3%, and 3%, respectively. **Conclusion:** A deep learning algorithm detected lung cancers on CXRs with a performance comparable to that of radiologists, which will be helpful for radiologists in healthy populations with a low prevalence of lung cancer.	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
5. Yoo H, Kim KH, Singh R et al. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. JAMA Network Open. 2020; 3(9): e2017135	**Diagnostic test study** This study assessed the performance of a deep learning–based AI algorithm for the detection of pulmonary nodules and lung cancer on CXRs using separate training (in-house) and validation (National Lung Screening Trial, NLST) data sets. The change in the performance of the AI algorithm compared with the performance of the NLST radiologists for the detection of nodules, lung cancer, and malignant pulmonary nodules was analysed.	II-3	Baseline (T0) posteroanterior chest radiographs from 5,485 participants (full T0 data set) were used to assess lung cancer detection performance, and a subset of 577 of these images (nodule data set) were used to assess nodule detection performance. Participants aged 55 to 74 years who currently or formerly smoked cigarettes for 30 pack-years or more were enrolled in the NLST at 23 US centres between August 2002 and April 2004. Information on lung cancer diagnoses was collected through December 31, 2009. Analyses were performed between August 20, 2019, and February 14, 2020.	Deep learning-based AI algorithm (Lunit INSIGHT CXR) (DCNN)	NLST radiologists' interpretation	Median follow-up duration of 6.5 years (interquartile range, 6.1-6.9 years).	**Nodule Detection Performance** The area under the ROC curve (AUROC) of the AI algorithm was 0.93 (95% CI: 0.90 to 0.96) for all CXRs, 0.99 (95% CI: 0.97 to 1.00) for digital radiographs, and 0.86 (95% CI: 0.79 to 0.93) for computed radiographs. The differences between the AI algorithm and the NLST radiologists in both sensitivity (86.2% [95% CI: 77.8% to 94.6%] vs. 87.7% [95% CI: 79.7% to 95.7%], respectively; p=0.80) and specificity (85.0% [95% CI: 81.9% to 88.1%] vs. 86.7% [95% CI: 83.8% to 89.7%]; p=0.42) were statistically nonsignificant in all CXRs at the operating point chosen from the internal validation set. The sensitivity and specificity of the AI algorithm were higher compared with those of the NLST radiologists for digital radiographs (sensitivity 96.0% [95% CI: 88.3% to 100.0%] vs. 88.0% [95% CI: 75.3% to 100.0%], respectively; p=0.32; specificity 93.2% [95% CI: 89.9% to 96.5%] vs. 82.8% [95% CI: 77.8% to 87.8%]; p= 0.001) but were lower compared with those of the NLST radiologists for computed radiographs (sensitivity 77.8% [95% CI: 64.2% to 91.4%] vs. 86.1% [95% CI: 74.8% to 97.4%]; p=0.37; specificity 78.8% [95% CI: 73.9% to 83.8%] vs. 90.4% [95% CI: 86.8% to 94.0%]; p<0.001). Of the 65 total non-calcified nodules or masses present in the nodule data set, 56 nodules or masses were detected by the AI algorithm (including 7 nodules or masses that were missed by NLST radiologists), 57 nodules or masses were detected by NLST radiologists (including 8 nodules or masses that were missed by the AI algorithm), 49 nodules or masses were detected by both, and 1 nodule or mass was missed by both.	

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
							Cancer Detection Performance The sensitivity, specificity, PPV, and NPV of the AI algorithm were 75.0% (95% CI: 62.8% to 87.2%), 83.3% (95% CI: 82.3% to 84.3%), 3.8% (95% CI: 2.6% to 5.0%), and 99.8% (95% CI: 99.6% to 99.9%) for the detection of all cancers in all CXRs of the full T0 data set. In digital radiographs of the full T0 data set, the AI algorithm and the NLST radiologists (as assessed by the cancer label) had similar sensitivity (76.0% [95% CI: 59.3% to 92.7%] vs. 80.0% [95% CI: 64.3% to 95.7%], respectively; p=0.65), similar specificity (90.0% [95% CI: 89.7% to 92.2%] vs. 91.1% [95% CI: 89.9% to 92.3%]; p=0.82), similar PPV (9.1% [95% CI: 5.2% to 13.0%] vs. 9.8% [95% CI: 5.7% to 13.8%]; p=0.62), and similar NPV (99.7% [95% CI: 99.4% to 99.9%] vs. 99.7% [95% CI: 99.5% to 100.0%]; p=0.65) for cancer detection. In computed radiographs of the full T0 data set, the AI algorithm had lower sensitivity (68.4% [95% CI: 47.5% to 89.3%] vs. 89.5% [95% CI: 75.7% to 100.0%]; p=0.10), lower specificity (76.7% [95% CI: 75.2% to 78.3%] vs. 91.4% [95% CI: 90.3% to 92.4%]; p<0.001), lower PPV (1.9% [95% CI: 0.9% to 3.0%] vs. 6.3% [95% CI: 3.5% to 9.5%]; p<0.001), and similar NPV (99.7% [95% CI: 99.5% to 99.9%] vs. 99.9% [95% CI: 99.8% to 100.0%]; p=0.07) compared with the NLST radiologists. Among all images of the 48 participants who received lung cancer diagnoses within 1 year of screening, 36 cases were detected by the AI algorithm, 41 cases were detected by the NLST radiologists, 33 cases were detected by both, and 4 cases were missed by both. In all radiographs of the full T0 data set, the sensitivity, specificity, PPV, and NPV of the AI algorithm were 94.1% (95% CI: 86.2% to 100.0%), 83.3% (95% CI: 82.3% to 84.3%), 3.4% (95% CI: 2.2% to 4.5%), and 100.0% (95% CI: 99.9% to 100.0%), respectively, for the detection of malignant pulmonary nodules. In digital radiographs of the full T0 data set, the AI algorithm had higher sensitivity (100.0% [95% CI: 100.0% to 100.0%] vs. 94.1% [95% CI: 82.9% to 100.0%]; p=0.32), similar specificity (90.9% [95% CI:	

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
							89.6% to 92.1%] vs. 91.0% [95% CI: 89.7% to 92.2%]; p=0.91), similar PPV (8.2% [95% CI: 4.4% to 11.9%] vs. 7.8% [95% CI: 4.1% to 11.5%]; p=0.65), and similar NPV (100.0% [95% CI: 100.0% to 100.0%] vs. 99.9% [95% CI: 99.8% to 100.0%]; p=0.32) compared with the NLST radiologists (as assessed by the cancer label). In computed radiographs of the full T0 data set, the AI algorithm had lower sensitivity (85.7% [95% CI: 67.4% to 100.0%] vs. 92.9% [95% CI: 79.4% to 100.0%]; p= 0.56), lower specificity (76.7% [95% CI: 75.2% to 78.3%] vs. 91.3% [95% CI: 90.2% to 92.3%]; p<0.001), lower PPV (1.8% [95% CI: 0.8% to 2.8%] vs. 5.0% [95% CI: 2.3% to 7.6%]; p<0.001), and similar NPV (99.9% [95% CI: 99.8% to 100.0%] vs. 100.0% [95% CI: 99.9% to 100.0%]; p=0.48) compared with the NLST radiologists. In all images of the 34 patients with malignant pulmonary nodules who received lung cancer diagnoses within 1 year after imaging, 32 cases were detected by the AI algorithm, 32 cases were detected by the NLST radiologists, 30 cases were detected by both, and 0 cases were missed by both.	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
6. Homayounieh F, Digumarthy S, Ebrahimian S et al. An artificial intelligence-based chest x-ray model on human nodule detection accuracy from a multicenter study. JAMA Network Open. 2021; 4(12): e2141096	**Diagnostic test study** To assess if a novel artificial intelligence (AI) algorithm can help detect pulmonary nodules on radiographs at different levels of detection difficulty. All images were processed with a novel AI algorithm, the AI Rad Companion Chest X-ray. Two thoracic radiologists established the ground truth and nine test radiologists from Germany and the US independently reviewed all images in two sessions (unaided and AI-aided mode) with at least a 1-month washout period. Each test radiologist recorded the presence of five findings (pulmonary nodules, atelectasis, consolidation, pneumothorax, and pleural effusion) and their level of confidence for detecting the individual finding on a scale of 1 to 10 (1 representing lowest confidence; 10, highest confidence). The analysed metrics for nodules included sensitivity, specificity, accuracy, and receiver operating characteristics curve area under the curve (AUC).	II-3	The study included 100 posterior-anterior (PA) CXRs belonging to 100 adult patients (mean age 55 ± 20 years; 64 men [64%], 36 women [36%]) from two different sources in 2019 - an ambulatory health care centre in Germany (site A) and the Lung Image Database Consortium in the US (site B). Included images were selected to represent nodules with different levels of detection difficulties (from easy to difficult), and comprised both normal and nonnormal control.	AI Rad Companion Chest X-Ray algorithm (Siemens Healthineers AG)	AI-aided versus AI-unaided between senior and junior radiologists' interpretations	-	The distribution of findings in the 100 images was 50% (50/100) pulmonary nodules, 12% (12/100) consolidation, 13% (13/100) atelectasis, and 10% (10/100) pleural effusion. A quarter of images (25 of 100) had no abnormal findings. **Sensitivity, specificity, and accuracy** The standalone performance of the AI for detection of pulmonary nodules included 32 true positive, 46 true negative, 18 false negative and 4 false positive identifications. The mean sensitivity, specificity, and accuracy for AI-aided interpretation increased by 11% (95% CI: 4% to 7%), 3% (95% CI: −2% to 5%), and 7% (95% CI: 2% to 11%) compared with unaided interpretation. **AUC analysis** Partial AUC within the effective interval range of 0 to 0.2 false positive rate improved by 5.6% (95% CI: -1.4 to 12.0%) with AI-aided interpretation (mean AUC 77.2%; 95% CI: 70.2% to 83.6%) compared with unaided interpretation (mean AUC 71.7%; 95% CI: 64.0% to 79.8%). **Junior versus senior readers** There were significant differences in performance of senior and junior radiologists. Junior radiologists had a 12% (95% CI: 4% to 19%) improvement for sensitivity for nodule detection with AI-aided interpretation over unaided interpretation. The corresponding improvement for senior radiologists was 9% (95% CI: 0.5% to 17%). On the other hand, senior radiologists (for 3 of 4 senior radiologists) witnessed a larger improvement in specificity with AI-aided interpretation (mean improvement 4%; 95% CI: -2% to 9%) compared with the junior group (for 1 of 5 junior radiologists; mean improvement 1%; 95% CI: -3% to 5%). Both groups had similar improvements in accuracy (6%) and partial AUC (5%) with AI-aided interpretation.	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
7. Koo YH, Shin KE, Park JS et al. Extravalidation and reproducibility results of a commercial deep learning-based automatic detection algorithm for pulmonary nodules on chest radiographs at tertiary hospital. J Med Imaging Radiat Oncol. 2021; 65(1): 15-22	**Diagnostic test study** To extra validate and evaluate the reproducibility of a commercial deep convolutional neural network (DCNN) algorithm for pulmonary nodules on CXRs and to compare its performance with radiologists. A DCNN performance was compared with two radiology residents and two thoracic radiologists. Abnormality assessment (using AUROC) and nodule detection (using JAFROC) were compared among three groups (DCNN only, radiologist without DCNN, and radiologist with DCNN). A subset of 56 paired cases, having two CXRs taken within a 7-day period, were assessed for intraobserver reproducibility using the intraclass correlation coefficient. Independent characteristics of pulmonary nodules detected by DCNN were assessed by multiple logistic regression analysis.	II-3	This retrospective study enrolled 434 CXRs (normal to abnormal ratio 246:188) from 378 patients that visited a tertiary hospital. Images with pulmonary nodules included 117 males and 71 females (mean age 63.9 ± 13.0 years); those without pulmonary nodules included 150 males and 96 females (mean age 59.8 ± 9.3 years). The total number of nodules was 202: 175 images had one nodule, 12 two and 1 three. There were 37 and 165 benign and malignant nodules, respectively. Adenocarcinomas (111, 67.3%) and tuberculosis (17, 45.9%) were the most common pathologies of malignant and benign nodules, respectively.	DCNN only Radiologist without DCNN Radiologist with DCNN	Between intervention	-	**External validation of the DCNN** The AUROC for DCNN CXR classification was 0.872 (95% CI: 0.838 to 0.903), the sensitivity was 88.3%, and the specificity was 86.18%. In terms of nodule detection, the JAFROC FOM was 0.926, the sensitivity was 89.1, and the false-positive rate was 0.138. **CXR classification and nodule detection by radiologists** When compared DCNN with radiologists without DCNN, all radiologists exhibited higher AUROCs than the DCNN ($p<0.001$). When compared radiologists working with and without DCNN data, all exhibited significant improvements in terms of classifying abnormal CXR images ($p=0.0002$ to 0.039; mean AUROC improvement 0.03 [range 0.01–0.05]). For nodule detection rates, JAFORC analyses revealed that all performance improved significantly (mean JAFROC FOM difference 0.035 (range 0.023 to 0.057), 0.929 vs. 0.964, $p<0.05$). The overall false-positive rate also fell. Sensitivity improved from 88.6 to 93.9 ($p=0.000$). **Reproducibility of the DCNN and radiologist** The DCNN evaluations of the paired CXRs were in good agreement (0.80 95% CI: 0.66 to 0.88); the software is stable and thus engenders user confidence. Radiologist reproducibility also increased when the DCNN data were available (as shown by the ICCs; $p<0.05$); the radiologists detected the same nodules in paired CXRs. The average time required to interpret the 434 CXRs was 211.25 ± 38.4 minutes without DCNN assistance and 171 ± 33.8 minutes with assistance ($p=0.068$).	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
8. Tam D, Dissez G, Morgan T et al. Augmenting lung cancer diagnosis on chest radiographs: positioning artificial intelligence to improve radiologist performance. Clin Radiol. 2021; 76(8): 607-614	**Diagnostic test study** To evaluate the role that artificial intelligence (AI) could play in assisting radiologists as the first reader of CXRs, to increase the accuracy and efficiency of lung cancer diagnosis by flagging positive cases before passing the remaining examinations to standard reporting. Examinations were reviewed by three FRCR radiologists and an AI algorithm to establish performance in tumour identification. AI and radiologist labels were combined retrospectively to simulate the proposed AI triage workflow. Classification performance of the AI algorithm and radiologists was assessed using overall accuracy, sensitivity, specificity, and precision. Further to this, agreement between radiologists was assessed.	II-3	The demographic breakdown of the tumour set comprised 92 male and 108 female patients (mean age 72.6 ± 10.4 years) and the control set comprised 87 male and 113 female patients (mean age 61.8 ± 15.6 years). Four were discarded due to incomplete data, leaving 396 examinations (198 positive for tumours, 198 negative).	AI classification Radiologists Radiologists with AI	Between intervention	-	**Radiologist performance** The mean accuracy of cancer detection is 87% (84% to 90%) and the overall mean sensitivity to cancer is 78% (69% to 86%). On a patient level, these performances correspond to between 136 and 171 patients being diagnosed correctly for tumours and between 62 and 27 patients with missed cancer pathologies. Notably, all radiologists have a low rate of false positives, between one and nine examinations in total (average precision 95.67%). Analysis of the statistical correlation between the radiologists' reports shows an average observed proportional agreement of 86.7% and corresponding average Cohen's kappa score of 0.72, denoting good overall agreement. **AI performance** When deployed as a standalone algorithm, the AI algorithm achieved an overall accuracy of 87% on this tumour dataset, equivalent to the mean performance of the three reviewing radiologists. The algorithm sensitivity was superior to two of three radiologists at 80% whilst specificity was marginally lower than radiologists at 93%. There was an increase in false-positive examinations, with an overall precision of 92%. **Radiologists plus AI** For all radiologists, overall accuracy and sensitivity for tumour identification were increased by combination with AI, improving average scores by +3.67% and +13.33%, respectively. False-negative cases, where cancer findings were missed, were reduced by between 15 and 40 cases. Combined performance did show an increase in false positive examinations in all cases, with an average precision change of -5.33% and specificity change of -6%. For all radiologists, improvements in combined scores were shown to be statistically significant when compared to their standalone performance ($p<0.05$). Furthermore, agreement between radiologists was improved when combined with AI, with all radiologist plus AI labels agreeing in 92% of cases (+12%). Average proportional agreement increased to 94.33% (+7.63%) and the average Cohen's Kappa score was 0.89 (+0.17), suggesting very good agreement. On average, cases of missed tumours were reduced by 60% by combination of a single radiologist with AI.	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
9. Ueda D, Yamamoto A, Shimazaki A et al. Artificial intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: a retrospective clinical validation study. BMC Cancer. 2021; 21: 1120	**Diagnostic test study** The purpose of the study was to validate a commercially available artificial intelligence (AI)-based computer-assisted detection (CAD) that achieved higher performance in detecting lung cancer from CXRs. To investigate the ability of this CAD as a support tool, they conducted a multi-vendor, retrospective reader performance test comparing both radiologist and general physicians' performance before and after using the CAD. Chest radiographs and their corresponding chest CT were retrospectively collected from one institution between July 2017 and June 2018. Two author radiologists annotated pathologically proven lung cancer nodules on the CXRs while referencing CT. Eighteen readers (nine general physicians and nine radiologists) from nine institutions interpreted the chest radiographs. The readers interpreted the radiographs alone and then reinterpreted them referencing the CAD output. The sensitivity, specificity, accuracy, PPV, and NPV of the readers' assessments were calculated.	II-3	A total of 312 radiographs (59 malignant radiographs from 59 patients and 253 non-malignant radiographs from 253 patients; mean age 59 ± 13 years) were used for the test dataset to examine reader performance.	AI-based computer-assisted detection (CAD)	A total of 18 readers (nine general physicians and nine radiologists)	-	**The deep learning-based CAD model performance** The standalone CAD sensitivity, specificity, accuracy, PPV, and NPV were 0.66 (0.53-0.78), 0.96 (0.92-0.98), 0.90 (0.86-0.93), 0.78 (0.64-0.88), and 0.92 (0.88-0.95) with mean false positive indications per image of 0.05, respectively. **Reader performance test** All readers improved their overall performance by referring to the CAD output. The overall increases for reader performance due to using the CAD for sensitivity, specificity, accuracy, PPV, and NPV were 1.22 (1.14-1.30), 1.00 (1.00-1.01), 1.03 (1.02-1.04), 1.07 (1.03-1.11), and 1.02 (1.01-1.03), respectively. General physicians benefited more from the use of the CAD than radiologists did. The performance of general physicians was improved from 0.47 to 0.60 for sensitivity, from 0.96 to 0.97 for specificity, from 0.87 to 0.90 for accuracy, from 0.75 to 0.82 for PPV, and from 0.89 to 0.91 for NPV while the performance of radiologists was improved from 0.51 to 0.60 for sensitivity, from 0.96 to 0.96 for specificity, from 0.87 to 0.90 for accuracy, from 0.76 to 0.80 for PPV, and from 0.89 to 0.91 for NPV. The rate of improvement was particularly high for general physicians. They were more likely to change their assessment from FN to TP by referencing correct positive CAD output (68 times [0.59] in general physicians, 49 [0.49] in radiologists) and from FP to TN by correct negative CAD output (29 times [0.36] in general physicians, 24 times [0.29] in radiologists). Radiologists were less likely to change their opinion than general physicians, and it was more difficult for radiologists to change their decisions from FP to TN (24 times) than from FN to TP (49 times). The less experienced the reader was, the higher the rate of sensitivity improvement. Conversely, the more experienced the readers were, the more limited the support capabilities of the CAD were.	

Evidence Table : Effectiveness/ safety/ organisational/ economic implication
Question : What is the diagnostic accuracy, effectiveness, safety, and cost-effectiveness of artificial intelligence-based chest x-ray for lung cancer screening?

Bibliographic Citation	Study Type/ Methods	LE	Number of Patients & Patient Characteristic	Intervention	Comparison	Length of Follow-up (if applicable)	Outcome Measures/ Effect Size	General Comments
10. Yoo H, Lee SH, Arru CD et al. AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset. Eur Radiol. 2021; 31(12): 9664-9674	**Diagnostic test study** To assess if deep learning-based AI algorithm improves reader performance for lung cancer detection on CXRs. Eight readers, including three radiology residents, and five board-certified radiologists, participated in the observer performance test. AI algorithm provided image-level probability of pulmonary nodule or mass on CXRs and a heatmap of detected lesions. Each reader reviewed each CXR twice: first, without AI and then with AI with at least four weeks of wash-out within the two reviews. Reader performance was compared with AUC, sensitivity, specificity, false-positives per image (FPPI), and rates of chest CT recommendations.	II-3	This reader study included 173 images from cancer-positive patients (n=98) and 346 images from cancer-negative patients (n=196) selected from National Lung Screening Trial (NLST).	Deep learning-based AI algorithm (Lunit INSIGHT CXR) (DCNN)	AI-aided versus AI-unaided between radiology residents and board-certified radiologists	-	**Observer performance assessment for visible lung cancer detection** Compared to that without AI, the average AUC for the detection of visible lung cancer increased significantly for radiology residents with AI (0.76 [95% CI: 0.67 to 0.86] vs. 0.82 [95% CI: 0.75 to 0.89]; p=0.003), but for radiologists, the average AUC was similar (0.82 [95% CI: 0.74 to 0.91] vs. 0.84 [95% CI: 0.79 to 0.89]; p=0.24). Compared to that without AI, the average sensitivity increased significantly for radiology residents (0.61 [95% CI: 0.55 to 0.67] vs. 0.72 [95% CI: 0.66 to 0.77]; p=0.016), but specificity was similar with AI (0.88 [95% CI: 0.86 to 0.90] vs. 0.88 [95% CI: 0.86 to 0.90]; p=0.89). For radiologists, average sensitivity (0.76 [95% CI: 0.72 to 0.81] vs. 0.76 [95% CI: 0.72 to 0.81]; p=1.00) was similar, but specificity increased with AI (0.79 [95% CI: 0.77 to 0.81] vs. 0.86 [95% CI: 0.84 to 0.87]; p< 0.001). Average FPPI without and with AI was similar for radiology residents (0.15 [95% CI: 0.11 to 0.18] vs. 0.12 [95% CI: 0.09 to 0.16]; p=0.13), but was significantly lower with AI for radiologists (0.24 [95% CI: 0.20 to 0.29] vs. 0.17 [95% CI: 0.13 to 0.20]; p<0.001). For patients with visible lung cancer on CXR, the average chest CT recommendation rate increased significantly for residents, but was similar for radiologists without and with AI (54.7% [95% CI: 48.2 to 61.2%] vs. 70.2% [95% CI: 64.2 to 76.2%]; p<0.001 for residents and 72.5% [95% CI: 68.0 to 77.1%] vs. 73.9% [95% CI: 69.4 to 78.3%]; p=0.68 for radiologists). Conversely, in patients without visible lung cancer, the average chest CT recommendation rate was similar without and with AI for residents, but decreased for radiologists (11.2% [95% CI: 9.6 to 13.1%] vs. 9.8% [95% CI: 8.0 to 11.6%]; p=0.32 for residents and 16.4% [95% CI: 14.7 to 18.2%] vs. 11.7% [95% CI: 10.2 to 13.3%]; p<0.001 for radiologists). When AI was used as a second reader, residents detected more missed lung cancer present in prior CXRs (39% vs. 71%, p = 0.021).	

ARTIFICIAL INTELLIGENCE-BASED CHEST X-RAY FOR LUNG CANCER SCREENING

CAWANGAN PENILAIAN TEKNOLOGI KESIHATAN (MAHTAS)

(ebook)