NIHR | National Institute for Health and Care Research

Health Technology Assessment

Check for updates

Extended Research Article

Software with artificial intelligence-derived algorithms for detecting and analysing lung nodules in CT scans: systematic review and economic evaluation

Julia Geppert,[1] Peter Auguste,[1] Asra Asgharzadeh,[1,4] Hesam Ghiasvand,[1,5] Mubarak Patel,[1] Anna Brown,[1] Surangi Jayakody,[1] Emma Helm,[2] Dan Todkill,[1] Jason Madan,[3] Chris Stinton,[1] Daniel Gallacher,[1] Sian Taylor-Phillips[1] and Yen-Fu Chen[1*]

[1]Warwick Evidence/Warwick Screening, Warwick Medical School, University of Warwick, Coventry, UK
[2]University Hospitals Coventry and Warwickshire, Coventry, UK
[3]Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK
[4]Population Health Science, University of Bristol, Bristol, UK
[5]Research Centre for Healthcare and Communities, Coventry University, Coventry, UK

[*]Corresponding author Y-F.Chen@warwick.ac.uk

Published May 2025
DOI: 10.3310/JYTW8921

Scientific summary

Software with artificial intelligence-derived algorithms for detecting and analysing lung nodules in CT scans: systematic review and economic evaluation

Health Technology Assessment 2025; Vol. 29: No. 14
DOI: 10.3310/JYTW8921

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

Lung nodules are found in different populations: (1) when people are referred for a computed tomography (CT) scan of the chest because they have signs and symptoms suggestive of lung cancer (symptomatic), (2) when people are investigated for conditions unrelated to lung cancer (incidental), or (3) through lung cancer screening programmes (screening). CT scans are also undertaken to assess whether the growth of previously identified nodules indicates malignancy and if further assessment or treatment is needed (surveillance). Nodules may be challenging to detect because of their small size, varying shape and proximity to other structures.

This assessment focuses on the use of software with artificial intelligence (AI)-derived algorithms to assist in the detection and analysis of lung nodules in CT chest scans.

Objectives

For the detection and analysis of lung nodules in symptomatic, incidental, screening or surveillance populations, the following key questions are asked.

Key question 1
What is the accuracy of CT image analysis assisted by AI software, and what are the practical implications and impacts on patient management?

Key question 2
What are the benefits and harms of CT image analysis assisted by AI software compared with unassisted reading?

Key question 3
What is the cost-effectiveness of CT image analysis assisted by AI software compared with unassisted reading?

Methods

Data sources
Databases including MEDLINE, EMBASE, Cochrane Database of Systematic Reviews, Cochrane CENTRAL, Health Technology Assessment (HTA) database (Centre for Reviews and Dissemination), International HTA database (INAHTA), Science Citation Index Expanded (Web of Science) and Conference Proceedings – Science (Web of Science) were searched from 1 January 2012 to January 2022. Preprints, trials registries, reference lists of included studies, relevant systematic reviews and forwards citations were also searched. Additional economics sources included NHS Economic Evaluation Database (NHS EED), Cost-Effectiveness Analysis registry (Tufts Medical Center), EconPapers and ScHARRHUD. Company submissions were accepted until 31 August 2022.

Eligibility criteria

Population
The population was (1) people undergoing a CT scan that included the chest with no known lung nodules or lung cancer and who were not receiving investigative or follow-up imaging for primary cancer elsewhere in the body; or (2) people having CT surveillance for a previously identified lung nodule.

Interventions
The intervention was analysis of chest CT images assisted by one of the 13 AI software specified by the National Institute for Health and Care Excellence (NICE).

Comparator
The comparator was CT image assessment without assistance by AI software, or no comparator.

Outcomes

- Accuracy of nodule detection; accuracy of measuring nodule diameter, volume or change in volume; characteristics of detected nodules; proportion of detected nodules that are malignant; technical failure rate; reading time; report turnaround time; impact of test result on clinical decision-making; number of people undergoing biopsy or excision or having CT scans as part of surveillance; number and stage of cancers detected; time to diagnosis; reader acceptability and experience of using AI software; concordance between readers with and without AI software, between readers using different AI software or between different AI software without human involvement; inter-observer variability; repeatability/reproducibility.
- Morbidity; mortality; health-related quality of life; patients' acceptance of use of the software.
- Cost-effectiveness covering incremental costs, incremental benefits, incremental cost-effectiveness ratio (ICER) and quality-adjusted life-years (QALYs).

Study selection, data extraction and quality appraisal
Two reviewers independently assessed articles for inclusion and assessed the articles' quality using the QUADAS-2 and QUADAS-C tools or the COSMIN Risk of Bias tool. A single reviewer extracted data, with a second reviewer checking. For cost-effectiveness, quality was independently assessed using the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) and Philips criteria.

Data synthesis
Narrative data synthesis was performed.

De novo cost-effectiveness analysis
Two decision trees, developed in TreeAge Pro (TreeAge Software Inc., Williamstown, MA, USA), were used to assess the cost-effectiveness of AI-assisted radiologist reading compared with unaided radiologist reading. The preliminary model followed current practice for identifying lung nodules that require further action (actionable nodules) based on morphology, nodule type and size. The full model followed the whole pathways of nodule surveillance and management as specified in the British Thoracic Society (BTS) guidelines. Associated costs of and health outcomes from the comparative strategies were estimated.

Information required to populate the models included the prevalence of lung nodules, risk of lung cancer with different nodule sizes, sensitivity and specificity for nodule detection, nodule type and size distributions in different population, resource use, costs and utilities. Where possible, parametrisation was driven by findings from the test accuracy review. This was supported by additional searches, clinical expert opinion and simulations to generate parameters otherwise not available. Assumptions and simplification were required for longer-term costs and health outcomes inputs to the full model.

Resource use and costs for both models were obtained from the cost-effectiveness literature and NHS reference cost schedule. Costs were reported in 2020/21 prices and discounted at 3.5% per annum.

The model estimated the mean costs incurred and benefits accrued associated with each strategy for people entering the model at 60 years old. Results are presented in the form of an ICER. The cost per correct detection of an actionable nodule was estimated in the preliminary model. The primary outcome measure for the full model was cost per QALY. The perspective was that of the NHS and Personal Social Services over a lifetime horizon. Secondary outcome measures were also analysed in the full model. Deterministic analysis for the base-case and scenario analyses as well as univariate and probabilistic sensitivity analyses were undertaken.

Results

Key question 1

Twenty-seven studies covering eight NICE-specified AI software and evaluating nodule detection or measurement accuracy/concordance, practical implications and/or impact on patient management were identified. All studies were rated as being at high risk of bias and had multiple applicability concerns. Twenty-four studies used retrospective data sets, 17 of which compared the performance of readers seeing and not seeing the findings of AI software concurrently ('concurrent AI'). Nine of them allowed comparison with stand-alone AI software without human input ('stand-alone AI'). One study evaluated readers with concurrent AI only (vs. a reference standard); five studies evaluated stand-alone AI only; and one further study compared stand-alone AI with unaided readers. Only three studies reported on prospective screening experiences based on a pilot trial conducted in the Republic of Korea: two studies reported on software-assisted reading only and one study used a before-and-after design.

Accuracy and reliability

Detection of any nodules

Three studies found that AI assistance significantly increased sensitivity of detecting people with nodules. Pooled per-person sensitivity varied from 0.43 to 0.68 for unaided reading and from 0.79 to 0.99 for AI-assisted reading. Average specificity decreased slightly in two studies while it improved slightly in one study (0.77–1.00 without and 0.81–0.97 with AI assistance). A fourth study reported improved average per-nodule sensitivity from 0.72 to 0.84 with no difference in false-positive rates with AI assistance.

Detection of actionable nodules

Three studies found that AI assistance significantly increased sensitivity of detecting actionable nodules (≥ 5 mm in diameter). In one study, specificity was significantly lower and the number of false-positive detections per image significantly increased with AI assistance. The other two studies also reported an increase in false-positive detections per scan, but no statistical test was performed.

Detection of malignant nodules

Three studies directly compared sensitivity, with two finding that AI assistance significantly increased sensitivity, and one also reporting lower specificity and higher false-positive detections per image. The remaining study only included one cancer case detected by readers both with and without AI assistance.

Modifiers for nodule detection accuracy

Estimated sensitivity and specificity for nodule detection varied substantially between studies, possibly due to heterogeneity in study designs, populations, reader experience and reader specialty.

Evidence from one UK reader study suggests that unaided, experienced radiologists in clinical practice (with 5% double reading) outperform inexperienced, trained radiographers assisted by concurrent AI who read the same screening CT images.

The detection performance of radiologists (with and without concurrent AI, respectively) was not significantly different between standard-dose and low-dose CT scans (one study).

Three studies that evaluated different AI software suggested that the accuracy of AI-assisted reading for detecting different types of nodules compared with unaided readers may vary depending on the performance of individual technology, but the evidence was insufficient for a firm conclusion to be drawn.

Nodule type determination

Inter-reader agreement in nodule type determination was similar in readers with and without software use (two studies).

Nodule size measurement

Nodule diameters were similar (two studies) or significantly larger (two studies) with software-aided measurements than with manual measurements. A significant correlation between software-aided and manual measurement was observed

(two studies). Inter-reader variability (three studies) and intra-reader variability (one study) in nodule size measurement was significantly reduced in readers with software use compared with manual measurement. However, the effect on measurement accuracy is unclear.

Classification into risk categories based on nodule type and size

AI-assisted readings showed a higher agreement with the consensus session (reference standard) than did unaided readings (one study). Inter-reader agreement in risk category classification based on BTS (one study), Lung-RADS (Lung CT Screening Reporting And Data System; two studies) and Fleischner (one study) consistently improved with concurrent AI. One study also reported reduced intra-reader variability with software use.

Whole read (detection and Lung-RADS categorisation)

One before-and-after study evaluated the performance of a whole read (with Lung-RADS category ≥ 3 classed as positive) for lung cancer detection. No significant difference in test accuracy was observed before and after software implementation. Positive predictive values differed significantly according to measurement planes (transverse, maximum orthogonal, any maximum).

Nodule growth

No study provided data comparing AI-assisted with unaided reading. The sensitivity of stand-alone software to detect nodule pairs in subsequent scans of the same patient was 100.0% (23/23), with no false-positive pairs (one study). The mean growth percentage discrepancy was similar for unaided chest radiologists and stand-alone software (one study). However, a single incorrect segmentation by stand-alone AI resulting in large measurement discrepancy led to the advice that human readers should visually verify nodule segmentation.

Practical implications

Segmentation failure ranged from 0% to 57% of nodules (eight studies). However, the observed nodule segmentation failure might be mostly due to radiologists rejecting segmentation results, rather than the system's inability to segment the nodule. Failure rates seem to be higher in ground-glass nodules (34%) and part-solid nodules (20%) than in solid nodules (7%) (one study). Manual modifications of segmentation were required in 29 to 59% of nodules (two studies).

Radiologist reading time reduced with concurrent AI by 11.3–78% compared with unaided reading (nine studies) but increased with the use of AI software after initial unaided reading ('2nd-read AI', + 26%, one study). When using software with vessel suppression function only, reading time was similar with and without software (one study).

Impact on patient management

- Among all detected nodules (true and false positives), the proportion of solid nodules was lower with concurrent AI than with unaided reading (87.1% vs. 90.6%) (one study). Additional true-positive nodules detected with software were 56–57% solid, due to larger improvements in the detection of subsolid nodules (two studies). Twenty-two per cent of additional true-positive nodules were ≥ 6 mm (one study).
- The proportion of detected actionable nodules that were malignant was lower with software use (two studies).
- With software use, readers tended to upstage rather than downstage Lung-RADS (three studies) or Fleischner risk categories (one study).
- The proportion of people classed as Lung-RADS category 3 or 4A increased with software use (two studies).
- Similar (one study) or slightly higher (one study) proportions of people were classed as Lung-RADS category 4B/4X, requiring biopsy or excision.
- One retrospective study showed that discrepancies (Lung-RADS category 1/2 vs. 4A/B) between readers would be reduced by half, and sensitivity for lung cancer would be improved with AI software use, which might translate into earlier diagnosis if confirmed in clinical practice.

Key question 2

No studies were identified that reported on the benefits and harms to patients of AI-assisted reading compared with current practice without AI assistance.

Key question 3

Of the 1,988 records identified, 15 were considered potentially relevant, but all were excluded at full-text stage. Two potentially relevant model-based economic analyses did not meet our inclusion criteria but were summarised as they provided some contextual evidence.

De novo cost-effectiveness analysis

Due to the complete absence of evidence related to clinical effectiveness, and substantial challenges in linking test accuracy evidence to clinical and economic outcomes, the findings presented here are highly uncertain and should be regarded as early indications and frameworks for future analyses. Our preliminary model suggested that AI-assisted radiologist reading dominates unaided reading in terms of cost per person with an actionable nodule correctly identified in the screening population. Our full model suggested that for symptomatic and incidental populations, AI-assisted CT image analysis dominates unaided radiologist reading for cost per correct detection of a person with an actionable nodule. However, when relevant costs and QALYs incurred throughout the full clinical pathway are considered, AI-assisted reading is dominated by the unaided reader. This is driven by costs and disutilities associated with false-positive results and CT surveillance. AI assistance was deemed cost-effective for both symptomatic and incidental populations in the scenario analyses from which disutility associated with false-positive results and CT surveillance were removed. In the screening population, AI assistance was cost-effective in the base case and all sensitivity and scenario analyses. This was driven by a more favourable profile of model inputs, including estimates of improved test specificity for AI assistance from a single study. Although more data were available to populate the screening population model, there was substantial uncertainty across all models.

Conclusions

AI-assisted detection and analysis of lung nodules increases consistency of nodule measurement and risk classification compared with unaided reading, but its effect on measurement accuracy is unclear. AI assistance appears to improve sensitivity for lung nodule and cancer detection but can be accompanied by a decrease in specificity and an increase in false-positive findings per scan, as well as raising risk categorisation. The reported performance of AI-assisted reading varies substantially among published studies (for any nodules: per-person sensitivity 0.79–0.99, per-person specificity 0.81–0.97), possibly attributable to heterogeneous study and reader populations, other study design features and risk of bias in addition to potential differences in the performance of individual technologies.

No eligible studies directly compared the performance of different AI software. Given the paucity of evidence, it is currently not possible to reliably establish the cost-effectiveness of AI-assisted reading compared with unaided reading, or the relative effectiveness and cost-effectiveness of strategies adopting different AI software to assist nodule detection and analysis. However, our preliminary results suggest that AI-assisted reading is dominant for the screening population, but reading without AI assistance dominates for symptomatic and incidental populations.

Published studies have largely been conducted retrospectively in a research rather than a clinical environment. All studies in this assessment were rated as being at high risk of bias and had multiple applicability concerns for UK settings. No studies evaluating downstream clinical outcomes were identified. Further studies are required.

Study registration

This study is registered as PROSPERO CRD42021298449.

Funding

This award was funded by the National Institute for Health and Care Research (NIHR) Evidence Synthesis programme (NIHR award ref: NIHR135325) and is published in full in *Health Technology Assessment*; Vol. 29, No. 14. See the NIHR Funding and Awards website for further award information.

Health Technology Assessment

NICE TAR and DAR

ISSN 2046-4924 (Online)

Impact factor: 3.5

A list of Journals Library editors can be found on the NIHR Journals Library website

Launched in 1997, *Health Technology Assessment* (HTA) has an impact factor of 3.5 and is ranked 30th (out of 174 titles) in the 'Health Care Sciences & Services' category of the Clarivate 2022 Journal Citation Reports (Science Edition). It is also indexed by MEDLINE, CINAHL (EBSCO Information Services, Ipswich, MA, USA), EMBASE (Elsevier, Amsterdam, the Netherlands), NCBI Bookshelf, DOAJ, Europe PMC, the Cochrane Library (John Wiley & Sons, Inc., Hoboken, NJ, USA), INAHTA, the British Nursing Index (ProQuest LLC, Ann Arbor, MI, USA), Ulrichsweb™ (ProQuest LLC, Ann Arbor, MI, USA) and the Science Citation Index Expanded™ (Clarivate™, Philadelphia, PA, USA).

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta.

Criteria for inclusion in the *Health Technology Assessment* journal

Manuscripts are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

Health Technology Assessment (HTA) research is undertaken where some evidence already exists to show that a technology can be effective and this needs to be compared to the current standard intervention to see which works best. Research can evaluate any intervention used in the treatment, prevention or diagnosis of disease, provided the study outcomes lead to findings that have the potential to be of direct benefit to NHS patients. Technologies in this context mean any method used to promote health; prevent and treat disease; and improve rehabilitation or long-term care. They are not confined to new drugs and include any intervention used in the treatment, prevention or diagnosis of disease.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

This article

The research reported in this issue of the journal was commissioned and funded by the Evidence Synthesis Programme on behalf of NICE as award number NIHR135325. The protocol was agreed in December 2021. The draft manuscript began editorial review in December 2022 and was accepted for publication in May 2024. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' manuscript and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this article.

This article presents independent research funded by the National Institute for Health and Care Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care.

This article was published based on current knowledge at the time and date of publication. NIHR is committed to being inclusive and will continually monitor best practice and guidance in relation to terminology and language to ensure that we remain relevant to our stakeholders.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Newgen Digitalworks Pvt Ltd, Chennai, India (www.newgen.co).