NIHR | National Institute for Health and Care Research

Journals Library

Check for updates

Health Technology Assessment

*Volume 29 • Issue 14 • May 2025*

*ISSN 2046-4924*

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

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

*DOI 10.3310/JYTW8921*

Health Technology Assessment

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

This report should be referenced as follows:

Geppert J, Auguste P, Asgharzadeh A, Ghiasvand H, Patel M, Brown A, *et al.* Software with artificial intelligence-derived algorithms for detecting and analysing lung nodules in CT scans: systematic review and economic evaluation. *Health Technol Assess* 2025;**29**(14). https://doi.org/10.3310/JYTW8921

# 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).

# Abstract

**Background:** Lung cancer is one of the most common types of cancer and the leading cause of cancer death in the United Kingdom. Artificial intelligence-based software has been developed to reduce the number of missed or misdiagnosed lung nodules on computed tomography images.

**Objective:** To assess the accuracy, clinical effectiveness and cost-effectiveness of using software with artificial intelligence-derived algorithms to assist in the detection and analysis of lung nodules in computed tomography scans of the chest compared with unassisted reading.

**Design:** Systematic review and de novo cost-effectiveness analysis.

**Methods:** Searches were undertaken from 2012 to January 2022. Company submissions were accepted until 31 August 2022. Study quality was assessed using the revised tool for the quality assessment of diagnostic accuracy studies (QUADAS-2), the extension to QUADAS-2 for assessing risk of bias in comparative accuracy studies (QUADAS-C) and the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist. Outcomes were synthesised narratively. Two decision trees were used for cost-effectiveness: (1) a simple decision tree for the detection of actionable nodules and (2) a decision tree reflecting the full clinical pathways for people undergoing chest computed tomography scans. Models estimated incremental cost-effectiveness ratios, cost per correct detection of an actionable nodule, and cost per cancer detected and treated. We undertook scenario and sensitivity analyses.

**Results:** Twenty-seven studies were included. All were rated as being at high risk of bias. Twenty-four of the included studies used retrospective data sets. Seventeen compared readers with and without artificial intelligence software. One reported prospective screening experiences before and after artificial intelligence software implementation. The remaining studies either evaluated stand-alone artificial intelligence or provided only non-comparative evidence. (1) Artificial intelligence assistance generally improved the detection of any nodules compared with unaided reading (three studies; average per-person sensitivity 0.43–0.68 for unaided and 0.79–0.99 for artificial intelligence-assisted reading), with similar or lower specificity (three studies; 0.77–1.00 for unaided and 0.81–0.97 for artificial intelligence-assisted reading). Nodule diameters were similar or significantly larger with semiautomatic measurements than with manual measurements. Intra-reader and inter-reader agreement in nodule size measurement and in risk classification generally improved with artificial intelligence assistance or were comparable to those with unaided reading. However, the effect on measurement accuracy is unclear. (2) Radiologist reading time generally decreased with artificial intelligence assistance in research settings. (3) Artificial intelligence assistance tended to increase allocated risk categories as defined by clinical guidelines. (4) No relevant clinical effectiveness and cost-effectiveness studies were identified. (5) The de novo cost-effectiveness analysis suggested that for symptomatic and incidental populations, artificial intelligence-assisted computed tomography image analysis dominated the unaided radiologist in cost per correct detection of an actionable nodule. However, when relevant costs and quality-adjusted life-years from the full clinical pathway were included, artificial intelligence-assisted computed tomography reading was dominated by the unaided reader. For screening, artificial intelligence-assisted computed tomography image analysis was cost-effective in the base case and all sensitivity and scenario analyses.

**Limitations:** Due to the heterogeneity, sparseness, low quality and low applicability of the clinical effectiveness evidence and the major challenges in linking test accuracy evidence to clinical and economic outcomes, the findings presented here are highly uncertain and provide indicators/frameworks for future assessment.

**Conclusions:** Artificial intelligence-assisted analysis of computed tomography scan images may reduce variability of and improve consistency in the measurement and clinical management of lung nodules. Artificial intelligence may increase nodule and cancer detection but may also increase the number of patients undergoing computed tomography surveillance unnecessarily. No direct comparative evidence was found, and nor was any direct evidence found on clinical outcomes and cost-effectiveness. Artificial intelligence-assisted image analysis may be cost-effective in screening for lung cancer but not for symptomatic populations. However, reliable estimates of cost-effectiveness cannot be obtained with current evidence.

**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.

# Contents

# List of tables

# List of figures

# List of boxes

# List of supplementary material

**Report Supplementary Material 1** Lists and further details of excluded studies

**Report Supplementary Material 2** Data extraction and quality assessment templates

**Report Supplementary Material 3** Summary of two potentially relevant economic analyses not meeting the review inclusion criteria

**Report Supplementary Material 4** R code for simulation

**Report Supplementary Material 5** Rationale for developing the Warwick Evidence (WE) model and comparison between the Exeter Natural history-Based economic model of Lung cancer screening (ENaBL) model and WE model

Supplementary material can be found on the NIHR Journals Library report page (https://doi.org/10.3310/JYTW8921).

Supplementary material has been provided by the authors to support the report and any files provided at submission will have been seen by peer reviewers, but not extensively reviewed. Any supplementary material provided at a later stage in the process may not have been peer reviewed.

# Glossary

**2 x 2 contingency table** A table with two rows and two columns that presents classifications of individuals with regard to presence/absence of a disease condition, usually by a new diagnostic test to be evaluated and a reference standard that is considered to reflect the true disease status in the following form:

| Index (new) test (↓) / Reference (gold) standard (→) | Yes | No |
|---|---|---|
| Yes | a = true positive | b = false positive |
| No | c = false negative | d = true negative |

**Cohen's kappa** Denoted as the Greek letter 'κ', a statistic used for assessing the level of agreement between different raters (inter-rater reliability) or between the rating (classification) made by the same rater at different time points (intra-rater reliability) that takes into account agreement by chance. Similar to correlation coefficients, it can range between −1 and +1, where +1 denotes perfect agreement and 0 denotes the agreement that can be expected from random chance.

**Concordance** The agreement between two variables.

**Concurrent artificial intelligence (Concurrent AI)** In this report, the use of artificial intelligence software at the same time when a radiologist is reading and analysing the computed tomography scan image. This is in contrast with second-read artificial intelligence (see *Second-read articial intelligence*).

**Correlation** The degree to which two continuous variables are linearly related.

**Dice similarity coefficient or Dice coefficient** An index of spatial overlap and a reproducibility validation metric when segmentation of a nodule between different readers/readings is compared. It ranges between 0 (no overlap) and 1 (perfect overlap). In the context of comparing two diagnostic tests, it can be regarded as a measure of similarity in the classification of disease between two tests, ignoring cases considered as negative by both tests.

$$DSC = \frac{2a}{2a+b+c} = \frac{a}{a+\frac{1}{2}(b+c)} = \frac{2TP}{2TP+FP+FN} \tag{1}$$

Dice coefficient ranges between 0 and 1, with 1 signifying the greatest similarity between the two tests. Also known as the F-score or the Sørensen–Dice coefficient.

**Dominate** When different options are being compared in a health economic evaluation, an option 'dominates' another option if the former is less costly and more effective than the latter. This can also be stated as the latter option 'being dominated by' the former option.

**False-negative value** The number of cases in which the index test has wrongly suggested that the patient is disease-free.

$$FN = c \tag{2}$$

**False-positive rate** The proportion of people who test positive for a disease among people who do not have the disease of interest; the ratio between the false-positive value and (true-negative value + false-positive value). Equals 1 – specificity. Sometimes used in the literature to describe the 'number of false-positive detections per image' (see *Number of false-positive detections per image*), which may cause confusion.

**False-positive value** The number of cases in which the index test has wrongly indicated that the patient has the disease.

$$FP = b \tag{3}$$

**Index test** A (new) test whose performance is being evaluated against a reference standard.

**Inter-rater reliability** The degree of agreement between independent observers who rate the same phenomenon.

**Intra-rater reliability** The degree of agreement among repeated administrations of a diagnostic test performed by a single rater. Not to be confused with inter-rater reliability.

**Limits of agreement** A range showing where the vast majority (95%) of the differences between two measurements (e.g. lung nodule size measured by two radiologists) is likely to lie. Smaller limits of agreement indicate better agreement in measurements. Also known as Bland–Altman method.

**Lin's concordance correlation coefficient** Also denoted as $\rho_{c,}$ or CCC, a measure of agreement between two continuous variables that takes into account both measurement bias and measurement consistency (see below). Its value ranges between −1 (perfect discordance) and 1 (perfect concordance).

**Measurement accuracy** The accuracy of a measurement of a quantity (e.g. size of a lung nodule) made by a person (e.g. radiologist) or a tool (e.g. computer software) compared with the 'true' quantity, for example, whether computer software tends to overestimate the size of a nodule compared with its 'true' size. Also known as 'measurement bias' or 'systematic measurement error'.

**Measurement precision** How well the estimated quantities agree with each other when a person or a tool measures the same quantity (e.g. the size of a nodule) multiple times (intra-rater reliability; see *Intra-rater reliability*) or when different people try to measure the same quantity (inter-rater reliability; see *Inter-rater reliability*). Also known as 'measurement consistency', 'measurement reliability' or 'random measurement error'.

**Negative predicted value** The percentage of patients with a negative index test result who are actually disease-free.

$$NVP = \frac{d}{c+d} = \frac{TN}{FN+TN} \tag{4}$$

**Number of false-positive detections per image** In nodule detection, a false-positive finding (recognising/reporting something as a nodule when in fact it is not) can be recorded multiple times in different locations of a computed tomography scan image. The number of false-positive detections per image represents the total number of false-positive findings across a set of computed tomography scan images divided by the total number of computed tomography scan images within this set. For example, if overall 15 false-positive findings are recorded among 10 computed tomography scan images being reviewed, the number of false-positive detections per scan/image would be 1.5. This number has no theoretical limit – unlike false-positive value and false-positive rate (see definitions above) in a per-person analysis, which are bounded by the total number of people without a nodule. The number is sometimes referred to in the literature as 'false-positive rate', which may cause confusion.

**Pearson's correlation coefficient** The measure of linear correlation between two sets of data; the ratio between the covariance of two variables and the product of their standard deviations. It can range between −1 and 1, with −1 indicating perfect negative correlation, 1 indicating perfect positive correlation and 0 indicating no correlation.

**Per-nodule analysis** Analysis of test accuracy results for nodule detection in which the unit of analysis is an individual nodule.

**Per-person (per-scan) analysis** Analysis of test accuracy results for nodule detection in which the unit of analysis is a person or a computed tomography scan image. As multiple nodules may be found in a computed tomography scan image for a person, this measure differs from per-nodule analysis and is more clinically relevant, as decision-making in nodule management often depends on the largest nodule or the nodule with most suspicious features rather than all nodules.

**Positive predictive value** The percentage of patients with a positive index test result who actually have the disease.

$$PPV = \frac{a}{a+b} = \frac{TP}{TP+FP} \tag{5}$$

**Receiver operating characteristic curve** A graph showing the sensitivity and specificity for every possible threshold of a test.

**Reference standard** The test, combination of tests, or procedure that is considered the best available method of categorising participants in a study of diagnostic test accuracy as having or not having a target condition.

**Risk-dominant nodule** The lung nodule that is judged to carry the highest risk (or probability) of being malignant and based on which the decision about clinical management is made for a patient with more than one nodule detected on the computed tomography image. It is usually the largest nodule without clearly benign features.

**Second-read artificial intelligence (2nd-read artificial intelligence)** Refers in this report to a radiologist first reading and analysing the computed tomography image independently, and then bringing up and considering findings produced with artificial intelligence assistance (as a 'second-reader') to make necessary changes and finalise nodule detection and analysis.

**Segmentation** A step in digital image processing in which small areas in an image (called pixels) are classified and labelled to facilitate further analysis. For example, segmentation enables an area on a computed tomography image that is likely to represent a lung nodule to be marked and separated from the rest of the image.

**Sensitivity** The proportion of people who test positive for a disease among people who have the disease of interest; the ratio between the true-positive value and (true-positive value + false-negative value).

$$Sensitivity = \frac{a}{a+c} = \frac{TP}{TP+FN} \tag{6}$$

**Specificity** The proportion of people who test negative for a disease among people who do not have the disease of interest. The ratio between the true-negative value and (true-negative value + false-positive value).

$$Specificity = \frac{d}{b+d} = \frac{TN}{TN+FP} \tag{7}$$

**True-negative value** The number of cases in which the index test has correctly indicated the patient as being disease-free.

$$TN = d \tag{8}$$

**True-positive value** The number of cases in which the index test has correctly indicated the patient as having the disease.

$$TP = a \tag{9}$$

# List of abbreviations

| | |
|---|---|
| AI | artificial intelligence |
| BTS | British Thoracic Society |
| CCC | Lin's concordance correlation coefficient |
| CHEERS | Consolidated Health Economic Evaluation Reporting Standards |
| CI | confidence interval |
| CT | computed tomography |
| DAR | diagnostic assessment report |
| EAG | external assessment group |
| FBP | filtered back projection |
| HTA | health technology assessment |
| ICER | incremental cost-effectiveness ratio |
| IQR | interquartile range |
| K-LUCAS | Korean Lung Cancer Screening |
| LDCT | low-dose computed tomography |
| LSUT | Lung Screen Uptake Trial |
| Lung-RADS | Lung CT Screening Reporting And Data System |
| MBIR | model-based iterative reconstruction |
| MDT | multidisciplinary team |
| MRMC | multireader multicase |
| NICE | National Institute for Health and Care Excellence |
| NLST | National Lung Screening Trial |
| PACS | picture archiving and communication system |
| PET-CT | positron emission tomography-computer tomography |
| PSS | Personal Social Services |
| QALY | quality-adjusted life-year |
| SD | standard deviation |
| TLHC | Targeted Lung Health Check |
| VDT | volume doubling time |

# Note

This monograph is based on the Diagnostic Assessment Report produced for NICE. The full report contained a considerable number of data that were deemed confidential. The full report was used by the Diagnostic Advisory Committee at NICE in their deliberations. The full report with each piece of confidential data removed and replaced by the statement 'confidential information (or data) removed' is available on the NICE website: www.nice.org.uk

The present monograph presents as full a version of the report as is possible while retaining readability, but some sections, sentences, tables and figures have been removed. Readers should bear in mind that the discussion, conclusions and implications for practice and research are based on all the data considered in the original full NICE report.

# Plain language summary

Lung cancer is one of the most common types of cancer in the UK. In the early stages, people may not have symptoms and so lung cancer is often diagnosed late. Identifying and monitoring lung nodules using computed tomography scans are the primary means of detecting lung cancer at earlier stages. If a nodule is found, it needs to be measured accurately so that the cancer risk can be assessed. Currently, images from computed tomography scans are interpreted without artificial intelligence software.

Artificial intelligence could help to detect and measure nodules more accurately and quickly. This report looks at the evidence on the benefits and harms of artificial intelligence in helping healthcare professionals to find and measure lung nodules. The report also looks at whether artificial intelligence offers value for money.

We did not find any studies that directly compared radiologists' performance with and without the help of artificial intelligence in the UK. All of the studies we did find were of low quality. Findings from these studies suggest pros and cons of using artificial intelligence:

- Artificial intelligence could improve nodule detection, with bigger improvements seen in detecting smaller nodules. However, artificial intelligence might increase the detection of both cancer as well as harmless nodules.
- With artificial intelligence, measuring nodule size and assessing cancer risk could be more consistent.
- In up to half of nodules, automatic size measurement needs manual adjustment.
- Radiologists' reading time could be reduced with artificial intelligence.

It has not yet been established how artificial intelligence would affect radiologists' performance in United Kingdom practice. Whether artificial intelligence offers good value for money is also uncertain because we lack good evidence. Our early assessment suggests that artificial intelligence software might be cost-effective for lung cancer screening but might not be cost-effective for people who have symptoms or who have a computed tomography scan for other reasons. This is because the balance between the benefit of detecting more cancers and the harm of worrying people with incorrect test results and adding unnecessary regular follow-ups may be different in different populations.

# 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.

# Chapter 1 Background and definition of the decision problem(s)

Sections of this report are reproduced from the final scope issued by National Institute for Health and Care Excellence (NICE) and Geppert *et al.*[1] © NICE 2021 *Software with Artificial Intelligence Derived Algorithms for Automated Detection and Analysis of Lung Nodules from CT Scan Images [DAP60]*. Available from www.nice.org.uk/guidance/dg55. All rights reserved. Subject to https://www.nice.org.uk/terms-and-conditions#notice-of-rights

National Institute for Health and Care Excellence guidance is prepared for the National Health Service in England. All NICE guidance is subject to regular review and may be updated or withdrawn. NICE accepts no responsibility for the use of its content in this product/publication.

## Lung nodules and lung cancer

Lung nodules are small rounded or irregular-shaped growths with a diameter ≤ 3 cm that are found inside the lung. They vary in size, and this variation is strongly associated with risk of malignancy albeit in a non-linear fashion.[2] A nodule with a diameter < 3 mm is referred to as a micronodule, and the measurement of these is not recommended due to accuracy limitations.[3] Lung nodules with a diameter < 5 mm have low probability of being lung cancer[4] and do not usually require further action if they are detected incidentally. We refer to nodules with a diameter ≥ 5 mm as 'actionable nodules'.

Most lung nodules on a computed tomography (CT) scan appear as solid structures, but some are subsolid. Subsolid nodules either have a solid part surrounded by a non-solid, cloud-like structure (part-solid nodules) or appear entirely non-solid (pure ground-glass nodules). While most lung nodules are benign (non-cancerous), some may be malignant (cancerous) or may develop into lung cancer.

Lung nodules are found (1) when people are referred for a CT scan that includes the chest because of signs and symptoms suggestive of lung cancer, (2) when people are investigated for other conditions unrelated to lung cancer, or (3) through lung cancer screening programmes. People with previously identified lung nodules can also have CT scans as part of surveillance to assess whether the growth of the nodules indicates malignancy and if further assessment or treatment is needed. Lung nodules may be challenging to detect because of their small size, varying shape and proximity to other structures.

Lung cancer is one of the most common types of cancer in the UK. Its incidence rises steeply from around the ages of 45–49 years.[5] Lung cancer causes symptoms, such as persistent cough, coughing up blood and feeling short of breath. People in the early stages of the disease may not have symptoms and so lung cancer is often diagnosed late. In 2018, > 65% of all 39,267 lung cancers in England were diagnosed at stage III ($n$ = 7,886) or IV ($n$ = 18,104).[6] The *NHS Long Term Plan*[7] sets out an ambitious target of diagnosing 75% of all cancers at an earlier stage, that is stage I or II, by 2028.

Although most lung nodules are non-cancerous, in a small number of cases they can be the first signs of an early cancer in the lung. In the absence of other specific and reliable signs and biomarkers, identifying and monitoring lung nodules using CT scans of the chest remains the primary means of detecting lung cancer at earlier stages.

## Diagnostic and care pathway

### *Pathway to computed tomography scan due to signs and symptoms suggestive of lung cancer*

People with signs and symptoms suggestive of lung cancer are often identified in primary care. The NICE guideline on recognition and referral for suspected cancer[8] recommends that people aged ≥ 40 years are offered an urgent chest

X-ray (within two weeks of referral) if they have two or more (or one or more if they have ever smoked) of the following unexplained symptoms:

- cough
- fatigue
- shortness of breath
- chest pain
- weight loss
- appetite loss.

An urgent chest X-ray should also be considered for people aged ≥ 40 years if they have persistent or recurrent chest infection, finger clubbing, enlarged lymph nodes near the collarbone or in the neck (supraclavicular lymphadenopathy or persistent cervical lymphadenopathy), chest signs consistent with lung cancer or increased platelet count (thrombocytosis).

If the chest X-ray findings suggest lung cancer, the patient should be referred to secondary care using a suspected cancer pathway referral for an appointment within two weeks. During the scoping of this technology appraisal, clinical experts noted that referral to secondary care for a CT scan may also be made if the X-ray findings do not show abnormalities, but an ongoing suspicion of lung cancer remains. People aged ≥ 40 years who present with unexplained coughing up of blood (haemoptysis) should be referred directly for a CT scan using the suspected lung cancer referral pathway, or for direct access to CT where this is available in primary care.

In secondary care, people with known or suspected lung cancer should be offered a contrast-enhanced chest CT scan to further establish the diagnosis and stage of the disease.[9]

### *Lung cancer screening*

In September 2022, the UK National Screening Committee recommended targeted lung cancer screening for people aged 55–74 years identified as being at high risk of lung cancer.[10] NHS England is evaluating the Targeted Lung Health Check (TLHC) programme in some areas of England,[11] which provides a feasible and effective starting point for the implementation of a targeted screening programme in England. In this programme, people aged between 55 and 74 years who have ever smoked are invited to receive a lung health check. The lung health check involves collecting information about the person's lung health, lifestyle and family and medical history, and measuring their height and weight. Following the lung health check, people assessed as being at high risk of lung cancer are offered a low-dose CT (LDCT) scan. The use of computer-aided detection (CAD) systems is not a requirement under this protocol, but software is being used as part of the TLHC programme.

### *Initial assessment and computed tomography surveillance of lung nodules*

In the NHS, the investigation of lung nodules follows the British Thoracic Society (BTS) guidelines for the investigation and management of pulmonary nodules and depends on the composition of the nodule (i.e. whether it is solid or subsolid).[12] The guidelines recommend the same diagnostic approach for nodules detected incidentally, symptomatically or through screening. The guidelines are for lung nodules in adults. Clinical expert opinion is that lung nodules in children are not currently routinely investigated to avoid unnecessary CT scans as these nodules are rarely malignant.

*Appendix 1*, *Figure 14* shows the recommended pathway for the initial assessment of solid lung nodules. When there are multiple nodules, the size of the largest nodule should be considered. For newly identified nodules above a specified size, malignancy risk is estimated using the Brock model.[13] The nodule size (in diameter) and the number of nodules detected are among the inputs to this multivariable prediction model.[14]

The initial assessment of subsolid nodules (part-solid and ground glass) follows a similar pathway (see *Appendix 1*, *Figure 15*), but because these nodules can sometimes disappear on their own, the pathway involves repeating the CT scan at 3 months before the Brock malignancy risk model is used. The Herder model[15] is not used for subsolid nodules.

*Appendix 1*, *Figure 16* shows the recommended pathway for CT surveillance of solid lung nodules. The overall aim of this approach is to use the presence and speed of the nodule growth to distinguish between benign and malignant nodules. The nodule's growth should be assessed by estimating its volume doubling time (VDT). The surveillance period for subsolid nodules is longer (4 years) than for solid nodules (1 year with volume and 2 years with diameter measurements).

The BTS guidelines are currently being updated.[16]

Outside the UK, the Lung CT Screening Reporting And Data System (Lung-RADS) developed by the American College of Radiology has also been widely used for stratifying cancer risk to inform the clinical management of lung nodules identified by screening programmes,[17] and it was adopted in some of the studies assessed in this report. Lung-RADS allows nodules to be categorised according to their size and features with increasing risk of lung cancer:

- Category 1: negative (no nodules and definitely benign nodules); risk of malignancy < 1%.
- Category 2: benign appearance or behaviour (nodules with a very low likelihood of becoming a clinically active cancer due to size or lack of growth); risk of malignancy < 1%.
- Category 3: probably benign (probably benign findings – short-term follow-up suggested; includes nodules with a low likelihood of becoming a clinical active cancer); risk of malignancy 1–2%.
- Category 4A: suspicious (findings for which additional diagnostic testing is recommended); risk of malignancy 5–15%.
- Category 4B and 4X: very suspicious (findings for which additional diagnostic testing and/or tissue sampling is recommended); risk of malignancy > 15%.

Lung-RADS uses different cut-off sizes for categorising lung nodules than the BTS guidelines;[12] for example, for solid nodules at baseline (initial) scan, a nodule size of ≥ 6 mm would be classified as Lung-RADS category 3 with a recommendation for CT follow-up (compared with ≥ 5 mm for CT surveillance in the BTS guidelines).

### *Current methods of detecting nodules and measuring nodule volume and growth on CT scans*

Currently, assistance with artificial intelligence (AI)-derived software is not routine in clinical practice in the UK. The healthcare professional reviewing the scan may be a specialist in reviewing chest CT images (such as a thoracic radiologist) or less specialised (such as a general radiologist in an accident and emergency department).

In the TLHC programme, the healthcare professionals reviewing the scans are radiologists specialised in reviewing chest CT images. They are either radiologists who regularly lead at their local lung cancer multidisciplinary team (MDT) or radiologists who yearly, as part of their normal clinical practice, report > 500 thoracic CT scans, of which a significant proportion should have been performed for the evaluation of lung cancer.[18] Software for the automated detection of lung nodules has been used in the TLHC programme. The British Society of Thoracic Imaging and the Royal College of Radiologists have published a summary of radiology-related considerations for the TLHC, including advice on software.[19]

The 2015 BTS guidelines for the investigation and management of pulmonary nodules recommend that the size of an identified nodule should be quantified as the volume of the nodule.[12] To do this, volumetry software needs to be used. In current practice, this software is often part of the picture archiving and communication system (PACS), or a module in a software that comes with the CT scanner. When measuring the size of the part-solid nodules, the diameter of the solid part of the nodule is considered. In ground-glass nodules, the diameter of the entire nodule is measured.

This volumetry software may or may not have the capability to compare sequential scans to automatically measure the VDT. When this feature is not available or not used, the VDT can be calculated by inputting the nodule volume measurements and the dates of the two scans into the BTS Pulmonary Nodule Risk Calculator.[14] In addition to growth, for ground-glass nodules, any later appearance of a solid part is assessed.

Where volumetry software is not available or measuring the nodule volume by the software is not possible because of the quality of the image or the location of the nodule within the lung, the largest diameter of the nodule is measured.

The VDT can then be estimated by inputting the diameter measurements and dates of the two scans using the BTS Pulmonary Nodule Risk Calculator.[14] During scoping, clinical experts reported that diameter measurements are still widely used in the NHS.

Mapping on to the BTS guidelines and current clinical practice, AI-software-assisted reading may impact on the detection and analysis of pulmonary nodule in a number of ways, as shown in *Figure 1*.

The relevant evidence concerning the potential impact of AI assistance at various points in the CT image analysis and nodule management process presented in this report and the incorporation of these pieces of evidence in our cost-effectiveness analysis are as follows.

1. Accuracy in the identification of nodules: evidence presented in *Nodule detection*; incorporated as a parameter for the economic model (see *Test accuracy*).
2. Accuracy in classification of nodule type: evidence presented in *Nodule type determination*; not included in the economic model as no clear evidence of an impact by AI software.
3. Accuracy and precision in measuring nodule size/volume: evidence presented in *Nodule diameter measurement* and *Nodule volume measurement*; incorporated into the model through simulation output (*Information required for the model* and *Appendix 7*, *Table 65* and text).
4. Number of nodules detected as an input to Brock model: no evidence found; not included in the economic model.
5. Accuracy and precision in measuring nodule growth: evidence presented in *Use case 2: nodule growth monitoring in people with previously identified lung nodules*; incorporated into the economic model through simulation output (*Information required for the model* and *Appendix 7*, *Table 65* and text).
6. Capability of measuring volume rather than diameter: incorporated into the model structure, which allows varying proportion between volumetry and diameter measurements.
7. Impact on reporting time: evidence presented in *Radiologist reading time (10 studies)*; incorporated as a parameter for the economic model.

### *Diagnosis and staging of lung cancer*

To guide the treatment of lung cancer, information about the cancer type and spread (stage) is needed. The NICE guideline on the diagnosis and management of lung cancer[9] recommends choosing investigations that give the most information about diagnosis and staging at the lowest risk to the person. The type and sequence of investigations may vary, but the investigations commonly include a contrast-enhanced CT of the chest, abdomen and pelvis, a positron emission tomography-computed tomography (PET-CT) scan and magnetic resonance imaging. Tissue diagnosis is often obtained by image-guided biopsy, endobronchial ultrasound-guided transbronchial needle aspiration and endoscopic ultrasound-guided fine-needle aspiration, respectively.

### *Treatment for lung cancer*

After diagnosis, treatment for lung cancer is based on several factors, such as the overall health of the patient and the type, size, position and stage of the cancer. The treatment may include surgery, chemotherapy, radiotherapy, immunotherapy or other targeted therapy drugs, or a combination of these.[9]

## Population and relevant subgroups

This diagnostic assessment included people who undergo any type of CT scan (e.g. with or without contrast, low dose or standard dose; excluding PET-CT) that includes part or all of the chest for the following reasons.

1. Use case 1 (nodule detection and analysis): people who have no confirmed lung nodules or lung cancer and are not having staging investigations or follow-up imaging for primary cancer elsewhere in the body:

   - because of signs or symptoms suggestive of lung cancer (symptomatic population)
   - for reasons unrelated to suspicion of lung cancer (incidental population)
   - who attend lung cancer screening (screening population)

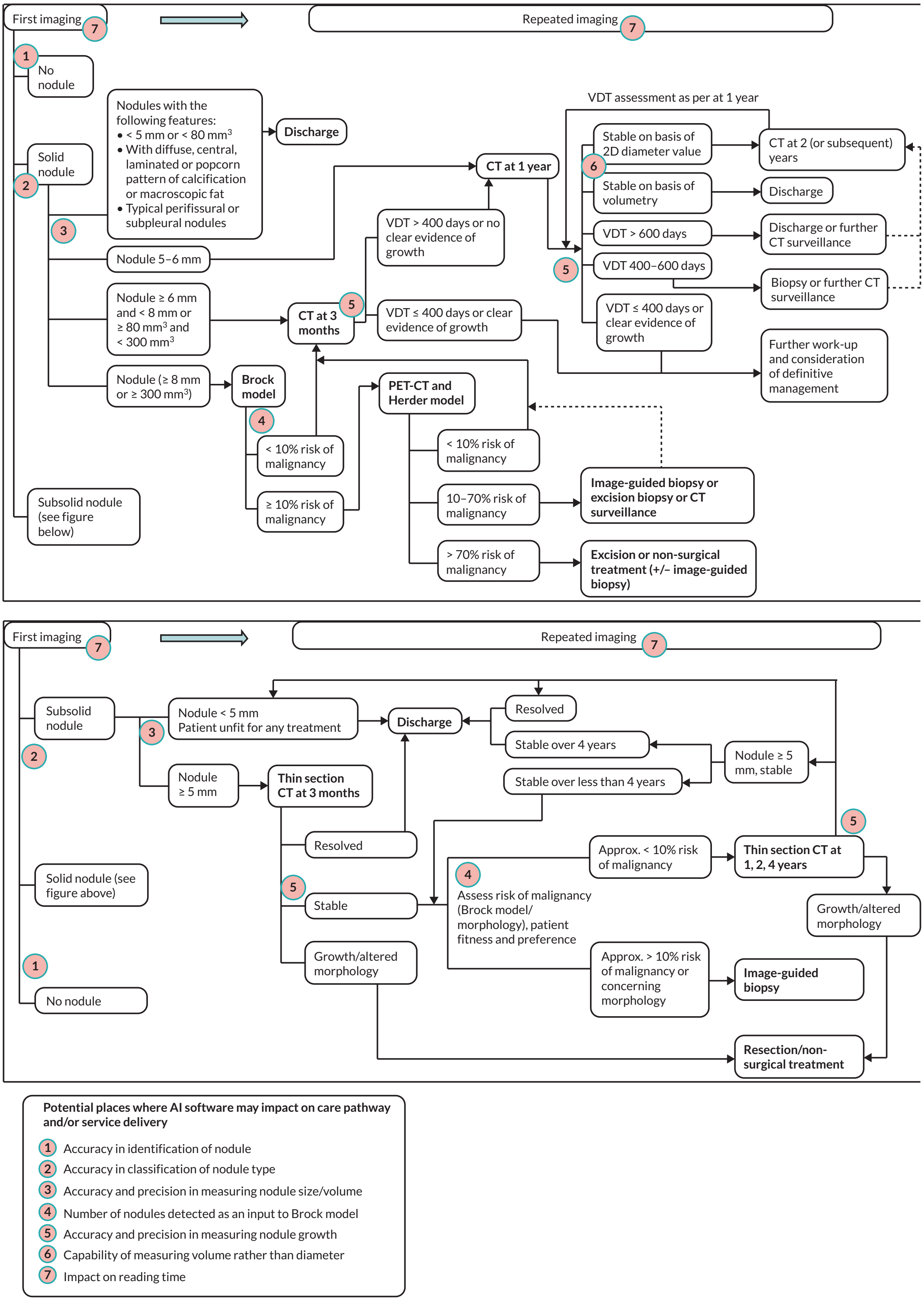

First imaging 7
Repeated imaging 7
1 No nodule
Solid nodule 2
3
Nodules with the following features:
• < 5 mm or < 80 mm³
• With diffuse, central, laminated or popcorn pattern of calcification or macroscopic fat
• Typical perifissural or subpleural nodules
Discharge
Nodule 5–6 mm
Nodule ≥ 6 mm and < 8 mm or ≥ 80 mm³ and < 300 mm³
Nodule (≥ 8 mm or ≥ 300 mm³)
CT at 3 months 5
Brock model
4
< 10% risk of malignancy
≥ 10% risk of malignancy
PET-CT and Herder model
< 10% risk of malignancy
10–70% risk of malignancy
> 70% risk of malignancy
Image-guided biopsy or excision biopsy or CT surveillance
Excision or non-surgical treatment (+/– image-guided biopsy)
Subsolid nodule (see figure below)
VDT > 400 days or no clear evidence of growth
VDT ≤ 400 days or clear evidence of growth
CT at 1 year
VDT assessment as per at 1 year
6
Stable on basis of 2D diameter value
Stable on basis of volumetry
VDT > 600 days
5
VDT 400–600 days
VDT ≤ 400 days or clear evidence of growth
CT at 2 (or subsequent) years
Discharge
Discharge or further CT surveillance
Biopsy or further CT surveillance
Further work-up and consideration of definitive management
First imaging 7
Repeated imaging 7
Subsolid nodule
2
3
Nodule < 5 mm Patient unfit for any treatment
Nodule ≥ 5 mm
Thin section CT at 3 months
Discharge
Resolved
Stable over 4 years
Stable over less than 4 years
Nodule ≥ 5 mm, stable
5
Resolved
5
Stable
Growth/altered morphology
4
Assess risk of malignancy (Brock model/ morphology), patient fitness and preference
Approx. < 10% risk of malignancy
Thin section CT at 1, 2, 4 years
Growth/altered morphology
Approx. > 10% risk of malignancy or concerning morphology
Image-guided biopsy
Resection/non-surgical treatment
Solid nodule (see figure above)
1
No nodule
Potential places where AI software may impact on care pathway and/or service delivery
1 Accuracy in identification of nodule
2 Accuracy in classification of nodule type
3 Accuracy and precision in measuring nodule size/volume
4 Number of nodules detected as an input to Brock model
5 Accuracy and precision in measuring nodule growth
6 Capability of measuring volume rather than diameter
7 Impact on reading time

**FIGURE 1** Points at which AI-derived software may have an impact on nodule detection and analysis and the relevant evidence in this report. PET-CT, positron emission tomography-computer tomography.

2. Use case 2 (nodule growth monitoring): people having CT surveillance for a previously identified lung nodule (surveillance population).

The use of the technologies for cancer staging and follow-up (including detection of metastasis to the lung) in people with extrathoracic primary cancers is outside the scope of this assessment.

### *Other subgroups of potential interest*

Across populations and use cases:

- parameters of the CT scan – with versus without contrast; low dose versus standard dose
- characteristics of the patient – different ethnicity
- characteristics of the lung nodule – solid nodules versus subsolid nodules
- characteristics of the reader – general radiologist (or other healthcare professional) versus radiologist (or other healthcare professional) with thoracic specialty
- within the incidental population – different reasons for the CT scan.

## Description of technologies under assessment

This diagnostic assessment focuses on the use of computer software with AI-derived algorithms for the automated detection and analysis of lung nodules from CT scan images of the chest. AI is a term that broadly refers to 'machines that perform tasks normally performed by human intelligence, especially when the machines learn from data how to do those tasks'.[20] The technologies included in this diagnostic assessment were defined in the NICE final scope and comprise computer software developed in a process that involves learning from data to detect and analyse lung nodules on CT scan images. The algorithms in the software are fixed but updated periodically.

Software is included in this diagnostic assessment if it has automated nodule detection and volume measurement capability. Some of the software can also compare subsequent scans to automatically measure VDT. In some of the software, the parameters can be changed to adjust the nodule detection performance (thus varying the sensitivity and specificity for nodule detection). Some include an integrated Brock model calculator.

Some of the software may only be able to analyse images of CT scans that include the entire lung. Some may be indicated for use only with a specific type of CT scan (e.g. scans without contrast or LDCT) or in specified populations (e.g. people without symptoms suggestive of lung cancer or people aged ≥ 18 years).

Thirteen relevant technologies have been identified by NICE. *Table 1* lists the specific technologies included in this assessment. Further descriptions of these technologies can be found in *Appendix 2*. These are reproduced from the final scope issued by NICE.

**TABLE 1** Summary of the included technologies

| Product name (manufacturer) | CE mark | Available to the NHS | CT scan types | Detection | Volumetry |
|---|---|---|---|---|---|
| AI-Rad Companion Chest CT (Siemens) | Class IIa[a] | To be confirmed | Low dose, regular dose with and without contrast[a] | Yes[a] | Yes[a] |
| AVIEW LCS+ (Coreline Soft) | Class IIa[a] | Yes | Low dose[a] | Yes | Yes |
| ClearRead CT (Riverain Technologies) | Class IIa | Yes | Low dose, regular dose with and without contrast | Yes | Yes |
| contextflow SEARCH Lung CT (contextflow) | Class IIa | Yes | With and without contrast | Yes | Yes |
| InferRead CT Lung (Infervision) | Class IIa | Yes | Low dose, regular dose with and without contrast | Yes | Yes |

**TABLE 1** Summary of the included technologies (*continued*)

| Product name (manufacturer) | CE mark | Available to the NHS | CT scan types | Detection | Volumetry |
|---|---|---|---|---|---|
| JLD-01K (JLK, Inc.) | Class I | To be confirmed | Without contrast | Yes | Yes |
| Lung AI (Arterys) | Class IIa[a] | To be confirmed | Low dose, regular dose with and without contrast[a] | Yes[a] | Yes[a] |
| Lung Nodule AI (Fujifilm) | To be confirmed | To be confirmed | To be confirmed | Yes | Yes |
| qCT-Lung (Qure.ai) | Class I[a] | To be confirmed | Without contrast[a] | Yes[a] | Research only[a] |
| SenseCare-Lung Pro (SenseTime) | Class IIb[a] | To be confirmed | Without contrast[a] | Yes[a] | Yes[a] |
| Veolity (MeVis) | Class IIa | Yes | Low dose, regular dose with and without contrast | Yes | Yes |
| Veye Lung Nodules (Aidence) | Class IIb | Yes | Low dose, regular dose with and without contrast | Yes | Yes |
| VUNO Med-LungCT AI (VUNO) | Class IIa[a] | To be confirmed | Low dose[a] | Yes[a] | Yes[a] |

a Information only from public domain.
Source: Reproduced from final NICE scope.[21]

## Proposed position of the intervention in the diagnostic pathway

*Figure 2* shows the simplified process of diagnosing lung cancer. In people who have no known pulmonary nodules (use case 1), the diagnostic process usually begins with a chest CT scan, where pulmonary nodules are identified (a). After nodules are detected, the nodule management pathway in accordance with the 2015 BTS guidelines[12] depends on two main criteria: nodule type (solid or subsolid; c) and nodule size (diameter or volume; d). Depending on the predicted malignancy risk (e), the guidelines recommend discharge, further CT surveillance or further clinical work-up and treatment.

During imaging follow-up of previously identified lung nodules (use case 2), the presence and speed of growth (e.g. VDT; f) as well as changes in nodule morphology are then used to predict the risk of malignancy (g) and make a decision on further patient management (i.e. discharge, further CT surveillance or further clinical work-up and treatment).

Software capable of automatically detecting and analysing lung nodules on chest CT scan images could be used to assist radiologists or other healthcare professionals when they review scan images. This could increase the detection of lung nodules that need further investigation or CT surveillance but could also increase the detection of benign nodules and lead to unnecessary follow-up investigations or CT surveillance. The same software could also help in assessing the growth of previously identified nodules that are being monitored with CT surveillance. Use of the software may impact on the recognition and recording of those lung nodule characteristics that are important for decisions on appropriate follow-up. It may also affect the time it takes to review and report the CT scan images. Although the software can automatically detect and analyse lung nodules on a CT scan image, the healthcare professional reporting the scan is still expected to review the findings of the software, and therefore no clinical decisions will be based on findings of the software alone. However, healthcare professionals reviewing CT scans may vary in their confidence to over-rule software depending on their experience and specialty (e.g. thoracic radiologists vs. general radiologists).

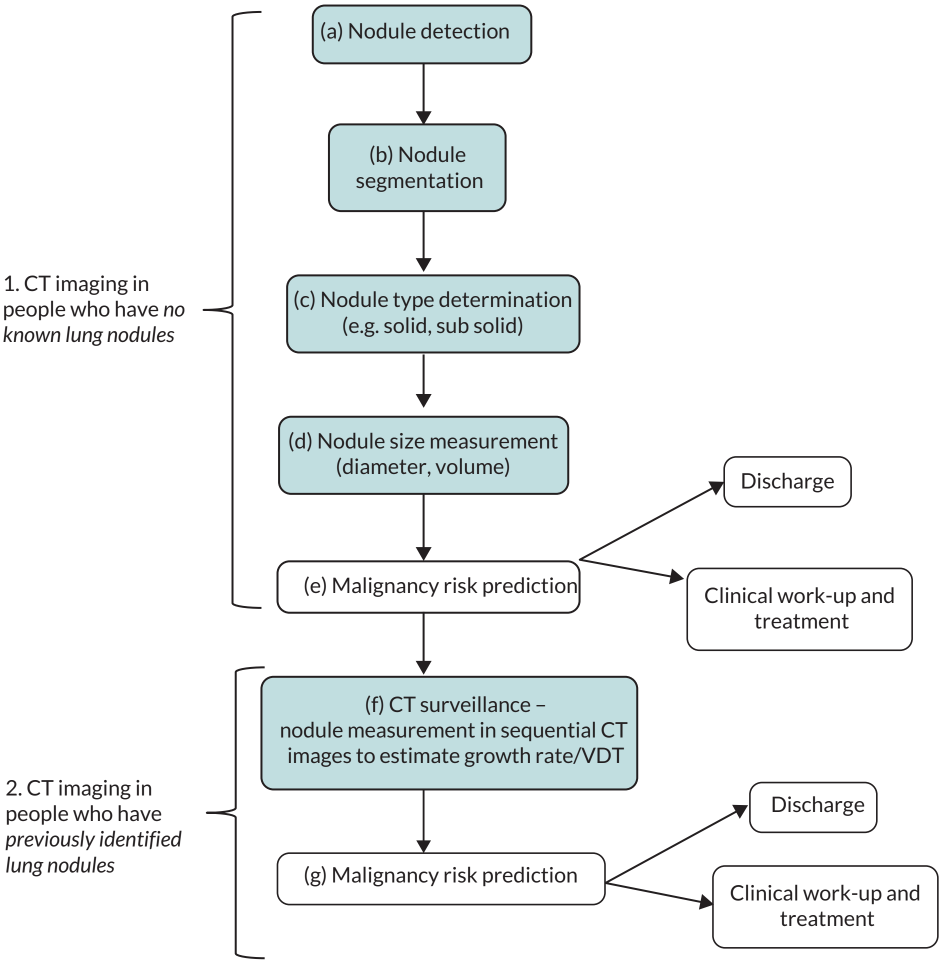

(a) Nodule detection
(b) Nodule segmentation
(c) Nodule type determination (e.g. solid, sub solid)
(d) Nodule size measurement (diameter, volume)
(e) Malignancy risk prediction
Discharge
Clinical work-up and treatment
1. CT imaging in people who have *no known lung nodules*
(f) CT surveillance – nodule measurement in sequential CT images to estimate growth rate/VDT
(g) Malignancy risk prediction
Discharge
Clinical work-up and treatment
2. CT imaging in people who have *previously identified lung nodules*

**FIGURE 2** Proposed roles of the intervention in the process of diagnosing lung cancer.

This diagnostic assessment considered the following specific locations in the diagnostic pathway where AI-based software for lung nodule detection and analysis could be used (shaded in blue in *Figure 2*):

1. CT images from people without previously identified lung nodules (use case 1)

   a. nodule detection
   b. nodule segmentation
   c. nodule type determination (solid or subsolid)
   d. nodule size measurement (diameter/volume).

2. CT images from people with previously detected lung nodules (use case 2)

   f. nodule size measurement in sequential CT images to estimate growth/VDT.

## Comparators

The comparator for this diagnostic assessment is review of chest CT scan images by a radiologist or another healthcare professional (e.g. a radiographer) without AI-based software for the automated detection and analysis of lung nodules. The reviewer of the scan may use software to help measure the volume of an identified lung nodule (see *Current methods of detecting nodules and measuring nodule volume and growth on computed tomography scans*), but this software does not automatically detect or measure lung nodules. When volumetric software is not used, nodule diameter is used to define the nodule size and nodule growth. The healthcare professional reviewing the scan may or may not be specialised in reviewing chest CT images.

During scoping, clinical experts highlighted that the experience of radiologists in reviewing CT scans for lung nodules will vary, for example between general, trauma and thoracic radiologists. They further commented that the expertise of the healthcare professional reviewing the scan may change the impact of the software. For example, less experienced reviewers may be more likely to act on nodules detected by the software, even if they disagree. For this reason, as highlighted in *Current methods of detecting nodules and measuring nodule volume and growth on computed tomography scans*, the standard protocol for the TLHC programme in England stipulates specific requirements for specialised readers reviewing the CT scans in the programme.[18]

## Outcomes

Key outcomes judged to be relevant to the assessment of the clinical effectiveness and cost-effectiveness of AI-based software for lung nodule detection and analysis, and the general diagnostic pathway for pulmonary nodules, are reported in detail in the study eligibility criteria for each key question (see *Study eligibility criteria*, *Identification and selection of studies* and *Identification and selection of studies*). In short, clinical effectiveness outcomes comprised test accuracy, reliability of the test, impact on patient management, practical implications and health outcomes. Health economic outcomes comprised incremental costs, incremental benefits, incremental cost-effectiveness ratio (ICER) and quality-adjusted life-years (QALYs). Owing to the limited nature of identified evidence base, many of these outcomes could only be evaluated using indirect evidence or could not be formally assessed.

## Objectives

The overall objectives of this diagnostic assessment are to assess the clinical effectiveness and cost-effectiveness of CT image analysis assisted by AI-based software capable of automated detection and analysis of lung nodules compared with unassisted CT image analysis in people undergoing CT scans of the chest due to symptoms suggestive of lung cancer, for purposes unrelated to suspicion of lung cancer, for lung cancer screening or for surveillance of previously identified lung nodules.

The key questions for this Diagnostic Assessment Report (DAR) are provided in *Box 1*.

**BOX 1** Key questions for this DAR

*Key question 1*

What is the accuracy of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules in people undergoing CT scans of the chest due to symptoms suggestive of lung cancer, for purposes unrelated to suspicion of lung cancer, for lung cancer screening or for surveillance of previously identified nodules, and what are the practical implications (e.g. test failure rate, reading time, acceptability) and the impact on patient management (e.g. stage of cancer detected, time to diagnosis, number of people referred for CT surveillance or having biopsy/excision)?

*Subquestions*

1. Does the accuracy of CT image analysis assisted by AI-based software for automated detection and analysis of lung nodules, its practical implications and impact on patient management differ between CT scans (1) with contrast and without contrast, (2) using a low-dose and a standard dose and (3) of solid nodules and subsolid nodules?
2. Does the accuracy of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules, its practical implications and impact on patient management differ by patient ethnicity?
3. Does the accuracy of CT image analysis assisted by AI-based software for automated detection and analysis of lung nodules, its practical implications and impact on patient management differ between general radiologists/health professionals and specialised thoracic radiologists/health professionals?
4. For the incidental population, does the accuracy of AI-based CT image analysis assisted by software for the automated detection and analysis of lung nodules, its practical implications and impact on patient management differ by reason for CT scan?

5.
   A. What is the concordance between readers with and without AI-based software support to detect and/or measure lung nodules from CT images?
   B. What is the concordance between readers using different AI-based software to detect and/or measure lung nodules from CT images?
   C. Does the use of AI-assisted CT image analysis impact on intra-observer and inter-observer variability in lung nodule detection and measurement?

*Key question 2*

What are the benefits and harms of using AI-based software for the automated detection and analysis of lung nodules from CT images compared with unassisted CT image analysis in people undergoing CT scans of the chest due to symptoms suggestive of lung cancer, for purposes unrelated to suspicion of lung cancer, for lung cancer screening or for surveillance of previously identified nodules?

*Subquestions*

1. Do the benefits and harms of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ between CT scans (1) with contrast and without contrast, (2) using a low-dose and a standard dose and (3) of solid nodules and subsolid nodules?
2. Do the benefits and harms of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ by patient ethnicity?
3. Do the benefits and harms of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ between general radiologists/healthcare professionals and specialised thoracic radiologists/healthcare professionals?
4. For the incidental population, do the benefits and harms of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ by reason for chest CT scan?

*Key question 3*

What is the cost-effectiveness of using AI-based software for the automated detection and analysis of lung nodules from CT images compared with unassisted CT image analysis in people undergoing CT scans of the chest due to symptoms suggestive of lung cancer, for purposes unrelated to suspicion of lung cancer, for lung cancer screening or for surveillance of previously identified nodules?

*Subquestions*

1. Does the cost-effectiveness of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ between CT scans (1) with contrast and without contrast, (2) using a low-dose and a standard dose and (3) of solid nodules and subsolid nodules?
2. Does the cost-effectiveness of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ by patient ethnicity?
3. Does the cost-effectiveness of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ between general radiologists/healthcare professionals and specialised thoracic radiologists/healthcare professionals?
4. For the incidental population, does the cost-effectiveness of CT image analysis assisted by AI-based software for the automated detection and analysis of lung nodules differ by reason for CT scan?

Ideally, priority would be given to 'end-to-end' studies that follow patients from testing through treatment to final health outcomes, such as morbidity and mortality. These studies can remove the need for separate searches for model parameters for cost-effectiveness modelling.[22] However, as no 'end-to-end' studies were found, we included and evaluated studies on test accuracy and practical implications, impact on patient management, costs and cost-effectiveness separately, and then synthesised the evidence using a linked evidence approach.[22]

# Chapter 2 Systematic review of test accuracy, practical implications and impact on patient management (key question 1): methods

Evidence required to address key question 1 was identified and assessed in a systematic review using the methods described in this chapter. The review followed the principles outlined in *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*[23] and NICE's *Diagnostics Assessment Programme Manual*.[22]

## Identification and selection of studies

### *Search strategy*

A comprehensive search was developed iteratively and undertaken in a range of relevant bibliographic databases. Searches combined keywords and, where appropriate, thesaurus (MeSH/EMTREE) terms relating to 'AI', 'lung nodules/lung cancer' and 'CT or screening'. Searches were limited to studies published in English as studies published in other languages were likely to be difficult to assess. No date limits were applied. An information specialist not otherwise involved in the project checked the draft MEDLINE search strategy for any omissions or errors. The final search strategies for all sources are provided in *Appendix 3*.

In January 2022, systematic searches were conducted in the following databases: MEDLINE All (via Ovid), EMBASE (via Ovid), Cochrane Database of Systematic Reviews (via Wiley), Cochrane CENTRAL (via Wiley), 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).

Records were exported to EndNote X9.3 (Clarivate Analytics, Philadelphia, PA, USA), where duplicates were systematically identified and removed.

To capture unpublished or ongoing studies, the MedRxiv preprint server (via the medrxivr app) and clinical trials registries [via ClinicalTrials.gov and the WHO International Clinical Trials Registry Platform (ICTRP) portal] were searched. The trials registry searches were highly focused, including search terms for the specific technologies of interest listed in the project scope and their manufacturing companies. Websites for the technologies and their manufacturers were also checked for further information, as were websites of selected organisations and conferences of interest (see *Appendix 3*). Reference lists of included studies and a selection of recent, relevant systematic reviews identified via the database searches were checked. Forwards citation tracking from key publications of included studies (to identify citing papers) was also undertaken, using Science Citation Index (Web of Science) and Google Scholar.

### *Study eligibility criteria*

Studies that satisfied the following criteria were included.

#### Population (all questions)

People who have no confirmed lung nodules or lung cancer and who are not having staging investigations or follow-up imaging for primary cancer elsewhere in the body, who have a CT scan that includes the chest:

- because of signs or symptoms suggestive of lung cancer (symptomatic population)
- for reasons unrelated to suspicion of lung cancer (incidental population)
- as part of lung cancer screening (screening population).

People having CT surveillance for a previously identified lung nodule (surveillance population).

Where data permit, the following subgroups may be considered:

- patient ethnicity
- people who have a CT scan (1) with or without contrast, (2) using a low-dose or a standard dose or (3) of solid nodules or subsolid nodules
- for the incidental population, by reason for CT scan.

### Target condition (all questions)

Lung nodules or lung cancer.

### Intervention (all questions)

CT scan review by a radiologist or another healthcare professional using any of the following software for the automated detection and analysis of lung nodules:

- AI-Rad Companion Chest CT (Siemens Healthineers, Erlangen, Germany)
- AVIEW LCS+ (Coreline Soft, Seoul, Republic of Korea)
- ClearRead CT (Riverain Technologies, Miamisburg, OH, USA) (indication: asymptomatic population only)
- contextflow SEARCH Lung CT (contextflow, Vienna, Austria) (indication: symptomatic population only)
- InferRead CT Lung (Infervision, Wiesbaden, Germany) (indication: asymptomatic population only)
- JLD-01K (JLK, Inc., Cheongju, Republic of Korea)
- Lung AI (Arterys, San Francisco, CA, USA)
- Lung Nodule AI (Fujifilm, Tokyo, Japan)
- qCT-Lung (Qure.ai, Mumbai, India)
- SenseCare-Lung Pro (SenseTime, Hong Kong)
- Veolity (MeVis, Bremen, Germany)
- Veye Lung Nodules (Aidence, Amsterdam, the Netherlands)
- VUNO Med-LungCT AI (VUNO, Seoul, Republic of Korea)

Where data permit, the following subgroups may be considered: general radiologist/other healthcare professional with AI-based software support versus radiologist/other healthcare professional with thoracic speciality with AI-based software support.

### Comparator (all questions)

CT scan review by a radiologist or another healthcare professional without AI-based software for the automated detection and analysis of lung nodules (using diameter or volume to measure nodule size) or no comparator.

Where data permit, the following subgroups may be considered: general radiologist/other healthcare professional without AI-based software support versus radiologist/other healthcare professional with thoracic speciality without AI-based software support.

### Reference standard

#### *Key question 1 and subquestions 1–4*

- Lung cancer confirmed by histological analysis of lung biopsy or health record review.
- CT surveillance (imaging follow-up) without significant growth, follow-up without lung cancer.
- Lung nodules: experienced radiologist reading (single reader or consensus of more than one reader)

#### *Subquestion 5*

None.

### Outcomes

#### *Key question 1 and subquestions 1–4*

- Accuracy to detect nodules (by nodule size and/or by nodule type; this may include, for example, the accuracy to detect nodules considered potentially significant as judged by experienced radiologist(s) and the accuracy to detect malignant nodules, respectively).

- Accuracy to measure diameter or volume of nodule or change in volume (when interventions are used as part of CT surveillance).
- Characteristics of detected nodules (e.g. size, type, location, spiculation).
- Proportion of detected nodules that are malignant.
- Technical failure rate.
- Radiologist reading time.
- Radiology report turnaround time.
- Impact of test result on clinical decision-making.
- Number of people having CT surveillance (this may also include, for example, the number of people with false-positive nodules having unnecessary CT surveillance).
- Number of CT scans taken as part of CT surveillance (this may also include, for example, the number of unnecessary CT surveillance scans resulting from false-positive nodules).
- Number of people having a biopsy or excision (this may also include, for example, the number of people with a negative biopsy resulting from false-positive nodules).
- Number of cancers detected.
- Stage of cancer at detection.
- Time to diagnosis.
- Acceptability and experience of using the software.

### Subquestion 5

- Concordance between readers with and without AI-based software.
- Concordance between readers using different AI-based software.
- Concordance between different AI-based software without human involvement.
- Inter-observer variability (e.g. positive and negative agreement, Cohen's kappa).
- Repeatability/reproducibility.

### Study design (all questions)

- Prospective test accuracy studies.
- Retrospective test accuracy studies.
- Randomised controlled trials.
- Cohort studies.
- Historically controlled trials.
- Before-and-after studies.
- Retrospective multireader multicase (MRMC) studies.
- Qualitative studies for user experience/acceptability.

### Publication type (all questions)

- Peer-reviewed papers.
- Conference abstracts and manufacturer data will be included. Only additional outcome data that have not been reported in peer-reviewed full-text papers will be extracted and reported.

### Language (all questions)

- English.

Papers fulfilling the following criteria were excluded:

- Studies using PET-CT scan images or lung phantom images or in which > 10% of CT scans were performed in patients with a primary cancer outside the lung (staging).
- Studies using index tests other than those specified in the inclusion criteria.
- Studies with no relevant outcomes reported.
- Non-human studies.

- Letters, editorials and communications were excluded unless they reported outcome data that had not been reported elsewhere, in which case they were to be handled in the same way as conference abstracts.
- Articles not available in the English language.
- Articles published before 2012. This cut-off date was based on expert advice, and all 13 companies were contacted for confirmation that no evidence relevant to their technology under investigation had been published before 2012.

### *Study screening and selection*

Two reviewers (JG/AA) independently screened the titles and abstracts of records identified from the searches and documents submitted by companies through NICE. Any disagreements were resolved through discussion, or retrieval of the full publication. Potentially relevant publications were obtained and assessed independently by two reviewers (JG/AA). Disagreements were resolved by consensus, with the inclusion of a third reviewer (CS, YFC) when required. Records excluded at full-text stage were documented, along with the reasons for their exclusion (see *Report Supplementary Material 1*, *Tables 1 and 2*).

## Data extraction and risk-of-bias assessment

### *Data extraction strategy*

Data were extracted by one reviewer (JG/AA) and checked by a second reviewer (JG/AA). All data extractions were entered into a piloted electronic data collection form (see *Report Supplementary Material 2*). Any disagreements were resolved through consensus, with the inclusion of a third reviewer (CS, YFC) when required.

### *Assessment of study risk of bias*

The risk of bias of test accuracy studies was assessed using a modified QUADAS-2 tool[24] combined with the QUADAS-C tool for comparative studies.[25] The COSMIN Risk of Bias tool was used to assess the risk of bias of studies focusing on evaluating reliability and errors of measurements on a continuous scale (e.g. nodule size and volume), in which test accuracy was not derived,[26] and for studies of agreements/concordance between readers where a reference standard could not be defined. The quality appraisal tools used in this DAR are tailored to the specific topic (see *Report Supplementary Material 2*). Two reviewers (JG/AA) independently undertook risk of bias assessment and critical appraisal. Disagreements were resolved by consensus, with the inclusion of a third reviewer (CS, YFC) if required.

## Methods of analysis/synthesis

First, test accuracy results are grouped by software functionality, for example, nodule detection, classification of nodule type (solid vs. subsolid nodule), nodule diameter and volume measurements. Comparative evidence between different testing strategies (e.g. AI-assisted readers, stand-alone AI, unaided readers) is then presented in preference to non-comparative evidence (e.g. individual testing strategy vs. a reference standard). The key comparison of interest (AI-assisted readers vs. unaided readers) is presented first, followed by other comparisons. Test accuracy results are also reported according to study population, the technology being evaluated and the type of nodules being detected.

Accuracy results are treated as binary (e.g. nodule present/absent; solid/subsolid nodule). Original data extracted from the studies were used to construct 2 × 2 tables. Pairs of sensitivities and specificities are also displayed in a paired forest plot to demonstrate scatter and uncertainty. Studies are grouped by the software and reading mode (e.g. stand-alone software, software-assisted reader).

Given the substantial heterogeneity of the study population, technologies being evaluated, reader speciality and experiences, reference standards and test accuracy outcome used and other study design features, no meta-analysis was carried out and findings are summarised narratively, with the results of data extraction presented in structured tables and plotted in figures where feasible.

Additionally, where data were available, we presented subgroup data and undertook subgroup analyses by:

- patient ethnicity
- reason for CT scan (within the incidental population)
- CT scans with versus without contrast
- CT scans using different radiation doses (e.g. ultra-low dose, low dose, standard dose)
- solid nodules versus subsolid nodules
- general radiologist (or other healthcare professional) versus specialised thoracic radiologist (or other healthcare professional).

Reliability outcomes as well as outcomes on patient management and practical implications are reported according to study population and the technology being evaluated. If applicable, comparative evidence between different reading modes (e.g. AI-assisted readers vs. unaided readers) are presented in preference to non-comparative evidence (e.g. individual testing strategy).

# **Chapter 3** Systematic review of test accuracy, practical implications and impact on patient management (key question 1): results

This report contains reference to confidential information provided as part of the NICE Diagnostic Assessment process. This information has been removed from the report and the results, discussions and conclusions of the report do not include the confidential information. These sections are clearly marked in the report.

Findings of the systematic review and company submissions answering key question 1 are presented in this chapter.

## Description of the evidence

### *Results of literature search*

Electronic database searches yielded 6,330 results, of which 4,886 had been published since 2012. Twenty-two records were judged to be relevant to key question 1 (see *Appendix 1*, *Figure 17*). An additional eight relevant records were identified through contacting authors of potentially relevant articles (*n* = 1)[27], searching company websites (*n* = 2)[28,29], company submissions (*n* = 3)[30–32], reviewers' Google searches for a published version of an unpublished manuscript (*n* = 1)[33] and tracking registered clinical trials (*n* = 1)[34], so 30 articles reporting 27 studies were included for key question 1.

The study by Murchison *et al.* is reported in two conference articles[28,29] and a journal article.[33] As the two conference articles from 2019 report only minimal additional information, in-text citations hereafter refer to the journal article by Murchison *et al.*[33] only. The study by Hall *et al.* is reported in a conference abstract[35] and a full journal article.[27] As the conference abstract from 2019 reports only minimal additional information, in-text citations hereafter refer to the journal article by Hall *et al.*[27] only.

Eleven articles evaluated relevant technologies but were excluded because the population comprised > 10% patients with extrathoracic cancer or previously diagnosed lung cancer.[36–46] These studies were not formally assessed, but the main study characteristics and outcome measures are summarised in *Report Supplementary Material 1*, *Table 3*.

### *Characteristics of included studies*

Twenty-seven studies were included for key question 1, evaluating 8 of the 13 NICE specified technologies (*Table 2*). Only two studies were conducted in the UK:

- AI-Rad Companion (Siemens Healthineers): three studies (USA, *n* = 2; Germany, *n* = 1)[47–49]
- AVIEW LCS+ (Coreline Soft): four studies (Republic of Korea, *n* = 3; Russia, *n* = 1)[32,50–52]
- ClearRead CT (Riverain Technologies): six studies [USA, *n* = 2; Taiwan (Province of China), *n* = 2; Japan, *n* = 1; Switzerland, *n* = 1][53–58]
- Contextflow SEARCH Lung CT (contextflow): one study (Austria, *n* = 1)[31]
- InferRead CT Lung (Infervision): three studies (China, *n* = 2; Japan, *n* = 1)[59–61]
- Veolity (MeVis): four studies [UK, *n* = 1; Republic of Korea, *n* = 2; USA (data)/Netherlands/Denmark (readers), *n* = 1][27,62–64]
- Veye Lung Nodules (Aidence): five studies (UK, *n* = 1; Netherlands, *n* = 3; USA, *n* = 1)[30,33,34,65,66]
- VUNO Med-LungCT AI (VUNO): one study [USA (data)/Republic of Korea (readers), *n* = 1].[67]

Sixteen studies were MRMC studies: eight[32,33,53,54,56,59,60,67] compared stand-alone AI software with human readers with and without concurrent AI software use under laboratory conditions. With 'concurrent' AI software use, the software results are simultaneously displayed to readers during the reading. (For brevity, in this report we describe human reading with concurrent AI software use as 'concurrent AI'.) The study by Hsu *et al.*[53] also assessed '2nd-read' AI

**TABLE 2** Characteristics of included studies (*n* = 27)

| Study, country | Study design | Target population | Index test | Comparator | Relevant outcomes reported |
|---|---|---|---|---|---|
| ***AI-Rad Companion (Siemens Healthineers) (three studies)*** | | | | | |
| Abadia *et al.* 2021,[47] USA | Retrospective test accuracy and MRMC study | Mixed (selected if ≥ 1 lung condition present and by nodule presence/ absence in radiology report) | [A] Stand-alone AI<br>[C] Concurrent AI (MRMC study) | [D] Unaided reader (MRMC study)<br>[E] Original radiologist reading (clinical practice) | Nodule detection accuracy<br>Nodule size measurement<br>Characteristics of nodules (FN, FP)<br>Reading times<br>Confidence in lung nodule detection |
| Chamberlin *et al.* 2021,[48] USA | Retrospective test accuracy study | Screening (random) | [A] Stand-alone AI | None | Nodule detection accuracy<br>Characteristics of detected nodules |
| Rückel *et al.* 2021,[49] Germany | Retrospective test accuracy study | Incidental (consecutive) | [A] Stand-alone AI | [E] Original radiologist reading (clinical practice) | Nodule detection accuracy<br>Characteristics of detected nodules |
| ***AVIEW LCS+ (Coreline Soft) (four studies)*** | | | | | |
| Hwang *et al.* 2021,[51] Republic of Korea | Before-and-after study | Screening (consecutive) | [A] Stand-alone AI for nodule detection<br>[B] 2nd-read AI for nodule detection<br>[C] Concurrent AI for nodule measurement and whole read including Lung-RADS categorisation | [E] Original radiologist reading (clinical practice) | Characteristics of detected nodules<br>% detected nodules being malignant<br>Nodule detection accuracy ([A])<br>Accuracy to detect lung cancer (whole read [C] with Lung-RADS)<br>Number of people with positive screening result<br>Technical failure rate |
| Hwang *et al.* 2021,[50] Republic of Korea | Retrospective analysis of prospective cohort study | Screening (consecutive) | [B] 2nd-read AI for nodule detection<br>[C] Concurrent AI for nodule measurement and whole read including Lung-RADS categorisation | None | Accuracy to detect lung cancer (whole read [C] with Lung-RADS)<br>Characteristics of detected nodules<br>% nodules being malignant<br>Number of people with positive screening result<br>Technical failure rate |
| Hwang *et al.* 2021,[52] Republic of Korea | Prospective screening cohort | Screening (consecutive) | [B] Assisted 2nd-read AI for nodule detection<br>[C] Concurrent AI for nodule measurement and whole read including Lung-RADS categorisation | None | Characteristics of detected nodules<br>Number of people having CT surveillance<br>Number of people having excision/biopsy<br>Technical failure rate |
| Lancaster *et al.* 2022,[32] Russia | MRMC study | Screening (nodule-enriched) | [A] Stand-alone AI<br>[C] Concurrent AI | [D] Unaided reader | Accuracy of nodule categorisation (< 100 $mm^3$, ≥ 100 $mm^3$)<br>Characteristics of detected nodules<br>Simulated radiologist workload reduction when stand-alone AI software would be used as pre-screen to rule out negative CT images |

continued

**TABLE 2** Characteristics of included studies ($n$ = 27) (*continued*)

| Study, country | Study design | Target population | Index test | Comparator | Relevant outcomes reported |
|---|---|---|---|---|---|
| ***ClearRead CT (Riverain Technologies) (six studies)*** | | | | | |
| Singh *et al*. 2021,[56] USA | MRMC study | Screening (nodule-enriched) | [A] Stand-alone AI-AD (with vessel suppression and autodetection of pulmonary nodules)<br>[C.1] Concurrent AI (with vessel suppression, without automatic nodule detection)<br>[C.2] Concurrent AI (with vessel suppression and autodetection of pulmonary nodules) | [D] Unaided reader | Nodule detection accuracy<br>Characteristics of detected nodules<br>Size measurement accuracy<br>Inter-observer agreement to detect the dominant nodule<br>Technical failure rate<br>Impact on clinical decision-making (change in Lung-RADS category) |
| Lo *et al*. 2018,[54] USA | MRMC study | Screening (nodule-enriched) | [A] Stand-alone AI<br>[C] Concurrent AI | [D] Unaided reader | Nodule detection accuracy<br>Radiologist reading time |
| Milanese *et al*. 2018,[55] Switzerland | MRMC study | Unclear (consecutive) | [C] Concurrent AI (vessel-suppressed CT images) using semiautomatic segmentation software (MM Oncology, Siemens Healthcare) | [D] Unaided reader (standard CT images) using semiautomatic segmentation software (MM Oncology, Siemens Healthcare) | Measurement accuracy<br>Inter-reader variability in nodule measurement<br>Impact on clinical decision-making (categorisation according to Fleischner Society guidelines[68]) |
| Hsu *et al*. 2021,[53] Taiwan | MRMC study | Mixed: clinical routine; screening (consecutive) | [A] Stand-alone AI<br>[B] Assisted 2nd-read AI<br>[C] Concurrent AI | [D] Unaided reader | Nodule detection accuracy<br>Radiologist reading time |
| Takaishi *et al*. 2021,[57] Japan | MRMC study | Mixed (unclear selection) | [C] Concurrent AI | [D] Unaided reader | Nodule detection accuracy<br>Reading time |
| Wan *et al*. 2020,[58] Taiwan | MRMC study | Mixed (selected only patients with subsequent nodule excision) | [A] Stand-alone AI | [D] Consensus of two radiologists measuring diameter manually | Nodule detection accuracy<br>Lung cancer detection accuracy<br>Characteristics of missed nodules<br>Measurement concordance between stand-alone AI and unaided reader consensus |
| ***Contextflow SEARCH Lung CT (contextflow) (one study)*** | | | | | |
| Röhrich *et al*. 2023,[31] Austria | MRMC study | Mixed (selected by presence/absence of diffuse parenchymal lung disease) | [C] Concurrent AI | [D] Unaided reader | Radiologist reading time<br>Technical failure rate |

**TABLE 2** Characteristics of included studies ($n$ = 27) (*continued*)

| Study, country | Study design | Target population | Index test | Comparator | Relevant outcomes reported |
|---|---|---|---|---|---|
| ***InferRead CT Lung (Infervision) (three studies)*** | | | | | |
| Kozuka *et al.* 2020,[59] Japan | MRMC study | Symptomatic (random) | [A] Stand-alone AI<br>[C] Concurrent AI | [D] Unaided reader | Nodule detection accuracy<br>Reading time<br>Characteristics of detected nodules |
| Liu *et al.* 2019,[60] China | MRMC study | Mixed (convenience sample) | Evaluation 1: [A] Stand-alone AI<br>Evaluation 4: [C] Concurrent AI | Evaluation 1: [D.1] Unaided reader<br>Evaluation 4: [D.2] Unaided reader | Nodule detection accuracy<br>Comparison of AI performance by radiation dose<br>Radiologist reading time |
| Zhang *et al.* 2021,[61] China | Retrospective test accuracy study and MRMC study | Screening (consecutive) | [C] Concurrent AI (MRMC study) | [E] Original radiologist reading (clinical practice) | Nodule detection accuracy<br>Characteristics of detected nodules |
| ***JLD-01K (JLK, Inc.)*** | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | |
| ***Lung AI (Arterys)*** | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | |
| ***Lung Nodule AI (Fujifilm)*** | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | |
| ***qCT-Lung (Qure.ai)*** | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | |
| ***SenseCare-Lung Pro (SenseTime)*** | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | |
| ***Veolity (MeVis) (four studies)*** | | | | | |
| Cohen *et al.* 2017,[62] Republic of Korea | MRMC study | Surveillance (preoperative CT scan for subsolid nodules and subsequent surgical resection) (consecutive) | [C] Concurrent AI (FBP vs. MBIR algorithms) | None | Diameter and volume measurement:<br>Technical failure rate<br>Inter-observer variability<br>Repeatability/reproducibility<br>Concordance between readers with software: FBP vs. MBIR |

continued

**TABLE 2** Characteristics of included studies (*n* = 27) (*continued*)

| Study, country | Study design | Target population | Index test | Comparator | Relevant outcomes reported |
|---|---|---|---|---|---|
| Kim *et al*. 2018,[63] Republic of Korea | MRMC study | Surveillance (preoperative CT scan for subsolid nodules and subsequent surgical resection) (consecutive) | [C] Concurrent AI | [D] Unaided reader | Diameter measurement:<br>Concordance between readers with and without software<br>Inter-observer variability<br>Repeatability/reproducibility<br>Technical failure rate<br>Nodule classification by size of solid portion:<br>Inter-observer variability<br>Repeatability/reproducibility |
| Hall *et al*. 2022,[27] UK | Retrospective test accuracy study and MRMC study | Screening (consecutive) | [C] Concurrent AI (MRMC study) | [E] Original radiologist reading (clinical practice) | Nodule detection accuracy<br>Lung cancer detection accuracy<br>Impact on decision-making<br>Radiologist reading time<br>Software acceptability and experience<br>Proportion of scans referred for CT surveillance<br>Proportion of scans referred to MDT<br>Characteristics of missed nodules<br>% detected nodules being malignant |
| Jacobs *et al*. 2021,[64] USA, Denmark, the Netherlands | MRMC study | Screening (selected by Lung-RADS category) | [C] Concurrent AI | [D] Unaided reader | Lung-RADS categorisation:<br>Inter-observer variability<br>Repeatability/reproducibility<br>Radiologist reading time<br>Technical failure rate<br>Impact on decision-making |
| ***Veye Lung Nodules (Aidence) (five studies – one study considered confidential was removed)*** | | | | | |
| Blazis *et al*. 2021,[65] the Netherlands | Retrospective test accuracy study | Mixed (unclear selection) | [A] Stand-alone AI | None | Nodule detection accuracy |
| Hempel *et al*. 2022,[34] the Netherlands | MRMC study | Mixed (incidentally detected nodules or no nodules, with or without prior CT) | [C] Concurrent AI | [D] Unaided reader | Accuracy of BTS[12] grade categorisation<br>Characteristics of detected nodules<br>Radiologist reading time<br>Technical failure rate<br>Inter-observer variability |
| Martins Jarnalo *et al*. 2021,[66] the Netherlands | Retrospective test accuracy study | Mixed (random) | [A] Stand-alone AI | None | Nodule detection accuracy<br>Nodule type accuracy (solid, subsolid)<br>Size measurement accuracy<br>Characteristics of detected (TP, FP) and missed (FN) nodules<br>Technical failure rate<br>Software acceptability and experience |

**TABLE 2** Characteristics of included studies (*n* = 27) (*continued*)

| Study, country | Study design | Target population | Index test | Comparator | Relevant outcomes reported |
|---|---|---|---|---|---|
| Murchison *et al.* 2022,[33] UK | MRMC study | Mixed – clinical routine mimicking a screening population in age and smoking history (selected) | [A] Stand-alone AI<br>[C] Concurrent AI | [D] Unaided reader | Nodule detection accuracy<br>Nodule type determination accuracy<br>Measurement (volume, diameter):<br>Inter-observer variability<br>Concordance between stand-alone software and readers without software<br>Technical failure rate<br>Growth rate:<br>Nodule registration accuracy<br>Inter-observer variability<br>Concordance between stand-alone software and readers without software |
| ***VUNO Med-LungCT AI (VUNO) (one study)*** | | | | | |
| Park *et al.* 2022,[67] USA, Republic of Korea | MRMC study | Screening (nodule-enriched) | [A] Stand-alone AI<br>[C] Concurrent AI | [D] Unaided reader | Nodule detection and Lung-RADS categorisation:<br>Lung cancer detection accuracy<br>Concordance between stand-alone software and readers<br>Inter-observer variability<br>Impact on decision-making |

AI, artificial intelligence software; BTS, British Thoracic Society; CT, computed tomography; EAG, External Assessment Group; FBP, filtered back projection; FN, false negative; FP, false positive; Lung-RADS, Lung CT Screening Reporting And Data System; MBIR, model-based iterative reconstruction; MDT, multi-disciplinary team; MRMC, multi-case multi-reader study; TP, true positive.

DOI: 10.3310/JYTW8921

software use, where the human reader first assessed the CT images without AI software and then opened the software results, revised their assessment and made the final decision. One MRMC study[58] compared the performance of stand-alone software with that of unaided readers, and six studies[31,34,55,57,63,64] compared the performance of readers with and without concurrent software use, with both reading sessions performed under laboratory conditions. The remaining MRMC study[62] compared software-assisted nodule measurement in CT images reconstructed with both filtered back projection (FBP) and model-based iterative reconstruction (MBIR) algorithms without comparison with unaided readers.

Five studies were retrospective test accuracy studies evaluating the performance of stand-alone software only[30,48,65,66] or in comparison with original unaided reading (clinical practice).[49]

Three studies were classed as retrospective test accuracy studies as well as MRMC studies. One study[47] performed a MRMC study comparing stand-alone AI and readers with concurrent AI with unaided reading, and additionally used the original radiologist reports as comparator. The other two studies[27,61] compared readers with concurrent AI with reading performed under laboratory conditions with unaided radiologists in clinical practice.

Three studies reported prospective screening experiences: two studies[50,52] included only software-assisted reading, whereas the remaining study[51] was a before-and-after study that evaluated the performance of stand-alone software as well as that of the original readers before and after software was implemented.

Regarding the relevance to the four target populations in this DAR:

- Symptomatic population (*n* = 1): one study was performed in a randomly selected symptomatic population.[59]
- Incidental population (*n* = 1): one study included a consecutive incidental population.[49]
- Screening population (*n* = 11): 11 studies included screening populations, of which 6 used consecutive or random sampling[27,48,50–52,61] and 5 were nodule-enriched (selection by nodule presence/absence, resulting in a higher nodule prevalence than expected in this population).[32,54,56,64,67]
- Surveillance population (*n* = 2): two studies included surveillance populations with applicability concerns: these two studies were performed in the same hospital and included potentially overlapping populations of consecutive patients with previously detected subsolid nodules who underwent preoperative CT scans and subsequent surgical resection.[62,63]
- ‘Mixed population’ (*n* = 11): in 11 studies, there were various indications for the chest CT scan: 3 studies[30,53,66] included consecutive or random sampling, 1 study[60] used convenience sampling, 5 studies[31,33,34,47,58] included enriched populations, and in the remaining 2 studies[57,65] the sampling method was unclear. The reasons for the CT scan are reported in *Appendix 4*, *Table 48*, so readers can decide if they want to consider the evidence from mixed populations for one of the four target populations.
- ‘Unclear population’ (*n* = 1): in one study,[55] the indication for the chest CT scan was not reported.

To help navigate the results section, *Tables 38–40* in *Appendix 1* present the number of studies identified and study details for each prespecified outcome and provide a link to the corresponding section of the report.

## Methodological quality of the evidence

The methodological quality of 22 studies[27,30–34,47–51,53–61,65,66] that reported test accuracy outcomes was assessed using QUADAS-2[24] and, if applicable, QUADAS-C.[25]

Four studies[62–64,67] reported concordance or agreement outcomes, and their quality was assessed using the COSMIN Risk of Bias tool (see *Assessment of study risk of bias*).[26] For the remaining study,[52] no quality appraisal was performed as the relevant outcomes for the DAR were related neither to accuracy nor to reliability/measurement error.

### *Risk of bias and applicability concerns according to QUADAS-2 and QUADAS-C*

The QUADAS-2 and QUADAS-C assessment results for 22 studies are summarised in *Appendix 1*, *Figures 18–20*.

### Risk of bias

Sixteen of the 22 studies were comparative test accuracy studies. In 12 (75%) studies the risk of bias according to QUADAS-C was considered 'high' in three or more domains. Among the remaining six non-comparative test accuracy studies, in one (17%) study the risk of bias (QUADAS-2) was considered high in three or more domains. No comparative or non-comparative test accuracy study was rated as being at 'low' or 'unclear' risk of bias in all four domains. The number and proportion of studies with 'low', 'high' and 'unclear' risk of bias, respectively, are presented in *Appendix 1*, *Figure 19*, for all 22 studies as well as separately for the 16 comparative studies (QUADAS-C) and the 6 non-comparative studies (QUADAS-2). The risk of bias in the four QUADAS-2 domains is discussed in more detail below.

#### *Patient selection domain*

The risk of bias in the patient selection domain was rated as 'high' in 15 (68%) out of 22 studies. The main reasons are as follows:

- no consecutive or random sample – eight studies[31–34,47,54,56,60]
- case–control design not avoided – eight studies[31–34,47,54,56,58]
- systematic exclusion of 'easy to read' CT images (e.g. exclusion of patients without other, non-nodule-related lung conditions) – two studies[31,47]
- exclusions by nodule number per image or unjustified (not based on management guidelines or minimal software manufacturer threshold) exclusion of certain nodule sizes – six studies[30,32,34,53,58,66]
- systematic exclusion of patients with other non-nodule-related lung pathology that can mimic or mask lung nodules (exclusion of 'difficult to read' CT images, e.g. severe pulmonary fibrosis, diffuse bronchiectasis, extensive inflammatory consolidation, pneumothorax and massive pleural effusion) – five studies[33,34,53,57,59]
- no fully paired or randomised design used – one study.[51]

In the patient selection domain, four studies[27,50,55,65] (18%) were classified as being at unclear risk of bias, and the remaining three studies[48,49,61] (14%) were classified as being at low risk of bias.

#### *Index test domain*

In the index test domain, three studies[48,50,51] (14%) were classified as being as low risk of bias. In 16 studies (73%), the risk of bias in this domain was considered 'high' for the following reasons:

- readers assessed the chest CT images outside clinical practice (MRMC studies) – 14 studies[27,31–34,47,53–57,59–61]
- AI software threshold not clearly pre-set by company or not prespecified in methods – four studies.[30,33,60,65]

In three studies, the risk of bias was rated as 'unclear' for the following reasons:

- unclear if there was no repeated application of AI to any of the same CT images, or use of the same CT images or images from the same patients for training – one study[66]
- unclear if the threshold was prespecified – two studies.[49,58]

#### *Reference standard domain*

Twenty-one of the 22 studies used a reference standard for lung nodules, and six studies had a reference standard for lung cancer.

For lung nodules, six of the of 21 studies (29%) were rated as being at low risk of bias.[30,54,56,58–60] The remaining 15 studies (71%) were rated as being at high risk of bias for the following reasons:

- no majority or consensus reading of (at least) three experienced thoracic radiologists – 11 studies[27,31,47–49,51,53,55,57,61,66]
- reference standard reader(s) part of the index test(s) or not blinded to index test markings/decisions – 13 studies.[27,32,33,47–49,51,53,55,57,61,65,66]

For lung cancer detection, two out of six studies were rated as being at low risk of bias.[54,58] Two studies were classified as high risk of bias as medical records were used as reference standard,[50,51] and the clinicians undertaking the diagnostic follow-up tests were not blinded to the results of the index test.[50,51] In the remaining two studies, the risk of bias was rated as 'unclear' as it was not stated how benign nodules were followed up[57] and no details about the reference standard were reported.[27]

### *Flow and timing domain*

Among the 21 studies evaluating lung nodule detection accuracy, the risk of bias was rated as 'low' in 12 (57%)[31,32,34,47,49,53,54,57–60,61,66] and as 'unclear' in 1 (5%).[30] A high risk of bias was present in the remaining eight studies (38%) for the following reasons:

- significant exclusions (> 10%; cut-off value determined pragmatically) after the point of selecting the cohort – six studies[27,33,55,56,60,65]
- number of CT images excluded due to software processing failures (e.g. segmentation failures) not reported – three studies.[33,48,51]

In the six studies reporting on lung cancer detection accuracy, the risk of bias was rated as 'low' in one study,[58] 'unclear' in two studies[27,51] and 'high' in three studies[50,54,57] for the following reasons:

- not all patients received a reference standard – one study[50]
- not all patients received the same reference standard – two studies.[54,57]

## Applicability concerns

Overall, all 22 studies had high applicability concerns in at least two of the three domains (i.e. population, index test, reference standard). The number and proportion of studies with low, high and unclear applicability concerns are presented in *Appendix 1*, *Figure 20*, separately for each evaluated index test.

### *Patient selection domain*

Applicability was assessed separately for the four target populations (i.e. symptomatic, incidental, screening and surveillance). There were high concerns regarding the applicability of the research identified to all relevant UK target populations in 20 out of the 22 (91%) included studies. The main reasons for the high applicability concerns are as follows:

- not a consecutive or random sample of patients/CT images – nine studies[31–33,47,54,57,60,65,69]
- enriched sample (e.g. inclusion/exclusion by nodule number, nodule type and nodule size) – eight studies[31,32,47,53,54,58,66,69]
- inclusion/exclusion by age – one study[33]
- study not performed in the UK or another north-western European country – 14 studies[30,32,47,48,50,51,53,54,57–61,69]
- > 10% of included people have different indication for the CT scan than the target population – 11 studies[30,31,33,47,53,55,57,58,60,65,66]
- CT image acquisition details (dose, contrast use, slice thickness) are different from UK practice for target population – eight studies[30,32,33,53,57,58,65,66]
- age of screening populations not between 55 and 75 years – six studies[27,32,48,55,58,61]
- nodule size < 5 mm or > 30 mm maximal diameter; < 80 $\text{mm}^3$ in a surveillance population – one study.[55]

Only one study[49] was classified as having low applicability concerns for the ‘incidental’ population. In another study,[55] the applicability to the ‘incidental’ and ‘symptomatic’ populations was 'unclear' as it was not reported if > 10% of included people had a different indication for the CT scan than the target population.

### *Index test domain*

Concerns regarding the applicability of the index test or the comparator to the situation in the UK were classified as high in all 22 included studies. The main reasons were:

- use of any prototype software versions that did not later become the commercially available version (e.g. applicability not confirmed by the company) – two studies[47,49]

- integration of software into pathway not applicable to UK (e.g. stand-alone AI performance instead of concurrent or 2nd-read software use) – 12 studies[30–33,47,51,58–60,65,66,69]
- reader had no access to maximum intensity projections and/or multiplanar reformations – six studies[31,33,54,57,60,69]
- study did not use a prespecified nodule size threshold similar to the UK 2015 BTS guidelines (i.e. ≥ 5 mm maximum axial diameter or ≥ 80 mm$^3$)[12] – 14 studies[31,32,48,50,51,53,54,57,58,60,61,65,66,69]
- other nodule types used than in the 2015 BTS guidelines (nodule type should be classified as solid, part-solid or pure ground glass)[12] – one study[66]
- for stand-alone AI, false-positive rate set to more than two false positives per image – three studies[30,33,65]
- for concurrent and assisted 2nd-read software use, more than one human reader involved per read – one study[61]
- for the unaided reader (comparator), human double reading instead of single human reader – two studies[27,61]
- human reader's experience and/or specialty not representative of UK clinical practice (five years training for radiologists, after which time they are considered 'fully trained') for target population – eight studies[27,31,53–55,57,59,61]
- software only had vessel suppression function, not nodule detection and measurement functions – one study.[55]

### *Reference standard domain*

Applicability concerns regarding the reference standard for lung nodules (21 studies) were rated as 'low' in three studies[27,33,55] and 'unclear' in one study.[31] The remaining 17 studies (81%) were rated as having high applicability concerns for the following reasons:

- for 'actionable' nodule present/absent, different nodule size from BTS 2015 guidelines definition ('actionable nodule' is ≥ 5 mm maximum axial diameter or ≥ 80 mm$^3$)[12] – 17 studies[30–32,47–49,51,53,54,57–61,65,66,69]
- other types used than in the BTS 2015 guidelines (nodule type should be classified as solid, part-solid or pure ground glass)[12] – one study[66]
- for nodule size measurement (volume/diameter), nodule size not measured as volume or, if volumetry segmentation is not possible, as maximum axial diameter – two studies.[55,58]

Applicability concerns regarding the reference standard for lung cancer (six studies) were rated as 'low' in two studies.[54,58] Two studies[27,57] were rated as having unclear applicability concerns as no details of the reference standard were reported in one study,[27] whereas in the other study[57] it was unclear if benign nodules were followed up for at least two years without lung cancer diagnosis. The remaining two studies[50,51] had high applicability concerns as there was no follow-up for at least two years for discharged patients (i.e. not receiving CT surveillance or biopsy/excision).

### *Risk of bias in reliability and measurement error (COSMIN tool)*

The COSMIN Risk of Bias tool[26] was used to assess the methodological quality of four studies[62–64,67] in terms of reliability and measurement error of outcome measurement instrument. All four studies received 'doubtful' final risk of bias ratings. The main reasons were 'doubtful' ratings for the following signalling questions:

- Was the time interval between the repeated measurements appropriate? - one study[62]
- Were there any other important flaws in the design or statistical methods of the study? - four studies[62–64,67]
- For continuous scores, were the standard error of measurement, smallest detectable change, limits of agreement or coefficient of variation calculated? - three studies[63,65,67]

## Use case 1: nodule detection and analysis in people with no known lung nodules

### *Nodule detection*

In this section we summarise the findings related to accuracy for nodule detection. Three main outcomes (targets of detection) are presented in each of the subsections: detection of any nodules, detection of actionable nodules and detection of malignant nodules (*Figure 3*). In each subsection, we focus on providing an overall summary of the comparative evidence between AI-assisted detection and unaided detection by human readers (the main comparison of interest in this DAR). Detailed descriptions of evidence from individual studies are provided in *Appendix 5*. Additional evidence on comparisons between stand-alone AI and unaided readers and non-comparative evidence, such as the accuracy of AI-assisted detection or detection by stand-alone AI compared with a reference standard, is presented *Appendix 6*.

Key characteristics, reported outcome measures and quality ratings for studies reporting comparative and non-comparative results are shown in *Tables 3* and *4*, respectively.

### Accuracy for identifying any nodules

a. Comparative results (seven studies)

Seven comparative studies[47,49,53,57,59–61] evaluated the accuracy for detecting any nodules. Of these, one included a screening population,[61] one included a symptomatic population,[59] one included an incidental population[49] and four included mixed populations.[47,53,57,60] The study by Hsu *et al.*[53] also reported accuracy data separately for the screening population subset.

Four of the comparative studies provided evidence on the comparison between AI-assisted reading and unaided reading, and the findings are presented in *Figure 4*. Reported sensitivity of AI-assisted reading (range 0.38–0.99) and unaided reading (range 0.21–0.72) varies widely between different studies, highlighting the heterogeneous nature of these studies. AI-assisted reading improved sensitivity compared with unaided readers across all studies, while the reported specificity for AI-assisted reading slightly worsened in two studies[59,61] and slightly improved in one study[53] compared with unaided readers. Findings from Kozuka *et al.*[59] show that the per-person sensitivity tends to be higher than the per-nodule sensitivity, but the differences between reading with and reading without AI support remain similar (*Figure 4*). Further details from individual studies are provided in *Appendix 5*.

### Accuracy for detecting actionable nodules

a. Comparative results (seven studies)

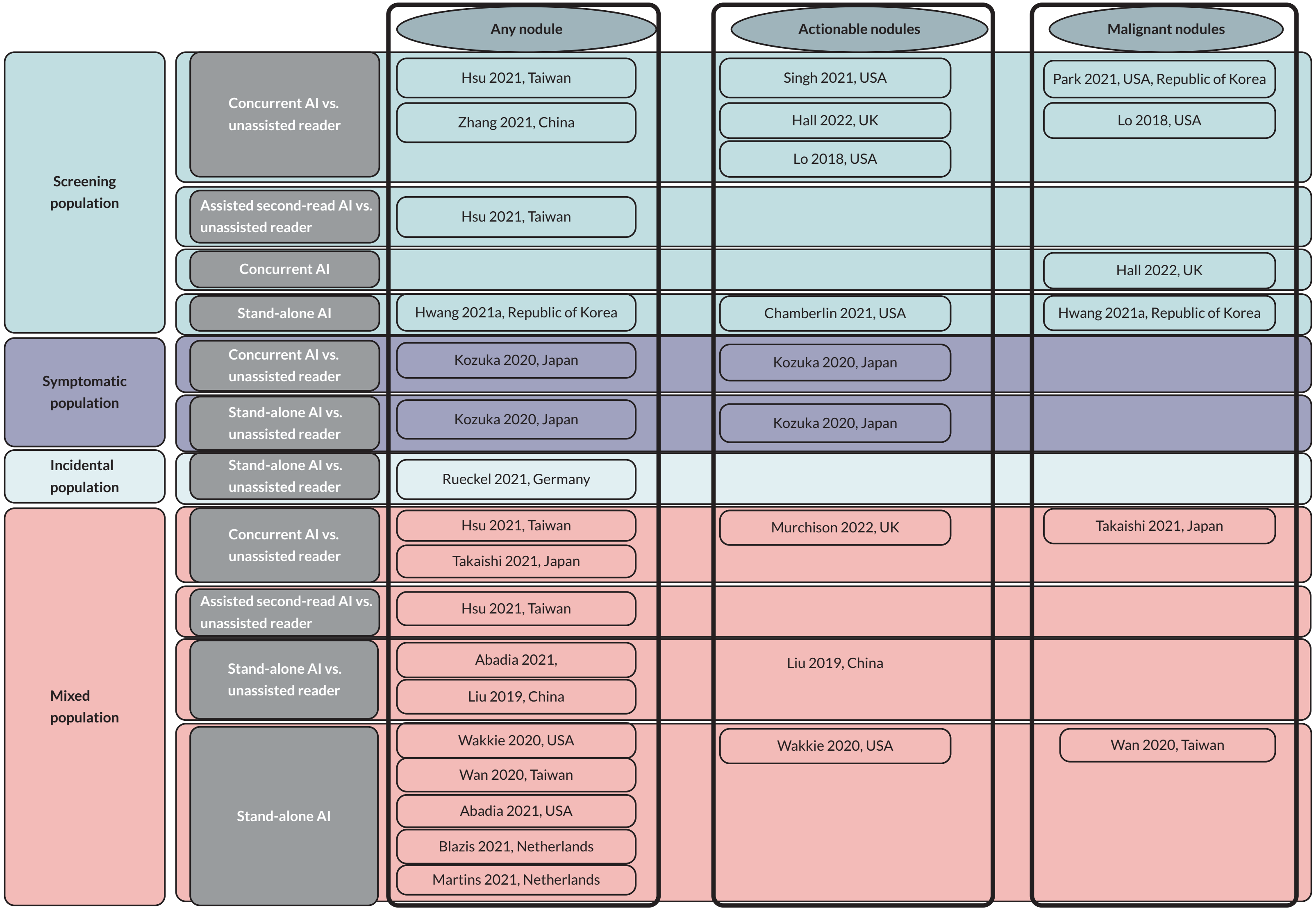
Six comparative studies[27,33,54,56,59,60] evaluated the accuracy for detecting actionable nodules (≥ 5 or 6 mm). Of these, three included a screening population,[27,54,56] one included a symptomatic population[59] and two included mixed populations.[33,60] Only one study[27] reported per-person analysis. Key results reported in these studies are shown in *Figure 5*. Reported sensitivity for concurrent AI ranged from 0.52 to 0.80 and was consistently higher than sensitivity for unaided readers of comparable experience (range 0.39–0.73). Only a small number of studies reported specificity or the number of false-positive detections per image. Where reported, the specificity was consistently lower, and false-positive detections per image were consistently higher, for concurrent AI than for unaided readers (*Figure 5*). Further details from individual studies are provided in *Appendix 5*.

One UK study[27] based on the Lung Screen Uptake Trial (LSUT) compared the use of concurrent AI by two radiographers (qualified in chest radiograph reporting but without prior experience in thoracic CT reporting) under research conditions with original reporting by experienced radiologists (5–28 years of experience in thoracic imaging, 5% double reading) without AI assistance. Both sensitivity (0.71 vs. 0.91) and specificity (0.92 vs. 0.97) were lower for AI-assisted, inexperienced radiographers than for unassisted, experienced radiologists (*Figure 5*).

### Accuracy for detecting malignant nodules

Evidence related to the accuracy for detecting malignant nodules is summarised in *Table 5*. It is worth highlighting that direct detection or classification of malignant nodules by AI-assisted reading is outside the scope of this assessment. The results presented in this section reflect the performance of AI-assisted reading or unassisted reading in identifying malignant nodule through the detection of actionable nodules and/or subsequent nodule management based on clinical guidelines following nodule detection.

Only one study[54] compared AI-assisted reading with unassisted reading and reported both sensitivity and specificity. The study found that sensitivity substantially increased (0.80 vs. 0.65) but specificity decreased (0.84 vs. 0.90) with



| | | Any nodule | Actionable nodules | Malignant nodules |
|---|---|---|---|---|
| Screening population | Concurrent AI vs. unassisted reader | Hsu 2021, Taiwan<br>Zhang 2021, China | Singh 2021, USA<br>Hall 2022, UK<br>Lo 2018, USA | Park 2021, USA, Republic of Korea<br>Lo 2018, USA |
| | Assisted second-read AI vs. unassisted reader | Hsu 2021, Taiwan | | |
| | Concurrent AI | | | Hall 2022, UK |
| | Stand-alone AI | Hwang 2021a, Republic of Korea | Chamberlin 2021, USA | Hwang 2021a, Republic of Korea |
| Symptomatic population | Concurrent AI vs. unassisted reader | Kozuka 2020, Japan | Kozuka 2020, Japan | |
| | Stand-alone AI vs. unassisted reader | Kozuka 2020, Japan | Kozuka 2020, Japan | |
| Incidental population | Stand-alone AI vs. unassisted reader | Rueckel 2021, Germany | | |
| Mixed population | Concurrent AI vs. unassisted reader | Hsu 2021, Taiwan<br>Takaishi 2021, Japan | Murchison 2022, UK | Takaishi 2021, Japan |
| | Assisted second-read AI vs. unassisted reader | Hsu 2021, Taiwan | | |
| | Stand-alone AI vs. unassisted reader | Abadia 2021,<br>Liu 2019, China | Liu 2019, China | |
| | Stand-alone AI | Wakkie 2020, USA<br>Wan 2020, Taiwan<br>Abadia 2021, USA<br>Blazis 2021, Netherlands<br>Martins 2021, Netherlands | Wakkie 2020, USA | Wan 2020, Taiwan |

**FIGURE 3** Visual map of included studies for detection accuracy based on population, comparison and reported outcomes (targets of detection).

**TABLE 3** Characteristics of included studies with comparative results for nodule detection accuracy, and their quality ratings (*n* = 12 studies)

| | | | | | | | Sensitivity/specificity/FP per scan | | | Quality of study | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Study, country** | **Population** | **Reading mode** | **Study design** | **Reader details** | **Nodule type** | **Nodule size** | **Any nodule** | **Actionable nodules** | **Malignant nodules** | **Risk of bias (QUADAS-C)** | **Applicability concerns** |
| Zhang *et al.* 2021,[61] China, InferRead CT Lung (Infervision) | Screening population | Concurrent AI vs. unassisted reader | Retrospective test accuracy study and MRMC study | One radiology resident with supervision from one experienced radiologist[a] | Solid, part-solid, GGN | Solid:<br>≤ 5 mm,<br>6–7 mm,<br>8–14 mm,<br>≥ 15 mm<br>GGN and part-solid: all sizes | Sensitivity (per patient)/ specificity (per patient) | | | P: low<br>I: high<br>RS (*N*): high<br>F&T (*N*): low | P: high<br>I: high<br>RS (*N*): high |
| Kozuka *et al.* 2020,[59] Japan, InferRead CT Lung (Infervision) | Symptomatic population | Concurrent AI vs. unassisted reader; stand-alone AI vs. unassisted reader | MRMC study | Two less experienced radiologists (1 and 5 years of diagnostic experience) | Any, solid, part-solid, GGN, calcified | ≥ 3 mm (3–6 mm, 6–10 mm, 10–15 mm, 15–20 mm, > 20 mm) | Sensitivity (per nodule)/ FP per scan; sensitivity (per patient)/ specificity (per patient) | Sensitivity (per nodule)/ FP per scan | | P: high<br>I: high<br>RS (*N*): low<br>F&T (*N*): low | P: high<br>I: high<br>RS (*N*): high |
| Takaishi *et al.* 2021,[57] Japan, ClearRead CT (Riverain Technologies) | Mixed population | Concurrent AI vs. unassisted reader | MRMC study | Three radiologists with < 10 years of experience | Any | ≥ 4 mm | Sensitivity (per nodule)/ FP per scan | | Sensitivity (per nodule)/ FP per scan | P: high<br>I: high<br>RS (*N*): high<br>RS (C): unclear<br>F&T (*N*): low<br>F&T (C): high | P: high<br>I: high<br>RS (*N*): high<br>RS (C): unclear |
| Liu *et al.* 2019,[60] Evaluation 4, China, InferRead CT Lung (Infervision) | Mixed population | Concurrent AI vs. unassisted reader | MRMC study | Two radiologists with approximately 10 years of experience | Any | NR | AUC | | | P: high<br>I: high<br>RS (*N*): low<br>F&T (*N*): high | P: high<br>I: high<br>RS (*N*): high |
| Liu *et al.* 2019,[60] Evaluations 1–3, China, InferRead CT Lung (Infervision) | Mixed population | Stand-alone AI vs. unassisted reader | MRMC study | Two radiologists with 5 and 10 years of experience, respectively | Any, solid, subsolid | Solid:<br>≤ 6 mm,<br>> 6 mm<br>Subsolid:<br>≤ 5 mm,<br>> 5 mm | Sensitivity (per nodule)/ FP per scan | Sensitivity (per nodule) | | P: high<br>I: high<br>RS (*N*): low<br>F&T (*N*): high | P: high<br>I: high<br>RS (*N*): high |

**TABLE 3** Characteristics of included studies with comparative results for nodule detection accuracy, and their quality ratings (*n* = 12 studies) (*continued*)

| | | | | | | | Sensitivity/specificity/FP per scan | | | Quality of study | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Study, country** | **Population** | **Reading mode** | **Study design** | **Reader details** | **Nodule type** | **Nodule size** | **Any nodule** | **Actionable nodules** | **Malignant nodules** | **Risk of bias (QUADAS-C)** | **Applicability concerns** |
| Hsu *et al.* 2021,[53] Taiwan, InferRead CT Lung (Infervision) | Mixed population (screening population reported separately) | Concurrent AI vs. unassisted reader;<br>2nd-read AI vs. unassisted reader | MRMC study | Six readers<br>Junior group: three residents in radiology (1–2 years of CT experience and ≥ 6 months of chest CT experience)<br>Senior group: three experienced chest radiologists (5, 10 and 25 years of experience, respectively) | Any | 3–10 mm | Sensitivity (per nodule)/ specificity (per patient) | | | P: high<br>I: high<br>RS (*N*): high<br>F&T (*N*): low | P: high<br>I: high<br>RS (*N*): high |
| Abadia *et al.* 2021,[47] USA, AI-Rad Companion (Siemens Healthineers) | Mixed population | Stand-alone AI vs. unassisted reader | Retrospective test accuracy and MRMC study | Clinical practice: one of five single expert chest radiologists<br>MRMC study: one expert chest radiologist (15 years of experience) | Any | ≥ 4 mm | Sensitivity (per nodule)/ FP per scan (for stand-alone AI only); sensitivity (up to three largest nodules)/PPV | | | P: high<br>I: high<br>RS (*N*): high<br>F&T (*N*): low | P: high<br>I: high<br>RS (*N*): high |

continued

**TABLE 3** Characteristics of included studies with comparative results for nodule detection accuracy, and their quality ratings (*n* = 12 studies) (*continued*)

| | | | | | | | Sensitivity/specificity/FP per scan | | | Quality of study | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Study, country** | **Population** | **Reading mode** | **Study design** | **Reader details** | **Nodule type** | **Nodule size** | **Any nodule** | **Actionable nodules** | **Malignant nodules** | **Risk of bias (QUADAS-C)** | **Applicability concerns** |
| Rückel *et al.* 2021,[49] Germany, AI-Rad Companion (Siemens Healthineers) | Incidental population | Stand-alone AI vs. unassisted reader | Retrospective test accuracy study | Clinical practice: Single board-certified radiologist alone (17%), or commonly reported by a radiology resident and a board-certified radiologist (83%). 25 different radiology residents and 18 different board-certified radiologists | Any | NR | Sensitivity (per nodule and per patient)/FP/scan (for stand-alone AI only) | | | P: low<br>I: unclear<br>RS (*N*): high<br>F&T (*N*): low | P: low<br>I: high<br>RS (*N*): high |
| Singh *et al.* 2021,[56] USA, ClearRead CT (Riverain Technologies) | Screening population | Concurrent AI vs. unassisted reader | MRMC study | Two radiologists (5 years and 10 years of thoracic CT experience) | GGN, part-solid, subsolid | ≥ 6 mm | | Sensitivity (per nodule)/specificity (per patient) | | P: high<br>I: high<br>RS (*N*): low<br>F&T (*N*): high | P: high<br>I: high<br>RS (*N*): high |
| Lo *et al.* 2018,[54] USA, ClearRead CT (Riverain Technologies) | Screening population | Concurrent AI vs. unassisted reader | MRMC study | 12 general radiologists certified by the American Board of Radiology (6–26 years of experience) | Any | 5–44 mm | | Sensitivity (per patient)/specificity (per patient) | Sensitivity (per patient)/specificity (per patient) | P: high<br>I: high<br>RS (*N*): low<br>RS (C): low<br>F&T (*N*): low<br>F&T (C): high | P: high<br>I: high<br>RS (*N*): high<br>RS (C): low |
| Hall *et al.* 2022,[27] UK, Veolity (MeVis) | Screening population | Concurrent AI vs. unassisted reader | Retrospective test accuracy study and MRMC study | [C] Two radiographers without prior experience in chest CT reporting (MRMC study); [E] five radiologists (5–28 years of experience; 5% double reading) (clinical practice) | Any | ≥ 5 mm,<br>≥ 6 mm | | Sensitivity (per patient)/specificity (per patient) | | P: unclear<br>I: high<br>RS (*N*): high<br>F&T (*N*): high | P: high<br>I: high<br>RS (*N*): low |

**TABLE 3** Characteristics of included studies with comparative results for nodule detection accuracy, and their quality ratings ($n$ = 12 studies) (*continued*)

| | | | | | | | Sensitivity/specificity/FP per scan | | | Quality of study | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Study, country** | **Population** | **Reading mode** | **Study design** | **Reader details** | **Nodule type** | **Nodule size** | **Any nodule** | **Actionable nodules** | **Malignant nodules** | **Risk of bias (QUADAS-C)** | **Applicability concerns** |
| Murchison *et al*. 2022,[33] UK, Veye Lung Nodules (Aidence) | Mixed population | Concurrent AI vs. unassisted reader | MRMC study | Two thoracic radiologists (≥ 9 years of experience) | Any | 5–30 mm | | Sensitivity (per nodule)/ FP per scan | | P: high<br>I: high<br>RS (*N*): high<br>F&T (*N*): high | P: high<br>I: high<br>RS (*N*): low |
| Park *et al*. 2022,[67] USA, Republic of Korea, VUNO Med-LungCT AI (VUNO) | Screening population | Concurrent AI vs. unassisted reader | MRMC study | Five readers: one fourth-year resident and four board-certified radiologists (1, 4, 8 and 20 years of experience) | Any | NR | | | Sensitivity | Assessed by COSMIN risk of bias tool only (doubtful rating) | Not assessed |

F&T (C), flow and timing domain (lung cancer detection); F&T (*N*), flow and timing domain (lung nodule detection); FP, false positive; GGN, ground-glass nodules; I, index test domain; NR, not reported; P, patient selection domain; PPV, positive predictive value; RF (C), reference standard domain (lung cancer detection); RF (*N*), reference standard domain (lung nodule detection).

a [C] MRMC study: one radiology resident (5 years of experience) and one radiologist (20 years of experience); [E] clinical practice: a total of 14 radiology residents (2–5 years of experience) and 15 radiologists (10–30 years of experience).

**TABLE 4** Characteristics of included studies with non-comparative results for nodule detection accuracy and quality ratings (eight studies[a])

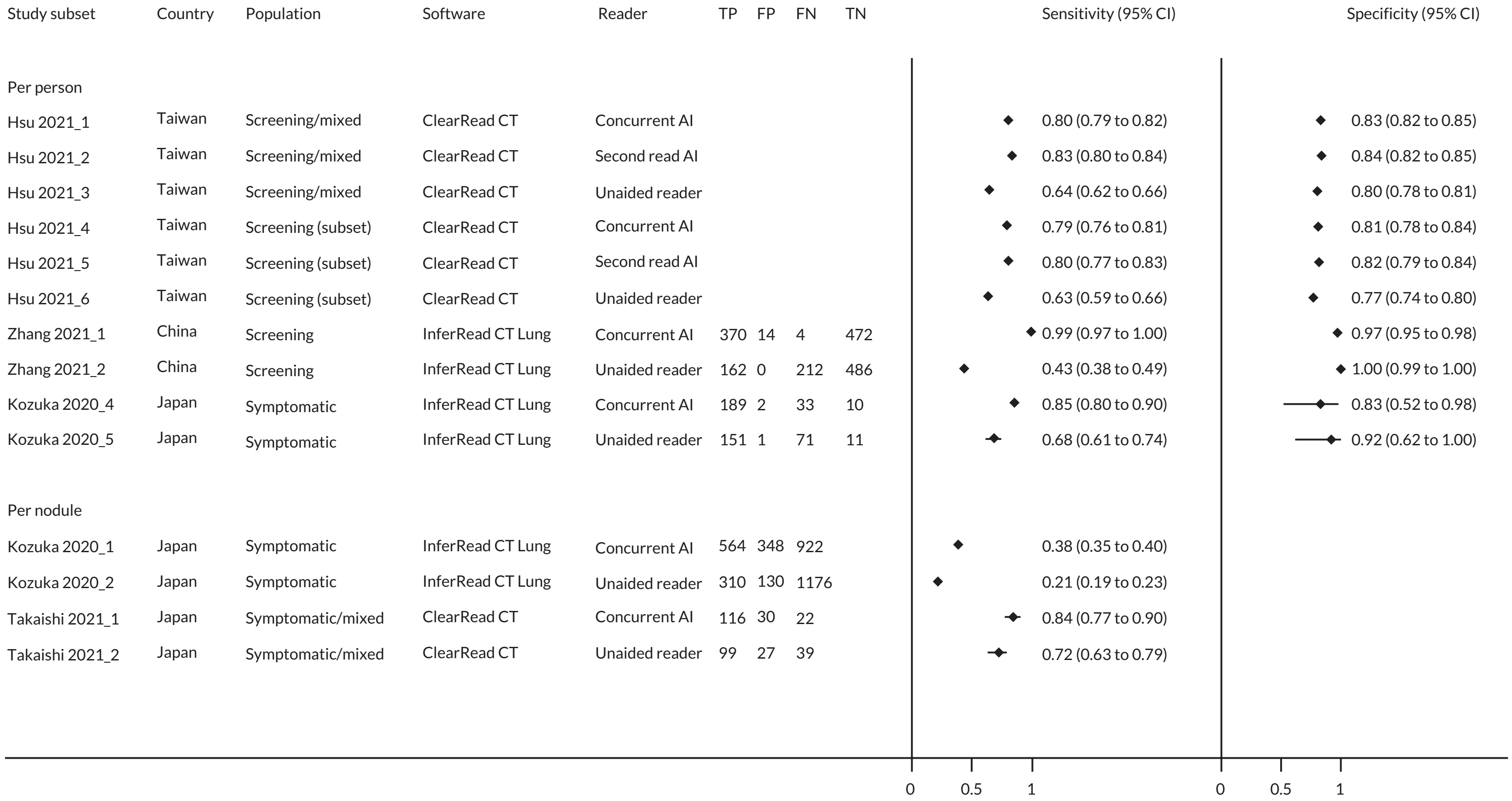
| | | | | | | Sensitivity/specificity/FP rate | | | Quality of study | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Study, country** | **Population** | **Reading mode** | **Study design** | **Nodule type** | **Nodule size** | **Any nodule** | **Actionable nodules** | **Malignant nodules** | **Risk of bias (QUADAS-2)** | **Applicability concerns** |
| Abadia *et al*. 2021,[47] USA, AI-Rad Companion (Siemens Healthineers) | Mixed population | Stand-alone AI | Retrospective test accuracy study | Any | ≥ 4 mm | Sensitivity (per patient)/specificity (per patient) | | | P: high<br>I: low<br>RS (*N*): high<br>F&T (*N*): low | P: high<br>I: high<br>RS (*N*): high |
| Chamberlin *et al*. 2021,[48] USA, AI-Rad Companion (Siemens Healthineers) | Screening population | Stand-alone AI | Retrospective test accuracy study | Any | > 6 mm | | Sensitivity (per nodule)/ FP per scan; sensitivity (per patient)/ specificity (per patient) | | P: low<br>I: low<br>RS (*N*): high<br>F&T (*N*): high | P: high<br>I: high<br>RS (*N*): high |
| Hwang *et al*. 2021,[51] Republic of Korea, AVIEW LCS+ (Coreline Soft) | Screening population | Stand-alone AI | Before-and-after study | Any, solid, GGN, part-solid | NR | Sensitivity (per nodule)/FP per scan | | Sensitivity (per nodule)/ FP per scan | P: unclear<br>I: low<br>RS (*N*): high<br>RS (C): high<br>F&T (*N*): high<br>F&T (*N*): unclear | P: high<br>I: high<br>RS (*N*): high<br>RS (C): high |
| Wan *et al*. 2020,[58] Taiwan, ClearRead CT (Riverain Technologies) | Mixed population | Stand-alone AI | MRMC study | Any | ≤ 2 cm | Sensitivity | | Sensitivity/ specificity | P: high<br>I: unclear<br>RS (*N*): low<br>RS (C): low<br>F&T (*N*): low<br>F&T (C): Low | P: high<br>I: high<br>RS (*N*): high<br>RS (C): low |
| Blazis *et al*. 2021,[65] the Netherlands, Veye Lung Nodules (Aidence) | Mixed population | Stand-alone AI | Retrospective test accuracy study | Any | > 4 mm or ≥ 30 mm$^3$ | Sensitivity (per nodule)/FP per scan | | | P: unclear<br>I: high<br>RS (*N*): high<br>F&T (*N*): high | P: high<br>I: high<br>RS (*N*): high |
| Martins Jarnalo *et al*. 2021,[66] the Netherlands, Veye Lung Nodules (Aidence) | Mixed population | Stand-alone AI | Retrospective test accuracy study | Any, solid, subsolid | 4–30 mm | Sensitivity (per nodule)/FP per scan | | | P: high<br>I: unclear<br>RS (*N*): high<br>F&T (*N*): low | P: high<br>I: high<br>RS (*N*): high |

**TABLE 4** Characteristics of included studies with non-comparative results for nodule detection accuracy and quality ratings (eight studies[a]) (*continued*)

| | | | | | | Sensitivity/specificity/FP rate | | | Quality of study | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Study, country** | **Population** | **Reading mode** | **Study design** | **Nodule type** | **Nodule size** | **Any nodule** | **Actionable nodules** | **Malignant nodules** | **Risk of bias (QUADAS-2)** | **Applicability concerns** |
| Hall *et al*. 2022,[27] UK, Veolity (MeVis) | Screening population | Concurrent AI | MRMC study: two radiographers without prior experience in chest CT reporting | Any | ≥ 5 mm | | | Sensitivity | P: unclear<br>I: high<br>RS (C): unclear<br>F&T (C): high | P: high<br>I: high<br>RS (C): unclear |

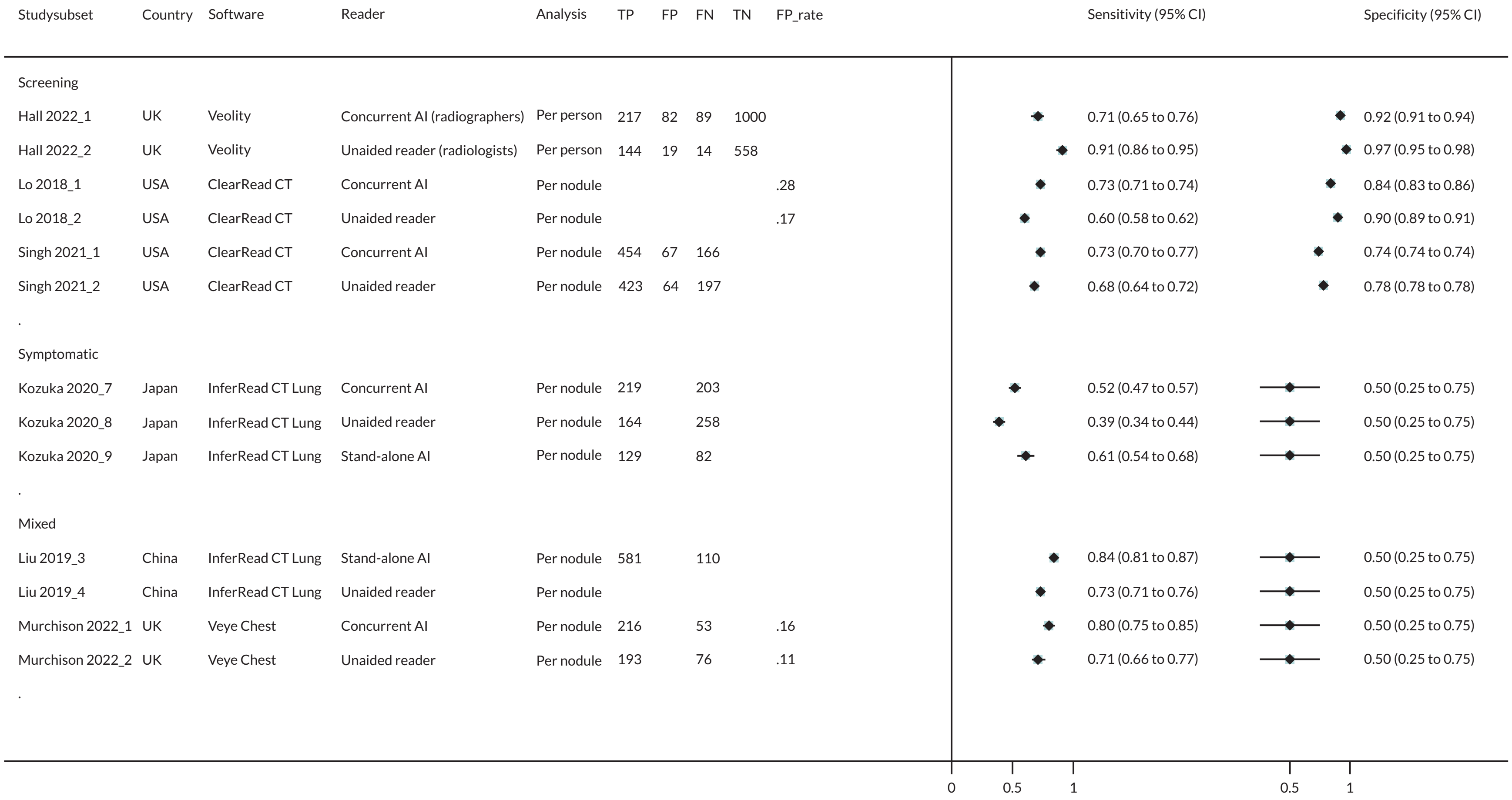
F&T (*N*), flow and timing domain (lung nodule detection); FP, false positive; GGN, ground-glass nodules; I, index test domain; NR, not reported; P, patient selection domain; RF (*N*), reference standard domain (lung nodule detection).

a One study considered confidential was removed from the table.



| Study subset | Country | Population | Software | Reader | TP | FP | FN | TN | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|---|---|---|---|---|---|
| Per person | | | | | | | | | | |
| Hsu 2021_1 | Taiwan | Screening/mixed | ClearRead CT | Concurrent AI | | | | | 0.80 (0.79 to 0.82) | 0.83 (0.82 to 0.85) |
| Hsu 2021_2 | Taiwan | Screening/mixed | ClearRead CT | Second read AI | | | | | 0.83 (0.80 to 0.84) | 0.84 (0.82 to 0.85) |
| Hsu 2021_3 | Taiwan | Screening/mixed | ClearRead CT | Unaided reader | | | | | 0.64 (0.62 to 0.66) | 0.80 (0.78 to 0.81) |
| Hsu 2021_4 | Taiwan | Screening (subset) | ClearRead CT | Concurrent AI | | | | | 0.79 (0.76 to 0.81) | 0.81 (0.78 to 0.84) |
| Hsu 2021_5 | Taiwan | Screening (subset) | ClearRead CT | Second read AI | | | | | 0.80 (0.77 to 0.83) | 0.82 (0.79 to 0.84) |
| Hsu 2021_6 | Taiwan | Screening (subset) | ClearRead CT | Unaided reader | | | | | 0.63 (0.59 to 0.66) | 0.77 (0.74 to 0.80) |
| Zhang 2021_1 | China | Screening | InferRead CT Lung | Concurrent AI | 370 | 14 | 4 | 472 | 0.99 (0.97 to 1.00) | 0.97 (0.95 to 0.98) |
| Zhang 2021_2 | China | Screening | InferRead CT Lung | Unaided reader | 162 | 0 | 212 | 486 | 0.43 (0.38 to 0.49) | 1.00 (0.99 to 1.00) |
| Kozuka 2020_4 | Japan | Symptomatic | InferRead CT Lung | Concurrent AI | 189 | 2 | 33 | 10 | 0.85 (0.80 to 0.90) | 0.83 (0.52 to 0.98) |
| Kozuka 2020_5 | Japan | Symptomatic | InferRead CT Lung | Unaided reader | 151 | 1 | 71 | 11 | 0.68 (0.61 to 0.74) | 0.92 (0.62 to 1.00) |
| Per nodule | | | | | | | | | | |
| Kozuka 2020_1 | Japan | Symptomatic | InferRead CT Lung | Concurrent AI | 564 | 348 | 922 | | 0.38 (0.35 to 0.40) | |
| Kozuka 2020_2 | Japan | Symptomatic | InferRead CT Lung | Unaided reader | 310 | 130 | 1176 | | 0.21 (0.19 to 0.23) | |
| Takaishi 2021_1 | Japan | Symptomatic/mixed | ClearRead CT | Concurrent AI | 116 | 30 | 22 | | 0.84 (0.77 to 0.90) | |
| Takaishi 2021_2 | Japan | Symptomatic/mixed | ClearRead CT | Unaided reader | 99 | 27 | 39 | | 0.72 (0.63 to 0.79) | |

**FIGURE 4** Evidence on AI-assisted reading compared with unaided reading for accuracy of detecting any nodules (four studies). FN, false negative; FP, false positive; TN, true negative; TP, true positive. Hsu 2021_4, 2021_5 and 2021_6 (*n* = 57 scans) were corresponding subsets of Hsu 2021_1, 2021_2 and 2021_3 (*n* = 93 scans) after non-screening mixed populations were excluded.



| Studysubset | Country | Software | Reader | Analysis | TP | FP | FN | TN | FP_rate | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Screening | | | | | | | | | | | |
| Hall 2022_1 | UK | Veolity | Concurrent AI (radiographers) | Per person | 217 | 82 | 89 | 1000 | | 0.71 (0.65 to 0.76) | 0.92 (0.91 to 0.94) |
| Hall 2022_2 | UK | Veolity | Unaided reader (radiologists) | Per person | 144 | 19 | 14 | 558 | | 0.91 (0.86 to 0.95) | 0.97 (0.95 to 0.98) |
| Lo 2018_1 | USA | ClearRead CT | Concurrent AI | Per nodule | | | | | .28 | 0.73 (0.71 to 0.74) | 0.84 (0.83 to 0.86) |
| Lo 2018_2 | USA | ClearRead CT | Unaided reader | Per nodule | | | | | .17 | 0.60 (0.58 to 0.62) | 0.90 (0.89 to 0.91) |
| Singh 2021_1 | USA | ClearRead CT | Concurrent AI | Per nodule | 454 | 67 | 166 | | | 0.73 (0.70 to 0.77) | 0.74 (0.74 to 0.74) |
| Singh 2021_2 | USA | ClearRead CT | Unaided reader | Per nodule | 423 | 64 | 197 | | | 0.68 (0.64 to 0.72) | 0.78 (0.78 to 0.78) |
| . | | | | | | | | | | | |
| Symptomatic | | | | | | | | | | | |
| Kozuka 2020_7 | Japan | InferRead CT Lung | Concurrent AI | Per nodule | 219 | | 203 | | | 0.52 (0.47 to 0.57) | 0.50 (0.25 to 0.75) |
| Kozuka 2020_8 | Japan | InferRead CT Lung | Unaided reader | Per nodule | 164 | | 258 | | | 0.39 (0.34 to 0.44) | 0.50 (0.25 to 0.75) |
| Kozuka 2020_9 | Japan | InferRead CT Lung | Stand-alone AI | Per nodule | 129 | | 82 | | | 0.61 (0.54 to 0.68) | 0.50 (0.25 to 0.75) |
| . | | | | | | | | | | | |
| Mixed | | | | | | | | | | | |
| Liu 2019_3 | China | InferRead CT Lung | Stand-alone AI | Per nodule | 581 | | 110 | | | 0.84 (0.81 to 0.87) | 0.50 (0.25 to 0.75) |
| Liu 2019_4 | China | InferRead CT Lung | Unaided reader | Per nodule | | | | | | 0.73 (0.71 to 0.76) | 0.50 (0.25 to 0.75) |
| Murchison 2022_1 | UK | Veye Chest | Concurrent AI | Per nodule | 216 | | 53 | | .16 | 0.80 (0.75 to 0.85) | 0.50 (0.25 to 0.75) |
| Murchison 2022_2 | UK | Veye Chest | Unaided reader | Per nodule | 193 | | 76 | | .11 | 0.71 (0.66 to 0.77) | 0.50 (0.25 to 0.75) |
| . | | | | | | | | | | | |

**FIGURE 5** Comparative evidence for accuracy of detecting actionable nodules (six studies). FN, false negative; FP, false positive; TN, true negative; TP, true positive.

AI-assisted reading compared with unassisted radiologist reading (*Table 5*). False-positive detections per image nearly doubled with AI-assisted reading (increase from 0.22 to 0.39).

The other five studies generally reported sensitivity of > 0.70 for the detection of malignant nodules with AI-assisted reading but did not provide information on specificity or false-positive detections per image. One study[58] reported high sensitivity (0.94) and low specificity (0.39) with stand-alone AI (*Table 5*). Further details from individual studies are provided in *Appendix 5*.

### Subquestions 1 to 4: potential factors influencing nodule detection accuracy

a. Subquestion 1-1: effect of contrast use.

No data was available to allow subgroup analysis based on contrast use.

b. Subquestion 1-2: effect of radiation dose (two studies).

Two studies performed in mixed populations from China[60] and Taiwan,[53] respectively, assessed the effect of radiation dose on nodule detection.

*Mixed population: ClearRead CT (Riverain Technologies) (one study)*

The study by Hsu *et al.*[53] reported accuracy results for the detection of any nodules for both standard-dose CT and LDCT images. It included 150 consecutive cases with lung nodules ≤ 1 cm or no nodules (93 standard-dose CT images from clinical routine and 57 LDCT images from lung cancer screening). Six readers participated in the MRMC study: three residents in radiology (junior group) and three experienced chest radiologists (senior group). For both AI-assisted and unaided reading, there was no significant difference between standard-dose and LDCT in terms of the mean sensitivity, specificity and area under the receiver operating curve for both junior and senior readers and all readers ($p > 0.05$; see *Appendix 1*, *Table 41*).

*Mixed population: InferRead CT Lung (Infervision) (one study)*

The study by Liu *et al.*[60] evaluated 187 LDCT and 942 standard-dose CT images. The deep-learning-based algorithm (InferRead CT Lung, Infervision) showed no dose-level dependence of nodule detection sensitivity ($x^2 = 1.1036$, $p = 0.9538$). The same result was observed for the two unaided radiologists (radiologist 1: $x^2 = 1.6562$, $p = 0.8944$; radiologist 2: $x^2 = 1.5293$, $p = 0.9097$). The false-positive rate of the stand-alone software was also independent of the dose ($x^2 = 0.5640$, $p = 0.4527$).

c. Subquestion 1-3: effect of nodule type (seven studies).

*Screening population: concurrent AI versus unaided reader (two studies)*

Two studies[56,61] reported detection accuracy for concurrent AI and unaided for different types of nodules (see *Appendix 1*, *Table 42*).

Zhang *et al.*[61] included 860 consecutive patients who underwent chest CT from November to December 2019 at one Chinese hospital as part of the Netherlands–China Big-3 disease screening (NELCIN-B3) project. One resident drafted the diagnostic report, and a board-certified radiologist supervised the final version without software use in clinical practice or with concurrent software use (InferRead CT Lung, Infervision) under laboratory conditions. The per-subject sensitivity of AI-assisted readers was 98.8% [95% confidence interval (CI) 96.5% to 99.8%] for solid nodules, 100.0% (95% CI 75.3% to 100.0%) for part-solid nodules and 99.1% (95% CI 95.1% to 99.9%) for ground-glass nodules. For the unaided readers in clinical practice, the per-subject sensitivity was 52.4% (95% CI 46.0% to 58.7%) for solid nodules, 23.1% (95% CI 5.0% to 53.8%) for part-solid nodules, and 25.2% (95% CI 17.5% to 34.4%) for ground-glass nodules.

**TABLE 5** Summary of evidence related to accuracy of AI-assisted reading and stand-alone AI for detecting malignant nodules (six studies)

| Study, country, image readers | Malignant nodules/ total scans | Measure of accuracy[a] | Index test[b] | Comparator[b] | Difference | p-value of difference |
|---|---|---|---|---|---|---|
| ***Screening population*** | | | | | | |
| Lo *et al.* 2018,[54] USA, 12 general radiologists (6–26 years) | 95/324 | Sensitivity | [C]: 0.800 (SD 0.039) | [D]: 0.647 (SD 0.039) | 0.154 (0.082 to 0.225) | < 0.0001 |
| | | Specificity | [C]: 0.844 (SD 0.020) | [D]: 0.899 (SD 0.020) | −0.055 (−0.090 to −0.019) | 0.0025 |
| | | False-positive detections per image | [C] 0.39 | [D] 0.22 | 0.17 (NR) | < 0.01 |
| Park *et al.* 2022,[67] USA/ Republic of Korea, five chest radiologists (1–20 years) | 31/200 | Sensitivity | [C]: 0.916 (0.817 to 0.964) | [D]: 0.852 (0.742 to 0.920) | 0.064 (NR) | 0.004 |
| Hwang *et al.* 2021,[51] Republic of Korea | 27/4666 | Sensitivity | [A] 0.704 (0.498 to 0.862) | NA | NA | NA |
| Hall *et al.* 2022,[27] UK, two radiographers | 33/716 | Sensitivity[c] | [C] 0.857 (0.746 to 0.933)[d] | NA | NA | NA |
| ***Mixed population*** | | | | | | |
| Takaishi *et al.* 2021,[57] Japan, three radiologists (2–8 years) | 1/61 | Sensitivity | [C] 1.00[e] | [D] 1.00[e] | 0 | NR |
| | | PPV | [C] 0.020 (1/49) | [D] 0.024 (1/42) | −0.004 | NR |
| Wan *et al.* 2020,[58] Taiwan | 47/50 | Sensitivity | [A] 0.936 (0.825 to 0.987) | NA | NA | NA |
| | | Specificity | [A] 0.393 (0.215 to 0.594) | NA | NA | NA |

NA, not applicable; NR, not reported; PPV, positive predictive value; SD, standard deviation.
a Data shown are based on per-nodule analysis unless otherwise indicated.
b [A]: Stand-alone AI; [C]: concurrent AI; [D]: unassisted reader.
c Per scan analysis.
d Calculated by review authors based on data provided in the original article.
e Only included one malignant nodule, which was detected by both concurrent AI and unaided reader.

**Notes**
Numbers shown in brackets are 95% CIs unless otherwise stated.
Technologies evaluated in the studies: Hall 2022: Veolity; Hwang 2021a: AVIEW LCS+; Lo 2018 and Takaishi 2021: ClearRead CT; Park 2022: VUNO Med-LungCT AI.

The per-subject specificity with concurrent software use was 99.2% (95% CI 98.1% to 99.7%) for solid nodules, 100.0% (95% CI 99.6% to 100.0%) for part-solid nodules and 98.8% (95% CI 97.7% to 99.5%) for ground-glass nodules. Without software use, the per-subject specificity was 100.0% (95% CI 99.4% to 100.0%) for solid, 100.0% (95% CI 99.6% to 100.0%) for part-solid and 100.0% (95% CI 99.5% to 100.0%) for ground-glass nodules.

With concurrent software use, the per-subject sensitivity and specificity seems not to vary by nodule type (95% CIs overlap), whereas without software use, the per-subject sensitivity for the detection of solid nodules seems to be higher than for part-solid nodules (the 95% CIs overlap, however) and ground-glass nodules (no overlap in 95% CIs). Concurrent software use seems to result in bigger sensitivity improvements for part-solid nodules (+76.9%) and ground-glass nodules (+73.9%) than for solid nodules (+46.4%).

Singh *et al.*[56] selected 150 LDCT from the National Lung Screening Trial (NLST): the first 125 patients with mixed attenuation or ground-glass nodules and the first 25 patients with no nodules. Two radiologists (with 5 and 10 years of thoracic CT experience) participated in a MRMC study to detect nodules ≥ 6 mm on vessel-suppressed CT images (ClearRead Vessel

Suppression, Riverain Technologies) as well as on standard CT images. However, the evaluated software did not possess nodule detection function. The study reported mean per-nodule sensitivities of 76% for part-solid nodules and 67% for ground-glass nodules on vessel-suppressed CT images. On standard CT images, the mean per-nodule sensitivities were 70% for part-solid and 67% for ground-glass nodules. The mean specificities were 85% for part-solid nodules and 78.5% for ground-glass nodules and 74% for all subsolid nodules on vessel-suppressed CT images (however, there might have been a mix-up in the table in the article). On standard CT images, the mean specificities were 76% for part-solid nodules, 84% for ground-glass nodules and 77.5% for all subsolid nodules.

*Symptomatic population: concurrent AI versus unaided reader (one study)*

Kozuka *et al.*[59] reported the per-nodule sensitivity of concurrent AI and unaided readers by nodule type (see *Appendix 1*, *Table 43*). This study was a retrospective analysis of 120 randomly selected chest CT images (117 cases included in analysis) from patients with suspected lung cancer. Two less experienced radiologists (1 and 5 years of experience) assessed the CT images with and without software use (InferRead CT Lung, Infervision). With software use, the pooled per-nodule sensitivities were 32.6% (95% CI 29.8% to 35.6%) for solid nodules, 58.4% (95% CI 49.5% to 67.0%) for part-solid nodules and 40.1% (95% CI 32.7% to 47.9%) for ground-glass nodules. In the unaided reading session, the pooled per-nodule sensitivity was 18.6% (95% CI 16.3% to 21.1%) for solid nodules, 31.5% (95% CI 23.7% to 40.3%) for part-solid nodules and 18.0% (95% CI 12.6% to 24.6%) for ground-glass nodules.

In contrast to the findings by Zhang *et al.*,[61] the study by Kozuka *et al.*[59] observed higher pooled per-nodule sensitivities for part-solid nodules than for solid nodules and ground-glass nodules, both with and without software use. Software use improved the pooled sensitivities by +14.0% for solid nodules ($p$ < 0.01), +26.9% for part-solid nodules ($p$ < 0.01), and +22.1% for ground-glass nodules ($p$ < 0.01) compared with the pooled unaided readers.

*Symptomatic population: stand-alone AI versus unaided reader (one study)*

Kozuka *et al.*[59] reported per-nodule and per-patient accuracy for stand-alone AI and unaided readers by nodule type (see *Appendix 1*, *Table 43*). This study was a retrospective analysis of 120 randomly selected chest CT images (117 cases included in analysis) from patients with suspected lung cancer. Two less experienced radiologists (1 and 5 years of experience) assessed the CT images with and without software use. For stand-alone AI (InferRead CT Lung, Infervision), the study observed per-nodule sensitivities of 68.1% (95% CI 63.9% to 72.1%) for solid nodules, 70.8% (95% CI 58.2% to 81.4%) for part-solid nodules and 72.1% (95% CI 61.4% to 81.2%) for ground-glass nodules. For the unaided readers, the pooled per-nodule sensitivity was 18.6% (95% CI 16.3% to 21.1%) for solid nodules, 31.5% (95% CI 23.7% to 40.3%) for part-solid nodules and 18.0% (95% CI 12.6% to 24.6%) for ground-glass nodules.

*Screening population: stand-alone AI (two studies)*

Hwang *et al.*[51] included 4,666 participants who had undergone lung cancer screening as part of the K-LUCAS (Korean Lung Cancer Screening) project after the implementation of the software AVIEW Lungscreen (Coreline Soft). The per-nodule sensitivity of stand-alone AI was 51% (95% CI 50% to 53%) for solid nodules, 49% (95% CI 36% to 61%) for part-solid nodules and 21% (95% CI 16% to 29%) for ground-glass nodules (see *Appendix 1*, *Table 44*).

The study by Lo *et al.*[54] included 324 LDCT (including 95 lung cancer cases) from the US-based NLST and two US hospitals; images with nodules (5–44 mm) and without nodules were selected in a ratio of 2 : 1. The per-nodule sensitivities of stand-alone AI (ClearRead CT, Riverain Technologies) were 84%, 85% and 67% for solid nodules, part-solid nodules and ground-glass nodules, respectively (see *Appendix 1*, *Table 44*).

*Mixed population: stand-alone AI alone (one study) or versus unaided reader (one study)*

Liu *et al.*[60] reported the per-nodule sensitivity of stand-alone AI (InferRead CT Lung, Infervision) as well as for two unaided readers for detecting nodules by type and size on conventional-dose CT and LDCT scans (see *Appendix 1*, *Table 45*). With LDCT, the per-nodule sensitivity of stand-alone AI was 71.9% for solid nodules ≤ 6 mm and 88.6% for solid nodules > 6 mm. With standard dose, the per-nodule sensitivity was 64.4% for solid nodules ≤ 6 mm and 87.9%

for solid nodules > 6 mm. When looking at subsolid nodules, the study observed that stand-alone software correctly detected 61.3% of nodules ≤ 5 mm and 85.2% of nodules > 5 mm on LDCT. With standard dose, the per-nodule sensitivity was 68.1% for subsolid nodules ≤ 5 mm and 81.1% for subsolid nodules > 5 mm.

Martins *et al.*[66] randomly selected 145 patients with 145 CT images from a large teaching hospital in the Netherlands. They reported 89.0% (65/73), 81.3% (13/16) and 100% (2/2) per-nodule sensitivity of stand-alone software (Veye Chest, Aidence) to detect solid, subsolid and mixed (solid/subsolid) nodules, respectively (see *Appendix 1*, *Table 45*).

d. Subquestion 2: effect of patient ethnicity

No subgroup analysis based on ethnicity was performed.

e. Subquestion 3: effect of radiologist speciality and experience (one study)

*Hsu et al. 2021,*[53] *Taiwan: ClearRead CT (Riverain Technologies)*

The study in a mixed population (with data for the screening subgroup reported separately) reported accuracy in detecting any nodules using concurrent AI compared with unaided reader for three residents in radiology (junior group; 1–2 years of CT experience and at least 6 months of chest CT experience) and three experienced chest radiologists (senior group; 5, 10 and 25 years of experience, respectively) separately. In the junior group, mean per-nodule sensitivity increased significantly from 52% (95% CI 47% to 57%) without software use to 74% (95% CI 70% to 78%) with concurrent AI ($p < 0.001$). The mean specificity did not change significantly and was 74% (95% CI 70% to 78%) with and 68% (95% CI 64% to 73%) without software use ($p = 0.442$). In the senior group, the mean per-nodule sensitivity increased significantly with concurrent software use from 73% (95% CI 69% to 77%) to 83% (95% CI 79% to 86%) ($p < 0.01$). The mean specificity was 88% (95% CI 85% to 91%) with and 86% (95% CI 83% to 90%) without software use ($p = 0.795$).

f. Subquestion 4: for the incidental population, effect of reason for CT scan

No study was identified that examined the accuracy of nodule detection by AI according to reasons for CT scan in the incidental population.

### Subquestion 5: concordance and variability in nodule detection

a. Concordance between readers with and without software (one study)

No study was identified that reported the concordance in nodule detection between readers with and without software use. However, one study reported the percentage agreement in nodule detection between stand-alone AI and the original unaided reading.[47]

*Mixed population: AI-Rad Companion CT Chest (Siemens Healthineers) (one study)*

Abadia *et al.*[47] found that across all included patients and lung conditions, the percentage of nodules found by the AI-Rad software that were also in the original radiology reports (original reading performed in clinical practice by one of five expert chest radiologists) was 75.8% (138/182). The highest agreement in nodule detection between AI-Rad software and the original radiology reports was achieved in the subpopulation with pulmonary embolism (87.2%; 34/39) and was lowest for patients with oedema (63.6%; 28/44).

b. Concordance between readers using different software (no study)

No study was identified that evaluated the agreement in nodule detection between readers using different AI-based software packages.

c. Intra-observer and inter-observer variability (one study)

One study reported on the inter-observer variability between unaided readers in the detection of the risk-dominant nodule.[56]

*Screening population: unaided readers (one study)*

The MRMC study by Singh *et al.*[56] found a Cohen's kappa of 0.63 for the detection of the risk-dominant nodule between the two unaided radiologists. Inter-observer agreement between the software-assisted radiologists assessing vessel-suppressed CT images (ClearRead CT, Riverain Technologies) was not reported.

## *Nodule type determination*

### Accuracy

No study was identified that compared the accuracy in nodule type determination between readers with and without software use. Non-comparative evidence is shown in *Appendix 6*, *Table 63*.

### Subquestions 1 to 4: potential factors influencing nodule type determination

No data were available to enable subgroup analyses of nodule type determination accuracy based on contrast use, dose, nodule type, patient ethnicity, radiologist speciality or reason for CT scan in the incidental population.

### Subquestion 5: concordance and variability in nodule type determination

a. Concordance between readers with and without software (no study)

No studies were identified.

b. Concordance between readers using different software (no study)

No studies were identified.

c. Intra-reader and inter-reader variability (two studies)

Two MRMC studies[64,67] were identified that reported on the inter-reader variability in nodule type determination in nodule-enriched screening populations in readers with and without software use. Both studies found that software use did not affect the proportion of disagreements in nodule type between the readers.

*Screening population: Veolity (MeVis) (one study)*

Jacobs *et al.*[64] found that the proportion of Lung-RADS disagreements due to different nodule type between seven readers was 1% (44/3,360 possible reader pairs; 21 readers pairs × 160 cases) when using the dedicated CT lung screening viewer with Veolity software and was also 1% (37/3,360 possible reader pairs) when using the standard PACS viewer.

*Screening population: VUNO Med-Lung CT AI (VUNO) (one study)*

Park *et al.*[67] reported that for all 2,000 possible paired observations among the five readers (10 reader pairs × 200 cases), the proportion of discordant pairs caused by different nodule type were similar between the sessions with (3.6%, 71/2,000) and without (3.4%, 68/2,000) software use ($p = 0.85$).

## *Nodule diameter measurement*

### Accuracy of measurement (three studies)

Three studies compared diameter measurements of stand-alone software[56,66] or readers with concurrent software use[55] with the measurements of a reference standard. The studies were performed in a screening population,[56] a mixed population[66] and a population with unclear indication for the chest CT scan,[55] respectively. Results on the diameter measurement accuracy of stand-alone software were inconsistent, with one study reporting significantly smaller nodule diameters measured by the software[55] and the other study reporting that in 83% of size disagreements the nodule size was overestimated by the software.[66] Substantial agreement with the reference standard was reported for semiautomated longest diameters measured on vessel-suppressed CT images in the third study.[55] Further details of the findings from these three studies are summarised in *Table 6* and the following text.

a. Non-comparative results (three studies)

*Screening population: ClearRead CT (Riverain Technologies) (one study)*

In a nodule-enriched screening population, Singh *et al.* found that for the same risk-dominant, subsolid nodule ($n$ = 100), the average diameter [(maximum dimension of the nodule in mm + orthogonal dimension in mm)/2] estimated by the stand-alone software was significantly smaller [mean 12 mm, standard deviation (SD) 3 mm] than the reference standard measurement obtained by consensus reading of two experienced chest radiologists, with a third experienced radiologist resolving discrepancies (mean 14 mm, SD 5 mm) ($p$ = 0.02).[56]

**TABLE 6** Main findings, risk of bias, applicability concerns and input into modelling

| Study, AI software, country | Population, design and sample | Main findings | Risk of bias, applicability concerns and input into modelling |
|---|---|---|---|
| Singh *et al.* 2021,[56] ClearRead CT, USA | Screening, MRMC, nodule-enriched sample<br>Risk-dominant subsolid nodule<br>$n$ = 100) | Average diameter[a]<br>**Stand-alone AI**: mean 12, SD 3 mm<br>**Radiologist consensus:**[b] mean 14, SD 5 mm; $p$ = 0.02 | **RoB**: research setting; excluded scans that could not be processed by the software ($n$ = 27)<br>**AppC**: research setting; subsolid nodules only<br>**Model**: no; stand-alone AI rather than concurrent AI |
| Martins Jarnalo *et al.* 2021,[66] Veye Chest, the Netherlands | Mixed, retrospective test accuracy study, randomly selected sample, 80 nodules (all nodule types, 4–30 mm) | Diameter measurements<br>**Stand-alone AI vs. unaided radiologist consensus:**[c] agreement (same millimetre): 67.5% (54/80)<br>+ 1 mm: 20.0% (16/80)<br>+ 2 mm: 2.5% (2/80)<br>+ 4 mm: 1.25% (1/80)<br>−1 mm: 2.5% (2/80)<br>−2 mm: 2.5% (2/80)<br>Failure: 3.75% (3/80) | **RoB**: research setting; scans with > 5 nodules were excluded<br>**AppC**: single hospital; stand-alone AI rather than concurrent AI<br>**Model**: yes, through EAG simulation. Randomly selected nodules covering all types; reported breakdown of discrepancies (differing by 1, 2 and 4 mm) between measurements by stand-alone AI and unaided radiologists, which allow measurement accuracy (bias) and precision (variation) of concurrent AI and unaided reading to be derived with some assumption (see *Appendix 8*) |
| Milanese *et al.* 2018,[55] ClearRead CT for vessel suppression; MM Oncology for semiautomatic measurement, Switzerland | Unclear, MRMC, consecutive sample, 65 solid nodules | **Lin's concordance correlation coefficient (CCC) vs. average of semi-automatic measurement on standard CT images:**[d] radiologist 1 on vessel-suppressed CT: 0.967; radiologist 2 on vessel-suppressed CT: 0.960 | **RoB**: research setting; index test readers are part of the reference standard<br>**AppC**: research setting; population characteristics unclear; solid nodules only; radiologists with < 5 years of experience; AI software only used for vessel suppression, not for measurement<br>**Model**: no; Lin's CCC does not allow the derivation of relative measurement accuracy or precision |

AppC, applicability concerns; EAG, External Assessment Group; model, input into modelling; MRMC, multi-reader multi-case study; RoB, risk of bias; SD, standard deviation.

a (maximum dimension of the nodule in mm + orthogonal dimension in mm)/2.

b Reference standard; consensus of two experienced chest radiologist, with a third experienced radiologist resolving discrepancies.

c Consensus reading of one experienced radiologist and one resident radiologist, with discrepancies resolved by a third experienced chest radiologist.

d Compared with reference standard, which was the average semiautomatic measurements by the two readers on standard CT images (without AI for vessel suppression). Radiologists 1 and 2 had 3 years and 1 year of experience, respectively, in chest CT.

*Mixed population: Veye Chest (Aidence) (one study)*

Martins Jarnalo *et al.*[66] compared the diameter measurements of stand-alone software (Veye Chest, Aidence) with a reference standard of consensus reading of one experienced radiologist and one resident radiologist, with discrepancies resolved by a third experienced chest radiologist. In 80 nodules (all nodule types, 4–30 mm), the agreement (same millimetre) between the software measurement and the reference standard was 67.5% (54/80). Of the size discrepancies that were not due to software segmentation failures (23/26), 82.6% (19/23) were measured larger than the reference standard: 16 nodules were measured 1 mm larger, two nodules were measured 2 mm larger, and one nodule was measured 4 mm larger. Four out of 23 (17.4%) nodules were measured smaller than the reference standard: two nodules were measured 1 mm smaller, and two nodules were measured 2 mm smaller. For most of the 1-mm size discrepancies, the reason is not clear. For three nodules (1-, 2- and 4-mm discrepancy) an adjacent artery was also measured by the software. For one nodule with 2-mm discrepancy, the measurement was performed on the wrong section; for one (2-mm discrepancy) a subsolid part of the nodule was not measured; one (1-mm discrepancy) had surrounding spiculae, and another (2-mm discrepancy) was a cavitating nodule.

*Unclear indication for CT scan: ClearRead CT (Riverain Technologies) (one study)*

Milanese *et al.*[55] reported on 65 solid nodules measured independently by one radiologist (3 years of experience in chest CT) and one radiology resident (1 year of experience in chest CT) using the semiautomatic segmentation software 'MM Oncology' (Siemens Healthcare) on vessel-suppressed (ClearRead CT, Riverain Technologies) as well as on standard CT images, with the average of the largest diameters measured on standard CT images by the two readers used as reference standard. To determine the reliability between the performed measurements, Lin's concordance correlation coefficient (CCC) was calculated between each reader's measurement and the reference standard measurement. For semiautomated largest diameters measured on vessel-suppressed CT images, Lin's CCC was 0.967 for reader 1 and 0.960 for reader 2 (Lin's CCC ranges from 0 to ± 1, with a value of 1 meaning perfect concordance).

## Subquestions 1 to 4: potential factors influencing nodule diameter measurement accuracy

No data were available to enable subgroup analyses based on contrast use, dose, nodule type, patient ethnicity, radiologist speciality or reason for CT scan in the incidental population.

## Subquestion 5: concordance between and variability in nodule diameter measurement

a. Concordance between readers with and without software (four studies)

One study[63] evaluated the concordance of nodule diameter measurements between readers with and without software in patients with previously detected subsolid nodules (surveillance population with applicability concerns). Another three studies[33,47,58] reported the concordance of stand-alone software measurements compared with manual diameter measurements in mixed populations.

The studies found similar[58,63] or significantly larger[47] nodule diameters with semiautomatic measurements than with manual measurements. Two studies[47,58] reported a significant correlation between the measurements. One study[33] concluded that the segmentation of pulmonary nodules of stand-alone software and the resulting diameter measurements are comparable to that of manual measurement performed by experienced thoracic radiologists.

*Surveillance population with applicability concerns: Veolity (MeVis) (one study)*

Kim *et al.*[63] included 89 patients with subsolid nodules who underwent preoperative non-enhanced CT and subsequent surgical resection. The diameter of the 102 subsolid nodules was not statistically different between the semiautomated and manual measurements ($p > 0.05$ for both readers; paired $t$-test or Wilcoxon's test, as appropriate). When looking at the diameter measurement of the solid portion only, significant differences were observed between semiautomated and manual measurements for reader 1 (6.3 ± 4.9 mm vs. 5.4 ± 4.5 mm; $p < 0.001$) and the second read of reader 2 (6.5 ± 5.0 mm vs. 5.9 ± 4.5 mm; $p < 0.001$), with semiautomated diameter measurements being larger than manual measurements.

*Mixed population: AI-Rad Companion Chest CT (Siemens Healthineers) (one study)*

Abadia *et al.*[47] found that for the 233 nodules ≥ 4 mm detected by both stand-alone AI-Rad Companion Chest CT (Siemens Healthineers) and the unaided expert radiologist, the software measured the nodule diameter on average 19.7% larger (mean difference 1.7 mm), with these nodules yielding a median size of 8.6 mm [interquartile range (IQR) 6.5–11.5 mm] by AI-Rad and 6.6 mm (IQR 5.0–9.5 mm) by the expert radiologist ($p < 0.0001$). However, the size measurements between the software and the expert radiologist were also significantly correlated ($\rho = 0.821$, $p < 0.0001$).

*Mixed population: Veye Chest (Aidence) (one study)*

The UK-based reader study by Murchison *et al.*[33] included a mixed population of 314 current or ex-smokers and/or those with radiological evidence of emphysema between 55 and 74 years, mimicking a screening population. Two or three independent expert chest radiologists performed manual nodule segmentation using Apple Pencil. The segmentation overlap between each individual reader's segmentation and the software's (Veye Chest, Aidence) segmentation was calculated as the Dice coefficient (a value of 1 means 100% overlap and a value of 0 means 0% overlap) and averaged. For 95% of the 428 nodules between 3 and 30 mm, for which the software was able to create a segmentation, the average Dice coefficient for nodule segmentation between software alone and radiologists was 0.86 (95% CI 0.51 to 0.95). From each segmentation, the largest axial diameter was obtained, and the diameter difference between each individual reader and Veye Chest software was calculated. The geometric mean difference between Veye Chest and the radiologist's measurement was 1.17 mm (95% CI 1.01 to 1.69 mm), which was similar to the geometric mean difference observed between the individual expert radiologists (1.15 mm, 95% CI 1.00 to 1.58 mm).

*Mixed population: ClearRead CT (Riverain Technologies) (one study)*

Wan *et al.*[58] included LDCT images from 50 Taiwanese patients with mixed indications whose nodule(s) were subsequently excised. The study found that in 61 nodules ≤ 2 cm (13 solid, 20 part-solid, 28 ground-glass nodules) detected and measured by the software ClearRead CT (Riverain Technologies), there was no significant difference in diameters measured manually by two experienced radiologists in consensus or by the stand-alone software (mean 7.83, SD 3.06 mm, vs. mean 8.13, SD 3.49 mm; $p = 0.624$) with a Pearson's correlation coefficient of 0.926.

b. Concordance between readers using different software or between different software without human involvement (no study)

No study was identified that reported on the concordance between readers using different AI-based software or between different AI-based software without human involvement for nodule diameter measurements.

c. Intra-observer and inter-observer variability (five studies)

### *Inter-observer variability (five studies)*

Five MRMC studies[33,62–64,67] were identified that reported on the inter-observer variability in nodule diameter measurements. Three of them[63,64,67] compared the inter-reader variability between manual diameter measurements and semiautomatic measurements and consistently found reduced disagreements in nodule sizes between readers with software use. The variability in readers using semi-automatic software was similar in CT images reconstructed with FBP and images reconstructed with MBIR algorithms.[62]

*Screening population: Veolity (MeVis) (one study)*

The study by Jacobs *et al.*[64] included a nodule-enriched screening population. All seven observers read all 160 CT images twice: once in the dedicated CT lung screening viewer including the software Veolity (MeVis) and once in the standard viewer without software support. The study found 67% (207 vs. 68) fewer Lung-RADS category disagreement pairs due to different nodule diameter measurements when using the dedicated CT lung screening viewer with Veolity software.

*Screening population: VUNO Med-Lung CT AI (VUNO) (one study)*

Park *et al.*[67] included a nodule- and cancer-enriched screening population (200 baseline LDCT images) selected from the US-based NLST data set. Five readers with varying levels of experience assessed the LDCT images with and without concurrent software use (VUNO Med-Lung CT AI). With software use, the proportion of disagreements in Lung-RADS category due to different nodule size measurements was reduced from 5.1% (102/2,000) to 3.1% (62/2,000) for all 2,000 possible paired observations among the five readers ($p < 0.001$).

*Surveillance population with applicability concerns: Veolity (MeVis) (two studies)*

Two studies were performed at the same hospital in the Republic of Korea and included (potentially overlapping) surveillance populations with applicability concerns: 89[63] and 73 patients,[62] respectively, with preoperative CT scans for subsolid nodules. In both MRMC studies, two radiologists with concurrent use of the software Veolity (MeVis) independently performed nodule diameter measurements, but only one study[63] compared semiautomatic with manual diameter measurements.

Kim *et al.*[63] found that in 102 subsolid nodules measured by semiautomated segmentation software, the inter-reader variability of two experienced radiologists ranged from −1.9 mm (95% CI −2.3 to −1.6 mm) to 2.1 mm (95% CI 1.7 to 2.4 mm) for the whole nodule diameter and from −2.1 mm (95% CI −2.5 to −1.8 mm) to 2.1 mm (95% CI 1.7 to 2.5 mm) for the solid portion diameter. With manual measurement, inter-reader variability ranged from −2.8 mm (95% CI −3.3 to −2.4 mm) to 2.4 mm (95% CI 2.0 to 2.9 mm) for the whole nodule diameter and from −5.1 mm (95% CI −5.7 to −4.4 mm) to 2.8 mm (95% CI 2.1 to 3.5 mm) for the solid portion diameter. The inter-reader variability of semiautomatic measurement was significantly lower than that of manual measurement for both whole nodules and solid portion diameters ($p < 0.001$ for all).

Cohen *et al.*[62] compared semiautomatic measurement using CT images reconstructed with FPB and MBIR algorithm. This study did not include a 'manual measurement' comparator. Regarding the semi-automatic measurement of the longest diameter of the whole subsolid nodule ($n = 66$), the absolute and relative mean differences between the two readers were 0.48 mm and 3.3%, respectively, with FBP reconstruction algorithm, and 0.24 mm and 2%, respectively, with MBIR algorithm. For the diameter of the solid component of the subsolid nodules, the absolute and relative mean differences between the two readers were 0.01 mm and 6.4%, respectively, with FBP, and −0.31 mm and −3%, respectively, with MBIR. There were no significant differences in inter-reader variability between FBP and MBIR reconstructed CT images ($p > 0.05$).

*Mixed population: manual measurement (one study)*

The UK-based reader study by Murchison *et al.*[33] included a mixed population of 314 current or ex-smokers and/or those with radiological evidence of emphysema between 55 and 74 years, mimicking a screening population. The study reported the inter-reader variability between the unaided readers only. Two or three independent expert chest radiologists performed manual nodule segmentation using Apple Pencil. In 428 nodules between 3 and 30 mm, the average inter-reader Dice coefficient for nodule segmentation was 0.83 (95% CI 0.39 to 0.96), and the geometric mean diameter difference of the largest axial diameter was 1.15 mm (95% CI 1.00 to 1.58 mm).

### *Reproducibility/repeatability (two studies)*

Two studies[62,63] reported the intra-reader variability in nodule diameter measurements in patients with previously detected subsolid nodules. The intra-reader variability with semiautomatic measurement was significantly lower than that with manual measurement for the whole nodule diameter and the solid portion diameter, respectively,[63] and was similar in FBP and MBIR reconstructed CT images.[62]

*Surveillance population with applicability concerns: Veolity (MeVis) (two studies)*

Both the MRMC studies were performed at the same hospital in the Republic of Korea and comprised (potentially overlapping) surveillance populations with applicability concerns: 89[63] and 73 patients,[62] respectively, with preoperative CT scans for subsolid nodules.

In the study by Kim *et al.*,[63] one experienced radiologist performed the nodule diameter measurements twice with concurrent use of the software Veolity (MeVis) and twice without software use in 102 subsolid nodules. With semiautomatic measurement, the mean percentage relative difference between the two repeated measurements was 2.3 ± 4.9% for the whole nodule diameter and 8.9 ± 34.2% for the solid portion diameter. With manual measurement, the mean percentage relative difference was 7.0 ± 6.6% for the whole nodule diameter and 17.4 ± 34.3% for the solid portion. The intra-reader variability of semiautomatic measurement was significantly lower than those of manual measurement for the whole nodule diameter and the solid portion diameter, respectively ($p < 0.001$ for all).

In the study by Cohen *et al.*,[62] two radiologists with four and five years of experience performed the semiautomatic measurements with concurrent use of the software Veolity (MeVis), twice on FBP reconstructed CT images and twice on FBIR reconstructed CT images. In 66 subsolid nodules, the mean relative difference was −0.59% using FBP and 0.03% using MBIR for the longest diameter of the whole nodule ($p = 0.41$). The mean relative difference of the longest diameter of the solid portion was −0.17% for FBP and −4.12% for MBIR ($p = 0.08$). Intra-observer variability was similar ($p > 0.05$) for FBP and MBIR reconstructed CT images.

### *Nodule volume measurement*

#### Accuracy in nodule volume measurement (one study)

One MRMC study[55] reported on the accuracy of volume measurement in solid nodules and found that semiautomated volumetric measurements in vessel-suppressed CT images agreed substantially with the reference standard. The percentages of error of semiautomated volumetric measurement were similar in standard CT images and vessel-suppressed CT images.

a. Comparative results: reader with and without software (one study).

*Unclear indication for chest CT scan: ClearRead CT (Riverain Technologies) (one study)*

This MRMC study[55] included 93 consecutive patients referred for clinical non-enhanced, LDCT (unclear indication for the chest CT scan). One radiologist with three years of experience in chest CT and a radiology resident independently performed semiautomatic volume measurements of 65 solid nodules using the software 'MM Oncology' by Siemens Healthcare on vessel-suppressed (ClearRead CT, Riverain Technologies) as well as on standard CT images. After the independent reading by the two readers, the volumes measured on standard CT images by reader 1 and reader 2 for each nodule were averaged, and the resulting values acted as the reference standard. The agreement of semiautomatic volumetric measurement with the reference standard was assessed using Lin's CCC (value of 1 meaning perfect concordance and 0 meaning no concordance). Overall, Lin's CCC was 0.990 for reader 1's volume measurements and 0.985 for reader 2's volume measurements. For central nodules, Lin's CCC was 0.992 for both readers. For peripheral nodules, Lin's CCC was 0.959 for reader 1 and 0.956 for reader 2, and for subpleural/perifissural nodules, Lin's CCC was 0.981 and 0.960 for reader 1 and reader 2, respectively. Regarding nodules adjacent to a vessel, Lin's CCC was 0.992 for reader 1 and 0.990 for reader 2 on vessel-suppressed CT images and 0.990 for reader 1 and 0.992 for reader 2 on standard CT images. The percentages of error for the volumetric measurements compared with the reference standard were not statistically different between standard CT images and vessel-suppressed CT images ($p > 0.05$ for every pair of data sets). On standard CT images, the percentage error was 3.7% for reader 1 and −2.7% for reader 2, whereas on vessel-suppressed CT images, the percentage volume error was −1.4% for reader 1 and −6.4% for reader 2. Milanese *et al.*[55] concluded that vessel-suppressed CT data sets can be used for semiautomated measurements of solid pulmonary nodules.

#### Subquestions 1 to 4: potential factors influencing nodule volume measurement

No data were available to enable subgroup analyses based on contrast use, dose, nodule type, patient ethnicity, radiologist speciality or reason for CT scan in the incidental population.

#### Subquestion 5: concordance and variability in nodule volume measurement

a. Concordance between readers with and without software (one study).

No study was identified that reported on the concordance of volume measurements between readers with and readers without software. However, one study[33] evaluated the concordance of volume measurements between stand-alone software and unaided readers. The study concluded that the performance of the software for segmenting pulmonary nodules on chest CT is comparable to that of experienced thoracic radiologists.

*Mixed population: Veye Chest (Aidence) (one study)*

The UK-based reader study by Murchison *et al.*[33] comprised a mixed population of 314 current or ex-smokers and/or those with radiological evidence of emphysema between 55 and 74 years, mimicking a screening population. Nodules were manually segmented (using Apple Pencil) by two or three experienced thoracic radiologists. Software segmentation was successful in 95% of 428 nodules of all types between 3 and 30 mm. The average Dice coefficient between Veye Chest's and each individual radiologist's segmentation was 0.86 (95% CI 0.51 to 0.95). For the volumes derived from the segmentation, the geometric mean volumetric difference between the software and each individual radiologist was 1.38 mm$^3$ (95% CI 1.01 to 3.38 mm$^3$), which was similar to the volume difference observed between the expert radiologists (1.39 mm$^3$, 95% CI 1.01 to 3.19 mm$^3$).

b. Concordance between readers using different software or between different software without human involvement (no study)

No study was identified that reported on the concordance of volume measurements between readers using different AI-based software or between different software without human involvement.

c. Intra-reader and inter-reader variability (three studies)

### *Inter-observer variability (three studies)*

Three MRMC studies[33,55,62] reporting on the inter-observer variability in nodule volume measurement were identified. Between-reader agreement using semiautomatic software was almost perfect on both standard CT images and vessel-suppressed CT images.[55] The inter-reader variability of semiautomatic volumetric measurement was similar in FBP and MBIR reconstructed CT images.[62] The third study only reported inter-observer agreement between unaided readers.[33]

*Surveillance population with applicability concerns: Veolity (MeVis) (one study)*

The study by Cohen *et al.*[62] included a surveillance population with applicability concerns: 73 patients with preoperative CT scans for subsolid nodules. Two radiologists with four and five years of experience independently performed the semiautomatic measurements with concurrent use of the software Veolity (MeVis) on FBP reconstructed CT images as well as on MBIR reconstructed CT images. In 66 subsolid nodules, the mean absolute (relative) differences between the two readers for the whole nodule volume was 199.8 mm$^3$ (9.6%) with FBP and 92.6 mm$^3$ (5.5%) with MBIR ($p$ = 0.13). The mean absolute (relative) volume differences between the two readers for the solid portion were −4.9 mm$^3$ (1.6%) with FBP and −21.4 mm$^3$ (−12.7%) with MBIR ($p$ = 0.11).

*Mixed population: unaided readers (one study)*

The UK-based reader study by Murchison *et al.*[33] comprised a mixed population of 314 current or ex-smokers and/or those with radiological evidence of emphysema between 55 and 74 years, mimicking a screening population. Nodules were manually segmented (using Apple Pencil) by two or three experienced thoracic radiologists. In 428 nodules between 3 and 30 mm, the average Dice coefficient between each reader's segmentation and the segmentation from the other readers was 0.83 (95% CI 0.39 to 0.96). The geometric mean volumetric discrepancy between radiologists was 1.39 mm$^3$ (95% CI 1.01 to 3.19 mm$^3$).

*Unclear indication for CT scan: ClearRead CT (Riverain Technologies) (one study)*

This MRMC study by Milanese *et al.*[55] comprised 93 consecutive patients referred for clinical non-enhanced, chest LDCT (unclear indication). One radiologist with three years of experience in chest CT and a radiology resident independently performed semiautomatic volume measurements of 65 solid nodules using the software 'MM Oncology' by Siemens Healthcare on vessel-suppressed (ClearRead CT, Riverain Technologies) as well as on standard CT images. Between-readers agreement was assessed using Lin's CCC and found to be 0.994 on both standard CT images and vessel-suppressed CT images (Lin's CCC of 1 meaning perfect concordance and 0 meaning no concordance). On standard CT images, the two readers measured identical volumes in eight cases (12.3%). On vessel-suppressed CT images, reader 1 and reader 2 measured identical volumes in 11 cases (16.9%). The upper and lower limits of agreement between reader 1 and reader 2 were 15.5 mm$^3$ and −21.4 mm$^3$, respectively, on vessel-suppressed CT images and 16.3 mm$^3$ and −22.4 mm$^3$, respectively, on standard CT images.

#### *Repeatability/reproducibility (one study)*

One study[62] reported on the reproducibility of semiautomatic volume measurements and found similar intra-reader variability in FBP and MBIR reconstructed CT images.

*Surveillance population with applicability concerns: Veolity (MeVis) (one study)*

Cohen *et al.*[62] included 73 patients with preoperative CT scans for subsolid nodules from a single hospital in the Republic of Korea. Two radiologists performed the semiautomatic measurements with concurrent use of the software Veolity (MeVis), twice on FBP reconstructed CT images and twice on MBIR reconstructed CT images. In 66 subsolid nodules, the mean relative difference in the whole nodule volume was −1.23% using FBP and 0.28% using MBIR ($p$ = 0.16). For the volume of the solid portion, the mean relative difference was 4.74% with FBP and −5.9% with MBIR ($p$ = 0.07). Intra-observer variability was similar ($p$ > 0.05) in FBP and MBIR reconstructed CT images.

## *Classification into risk categories based on nodule type and size*

### Accuracy for risk classification based on 2015 British Thoracic Society guidelines (one study)

One study[34] reported on the performance of readers with and without concurrent software use for identifying patients classed as BTS grade A (discharge recommended) on consensus. This study also reported on the agreement in nodule management recommendations (four grades based on the 2015 BTS guidelines[12]) between single readers (with/without software use) and the consensus read. It was performed in patients with incidentally detected nodules with and without prior CT imaging. Sensitivities and specificities for identifying patients that can be discharged were higher in software-aided readers than in unaided readers, but 95% CIs overlapped. Regarding all four possible nodule management recommendation categories, the aided readings of each radiologist showed a higher agreement with the consensus session than when readings were unaided, but no level of significancy or 95% CIs were reported.

a. Comparative results: reader with and without software (one study)

*Mixed population: Veye Chest (Aidence) (one study)*

Hempel *et al.*[34] selected 50 chest CT scans with incidentally detected nodules (35 with and 10 without prior imaging) or no nodules ($n$ = 5) from one hospital in the Netherlands. For this MRMC study, two experienced radiologists independently assessed the CT images to determine the nodule management recommendation grade based on the 2015 BTS guidelines[12] (A, discharge; B, CT at 3 months; C, Brock score; D, diagnostic work-up) twice, first unaided and then aided by Veye Chest software (Aidence). After both reading sessions had been completed, the consensus BTS grade of the two readers was used as the reference standard. With concurrent use of Veye Chest software, the sensitivities and specificities to identifying patients with BTS grade A (no clinical follow-up required) were higher in both readers, but 95% CIs overlapped (see *Appendix 1*, *Table 46*).

The software-aided readings of reader 1 and reader 2 also showed a higher agreement in nodule management recommendation grades with the consensus session (linear weighted kappa, 0.80 and 0.87, respectively) than the unaided readings (0.66 and 0.57, respectively), but no level of significance or 95% CI was reported.

### Accuracy for risk classification based on other risk categories (two studies)

Two studies were identified that evaluated the accuracy of stand-alone software[32] and software-assisted readers[32,55] in classifying solid nodules into other risk categories based on volume.

One study was performed in a selected screening population,[32] and in the other study the indication for the chest CT scan was not reported.[55]

'Excellent agreement' with the reference standard (which was based on the average volume measurement on standard CT images of the two index test readers) was reported for readers performing semiautomatic volumetric measurements in vessel-suppressed CT images in one study using three volume-based risk categories.[55] Using two volume-based risk categories, another study found misclassifications by stand-alone software in 22% and by software-assisted readers in 10–15% of cases.[32]

a. Comparative results: reader with and without software (one study)

*Screening population: AVIEW LCS (Coreline Soft) (one study)*

Lancaster *et al.*[32] included 283 participants who underwent a baseline ultra-LDCT thorax scan and had at least one solid nodule of any size. In a MRMC study, five thoracic radiologists with > 7 years of experience independently interpreted the CT images with visual nodule detection and software use for semiautomated volume measurement (readers 1–3, AVIEW LCS from Coreline Soft; reader 4, AGFA Enterprise 8.0 Imaging software; reader 5, Syngo.via MM Oncology VB20) and classified nodules based on the NELSON-plus/EUPS protocol volume threshold of 100 mm$^3$. The performance of stand-alone software (AVIEW LCS from Coreline Soft) in automatically detecting, measuring and classifying solid nodules was also evaluated. As reference standard, an independent consensus read of the 283 largest nodules was performed by a panel of three radiologists with > 10 years' experience and an experienced information technologist. Compared with the reference standard, the stand-alone software had 61 (21.6%; false positive, *n* = 53; false negative, *n* = 8) misclassifications reported, compared with 43 discrepancies (15.1%; false positive, *n* = 22; false negative, *n* = 21) for reader 1, 36 (12.7%; false positive, *n* = 25; false negative, *n* = 11) for reader 2, 29 (10.2%; false positive, *n* = 25; false negative, *n* = 4) for reader 3, 28 (9.9%; false positive, *n* = 6; false negative, *n* = 22) for reader 4 and 50 (17.7%; false positive, *n* = 15; false negative, *n* =35) discrepancies for reader 5.

b. Non-comparative results (one study)

*Unclear indication for CT scan: ClearRead CT (Riverain Technologies) (one study)*

The MRMC study by Milanese *et al.*[55] comprised 93 consecutive patients referred for clinical non-enhanced, low-dose chest CT (unclear indication). One radiologist with three years of experience in chest CT and a radiology resident independently performed semiautomatic volume measurements of 65 solid nodules using the software 'MM Oncology' by Siemens Healthcare on vessel-suppressed (ClearRead CT, Riverain Technologies) as well as on standard CT images. They categorised nodules according to Fleischner Society Guidelines into < 100 mm$^3$, 100–250 mm$^3$ and > 250 mm$^3$.[68] After the independent reading was performed by the two readers, volumes measured on standard CT images by reader 1 and reader 2 for each nodule were averaged and the resulting values acted as the reference standard. The agreement between the Fleischner management categories[68] based on semiautomated volumetric measurements performed on vessel-suppressed CT images and the reference standard was reported as 'excellent' (see *Appendix 1*, *Table 47*).

### Subquestions 1 to 4: potential factors influencing risk classification

No data were available to enable subgroup analysis based on contrast use, dose, nodule type, patient ethnicity, radiologist speciality or reason for CT scan in the incidental population.

### Subquestion 5: concordance and variability in risk classification

a. Concordance between readers with and without software use (two studies)

One study[64] was identified that reported on the concordance in Lung-RADS categorisation between readers with and readers without software use. A second study[67] reported on the concordance in Lung-RADS categorisation between stand-alone software and readers with and readers without software use. Both studies were performed in nodule-enriched screening populations. The agreement in Lung-RADS categorisation between each reader with and without software as assessed by mean Cohen's weighted kappa value was 0.67.[64] The agreement between stand-alone software and each reader increased with software use.[67]

*Screening population: Veolity (MeVis) (one study)*

In the study by Jacobs *et al.*,[64] seven observers read all 160 CT images twice: once in the dedicated CT lung screening viewer including Veolity Lung CAD (MeVis) and once in the standard viewer without software support. The intra-observer agreement in Lung-RADS categorisation for each reader with and without software use was assessed using mean Cohen's weighted *k* value and constituted 0.67 (range 0.59–0.76 for individual readers).

*Screening population: VUNO Med-Lung CT AI (VUNO) (one study)*

Park *et al.*[67] investigated the agreement in nodule Lung-RADS categorisation of 200 LDCT images between stand-alone software and five readers with and without software use (VUNO Med-Lung CT AI from VUNO). Agreement in Lungs-RADS categorisation between stand-alone software and each unaided reader was assessed using Cohen's kappa, ranging from 0.45 (95% CI 0.34 to 0.57) to 0.57 (95% CI 0.46 to 0.67). Overall, the agreement in Lung-RADS categorisations between stand-alone software and each reader increased with software use, with Cohen's kappa ranging from 0.58 (95% CI 0.48 to 0.68) to 0.70 (95% CI 0.62 to 0.78).

b. Concordance between readers using different software or between different software without human involvement (no study)

No study was identified that reported on the concordance between readers using different AI-based software or between different AI-based software without human involvement for risk categorisation based on nodule type and size.

c. Intra-reader and inter-reader variability (five studies)

### *Inter-reader variability (five studies)*

**Categorisation based on 2015 BTS guidelines (one study)** One study[34] reported on the inter-reader agreement in nodule management recommendations based on the 2015 BTS guidelines.[12] It was performed in patients with incidentally detected nodules with and without prior CT imaging and found higher inter-reader agreement with concurrent software use, but no level of significance or 95% CIs were reported.

*Mixed population: Veye Chest (Aidence) (one study)*

Hempel *et al.*[34] selected 50 chest CT scans with incidentally detected nodules (35 with and 10 without prior imaging) or no nodules (*n* = 5) from one hospital in the Netherlands. For this MRMC study, two experienced radiologists independently assessed the CT images to determine the nodule management recommendation grade based on the 2015 BTS guidelines[12] (A, discharge; B, CT at 3 months; C, Brock score; D, diagnostic work-up) twice, first unaided and then aided by Veye Chest software (Aidence). The inter-reader agreement in nodule management recommendation grades was higher in readers with concurrent software use (linear weighted kappa 0.84) than in unaided readers (linear weighted kappa 0.61), but no level of significance or 95% CIs were reported.

**Categorisation based on Lung-RADS categories (two studies)** Two studies[64,67] were identified that reported on the inter-reader variability in nodule Lung-RADS categorisation. Both studies were performed in nodule-enriched screening populations and found marginally improved[67] and improved[64] inter-reader agreement with software use.

*Screening population: Veolity (MeVis) (one study)*

In the study by Jacobs *et al.*,[64] seven observers read all 160 CT images twice: once in the dedicated CT lung screening viewer including the software Veolity (MeVis) and once in the standard viewer without software support. When using the standard PACS-like viewer without software support, the inter-reader agreement in Lungs-RADS categorisation had a Fleiss' kappa value of 0.58 (95% CI 0.55 to 0.60). When readers used the dedicated CT lung screening viewer with Veolity software, the Fleiss' kappa value increased to 0.66 (95% CI 0.64 to 0.68). The mean pairwise Cohen's weighted kappa values of each reader with the remaining six readers ranged from 0.63 to 0.73 without software use and from 0.61 to 0.74 with software use. Disagreements regarding Lung-RADS categories occurred in 29% (971/3,360) of unaided readings and in 25% (853/3,360) of readings when using the dedicated CT lung screening viewer with integrated Veolity software, but no level of significance or 95% CIs were reported. The study found 12% (118/971) fewer disagreements between observer pairs when using the dedicated CT lung screening viewer than when using the standard PACS-like viewer.

*Screening population: VUNO Med-Lung CT AI (VUNO) (one study)*

In the study by Park *et al.*,[67] five readers assessed the 200 LDCT images with and without software use (VUNO Med-Lung CT AI from VUNO). Inter-reader agreement of five readers for Lung-RADS categorisation as assessed by Fleiss' kappa was 0.60 (95% CI 0.57 to 0.63) without software use, and improved marginally to 0.65 (95% CI 0.63 to 0.68) with software use. The pairwise agreement between unaided readers found an average Cohen's kappa of 0.71 (range 0.59–0.78). Disagreements in Lung-RADS category among the 2,000 possible reading pairs between the five readers were observed in 18.6% (371/2,000). With software use, the pairwise agreement between readers was slightly higher than in unaided readers, with an average Cohen's kappa of 0.75 (range 0.68–0.79). Disagreements in Lung-RADS category were observed in 18.3% (365/2,000) of all possible reading pairs.

**Categorisation into other risk categories (two studies)** Two studies were identified that reported on the inter-reader variability in categorising subsolid nodules in accordance with Fleischner Society guidelines[70] into (1) pure ground glass, (2) part-solid with a solid component ≥ 5 mm or (3) part-solid with a solid component < 5 mm.[62,63] Semiautomatic segmentation significantly improved inter-reader variability compared with manual measurement ($p = 0.022$), especially the subclassification of part-solid nodules according to the diameter of the solid portion.[63] The inter-observer agreement in semiautomated measurements performed on FBP and MBIR reconstructed CT images was not statistically different ($p = 0.22$).[62]

*Surveillance populations with applicability concerns: Veolity (MeVis) (two studies)*

Both studies were performed at the same hospital in the Republic of Korea and comprised (potentially overlapping) surveillance populations with applicability concerns: 89[63] and 73 patients,[62] respectively, with preoperative CT scans for subsolid nodules. In both reader studies, two radiologists with concurrent use of the software Veolity (MeVis) independently performed nodule measurements and nodule classification into the three categories. In the study by Kim *et al.*,[63] the two readers also assessed CT images without software use performing manual diameter measurement.

In the study by Kim *et al.*,[63] the inter-reader variability (kappa) regarding the classification of 102 subsolid nodules was 0.861 (95% CI 0.769 to 0.953) for semiautomatic measurement and 0.683 (95% CI 0.561 to 0.805) for manual measurement ($p = 0.022$). Percentage inter-reader agreement was 92.2% (94/102) for semiautomatic measurement and 80.4% (82/102) for manual measurement.

Cohen *et al.*[62] found that the inter-observer variability in categorising 66 subsolid nodules as assessed by kappa values was 0.66 and 0.77 for FBP and MBIR, respectively. The inter-observer agreement for both image reconstruction algorithms was not statistically different ($p = 0.22$).

### *Repeatability/reproducibility (two studies)*

Two studies were identified that reported on the intra-reader reproducibility in categorising subsolid nodules according to Fleischner Society guidelines[70] into (1) pure ground glass, (2) part-solid with a solid component ≥ 5 mm or (3) part-solid with a solid component < 5 mm.[62,63] One study reported significantly higher intra-reader reproducibility with semiautomatic measurement than with manual measurement.[63] Readers with semiautomatic measurement had significantly higher intra-reader agreement with MBIR than with FPB reconstructed images.[62]

*Surveillance populations with applicability concerns: Veolity (MeVis) (two studies)*

Both studies were performed at the same hospital in the Republic of Korea and comprised (potentially overlapping) surveillance populations with applicability concerns: 89[63] and 73 patients,[62] respectively, with preoperative CT scans for subsolid nodules.

In the reader study by Kim *et al.*,[63] one experienced radiologist performed the nodule diameter measurements twice with concurrent use of the software Veolity (MeVis), and twice without software use in 102 subsolid nodules. The intra-reader reproducibility (kappa) of nodule classification was 0.894 (95% CI 0.812 to 0.976) for semiautomatic measurement and 0.750 (95% CI 0.632 to 0.868) for manual measurement ($p = 0.049$). The percentage intra-reader agreement was 94.1% (96/102) for semi-automatic measurement and 85.3% (87/102) for manual measurement.

In the reader study by Cohen *et al.*,[62] two radiologists with four and five years of experience performed semiautomatic measurements with concurrent use of the software Veolity (MeVis), twice on FBP reconstructed CT images and twice on MBIR reconstructed CT images. The intra-observer reproducibility (kappa) for the classification of the 66 subsolid nodules was 0.83 and 0.94 for FBP and MBIR, respectively. The intra-reader agreement was significantly higher when using the MBIR algorithm ($p = 0.04$).

### *Whole read (detection plus risk categorisation based on nodule type and size)*

#### Accuracy for lung cancer detection based on whole read (two studies)

Two studies were identified that reported the accuracy for lung cancer detection of a whole read (nodule detection and classification based on nodule type and size) performed by single experienced thoracic radiologists with[50,51] or without[51] concurrent software use (AVIEW, Lungscreen, Coreline Soft) in a prospective screening population from the Republic of Korea. Positivity was based on Lung-RADS category ≥ 3, and the reference standard was medical record review. The comparative study did not find a statistical difference in sensitivity, specificity, positive predictive value and negative predictive value before and after software implementation, when measurements were performed on transverse planes. After software implementation, positive predictive values differed significantly according to measurement planes used (transverse, maximum orthogonal, any maximum).

a. Comparative results: reader with and without software (one study)

*Screening population: AVIEW Lungscreen (one study)*

In a before-and-after study, Hwang *et al.*[51] included 6,487 consecutive participants of the K-LUCAS project: 1,821 participants were screened before the software was implemented (including 16 cases of lung cancer) and 4,666 participants received screening after the software was implemented (including 31 cases of lung cancer). Based on transverse plane diameter measurements, the Lung-RADS-based (version 1.0) sensitivity was 93.8% before the AVIEW Lungscreen software was implemented and 93.5% after software implementation ($p = 0.979$). The specificity was 90.9% before and 89.6% after the software was implemented ($p = 0.132$). There were also no significant differences in positive and negative predictive values ($p > 0.05$ for all). With software use, the specificity (89.6% on transverse planes, 86.5% on maximum orthogonal planes, 83.1% on any maximum planes) and positive predictive values (5.7% on transverse planes, 4.6% on maximum orthogonal planes, 3.7% on any maximum planes) of Lung-RADS differed significantly according to the measurement planes used ($p < 0.001$ for all).

Non-comparative results (one study) are reported in *Appendix 6*.

#### Subquestions 1 to 4: potential factors influencing accuracy for lung cancer detection based on whole read

No data were available to enable subgroup analyses based on contrast use, radiation dose, nodule type, patient ethnicity, radiologist speciality or reasons for CT scan (incidental population).

#### Subquestion 5: concordance and variability for whole read

No evidence was identified for subquestions 5(a)–(c).

## Use case 2: nodule growth monitoring in people with previously identified lung nodules

### *Detection of growing nodules (no study)*

No study was identified that evaluated the accuracy of AI-based software for detecting growing nodules based on VDT at thresholds according to BTS guidelines[12] or other thresholds.

### *Nodule registration and growth assessment*

#### Accuracy of nodule registration (one study)

No study was identified that compared the accuracy of nodule registration between readers with and without AI software use. However, Murchison *et al.*[33] evaluated the accuracy of stand-alone AI (Veye Chest, Aidence) to detect nodule pairs in subsequent scans of the same patient. The study found a sensitivity for detecting nodule pairs of 100.0% (23/23), with no false-positive pairs (see *Appendix 6*).

#### Subquestions 1 to 4: potential factors influencing accuracy of nodule registration or growth rate estimation

No data were available to enable subgroup analyses based on contrast use, radiation dose, nodule type, patient ethnicity, radiologist speciality or reasons for CT scan (incidental population).

#### Subquestion 5: concordance and variability for nodule registration or growth assessment

a. Concordance between readers with and without AI software use (one study)

No study was identified that reported on the concordance of readers with and without AI software use. However, the same study mentioned above[33] reported on the mean growth percentage difference between stand-alone AI (Veye Chest, Aidence) and unaided expert radiologists.[33] The geometric mean growth rate difference was similar in stand-alone AI and unaided readers. However, due to a single incorrect segmentation of the stand-alone AI, the upper end of its CI is twice as high as that of readers, illustrating that visual verification of the nodule segmentation by human readers is still advised.

b. Concordance between readers using different software or between different software without human involvement (no study)

No study was identified that reported on the concordance in growth rate between readers using different AI-based software or between different AI-based software without human involvement.

c. Intra-reader and inter-reader variability (one study)

One study was identified that reported on the inter-reader variability in nodule growth assessment between unaided readers.[33] The mean growth rate difference for 23 nodule pairs between two unaided expert radiologists was 1.30%.

*Mixed population: unaided readers (one study)*

Murchison *et al.* included a routine cohort of current or ex-smokers and/or those with radiological evidence of pulmonary emphysema between 55 and 74 years (to mimic a screening population) who underwent chest CT for non-screening purposes at a single centre in Edinburgh (UK).[33] Forty-six CT scans from 23 patients undergoing CT surveillance of a pulmonary nodules (23 baseline CT scans and 23 follow-up CT scans) were included in the analysis of nodule registration and growth rate assessment. The mean growth rate difference for 23 nodule pairs between two unaided expert radiologists was 1.30% (95% CI 1.02 to 2.21).

## Practical implications

### *Technical failure rate (12 studies)*

Twelve records were identified that reported on the technical failure rate of AI-based software assessing chest CT images.[27,31,33,34,50–52,56,62–64,66] Six studies were performed in a screening population,[27,50–52,56,64] two studies were performed in a surveillance population with applicability concerns,[62,63] and the remaining four studies comprised mixed populations.[31,33,34,66] The identified studies used five different technologies: Veye Chest (Aidence) as stand-alone software[33,66] or in concurrent mode,[34,62,63] Veolity (MeVis) in concurrent mode,[27,62–64] ClearRead CT (Riverain Technologies) as stand-alone software,[56] AVIEW Lungscreen (Coreline Soft) in concurrent mode,[50–52] and contextflow SEARCH Lung CT (contextflow) in concurrent mode.[31] Segmentation failure ranged from 0% to 57% of nodules (eight studies; *Table 7*). However, one study discussed that the observed nodule segmentation failure was mostly due to the rejection of segmentation results by radiologists, rather than the inability of the system to segment the nodule. Failure rates seem to be higher in pure ground-glass nodules (34%) and part-solid nodules (19.7%) than in solid nodules (7%) (one study). Manual modifications of the segmentation were required in 29–59% of nodules (two studies).

*Screening population: Veolity (MeVis) (two studies)*

The MRMC study by Jacobs *et al.*[64] comprised a nodule-enriched screening population. Seven observers read all 160 CT images twice: once in the dedicated CT lung screening viewer including Veolity Lung CAD (MeVis) and once in the standard viewer without software support. The study found that a satisfactory nodule segmentation was achieved for almost all nodules shown in the dedicated CT lung screening viewer. In 28% of nodule segmentations, the readers manually tuned the segmentation parameters. Manual diameter measurement was deemed necessary for 1.9% (3/160; one observer) or 1.3% (2/160; two observers) nodules.

The study by Hall *et al.*[27] was performed in London (UK) and is a substudy of the LSUT trial. In a MRMC study, two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCT images with concurrent software use (Veolity, MeVis). Issues with the nodule detection software (no interpretation, processing failure) were reported by reader 1 in 9/770 (1.2%) and by reader 2 in 18/770 (2.3%) cases.

*Screening population: ClearRead CT (Riverain Technologies) (one study)*

Singh *et al.*[56] included a nodule-enriched screening population. Using ClearRead CT from Riverain Technologies in stand-alone and concurrent mode, 27 out of 150 (18%) chest CT examinations could not be processed with the AI algorithm as they had artifacts, thicker sections and/or missing images in the downloaded data sets.

*Screening population: AVIEW Lungscreen (Coreline Soft) (three studies)*

All three identified studies by Hwang *et al.*[50–52] are based on the K-LUCAS project and possibly have overlapping patients and CT images. K-LUCAS is a prospective pilot programme of lung cancer screening in the Republic of Korea involving 14 institutions. The software AVIEW Lungscreen from Coreline Soft was used in concurrent mode by experienced thoracic radiologists to detect, measure and classify their Lung-RADS category in clinical practice.

The first included analysis from the K-LUCAS project comprises 4666 CT images taken between April 2017 and March 2018 containing 4990 lung nodules. Semiautomated segmentation failed in 13.4% (669/4990) of nodules.[51]

A second analysis[50] included 10,424 CT images taken between April 2017 and December 2018 with a total of 10,080 nodules identified. Ninety-one per cent of nodules (9206/10,080) were measured by semiautomated segmentation, while 9% (874/10,080) of nodules failed to be semiautomatically segmented and were measured manually. Segmentation failures occurred in 7.3% (688/9465) of solid nodules, 19.7% (31/157) of part-solid nodules and 33.8% (155/458) of ground-glass nodules.

A third analysis[52] of the K-LUCAS project including 3,353 CT images conducted between April 2017 and December 2017 evaluated the inter-institutional and inter-radiologist variability in the frequency of segmentation failure

**TABLE 7** Technical failure rate of AI-based software for lung nodule detection and analysis, by target population and technology (12 studies)

| Reference and country | Population/nodule characteristics/slice thickness | Technology | Details of technical failure | Failure rate |
|---|---|---|---|---|
| ***Screening population (six studies)*** | | | | |
| Hwang *et al.* 2021,[51] Republic of Korea | K-LUCAS (Korea)<br>4666 LDCT taken between April 2017 and March 2018; 4990 nodules<br>4686 (93.9%) solid; 78 (1.6%) part-solid; 226 (4.5%) pure ground glass<br>Non-enhanced CT, slice thickness < 1.5 mm | AVIEW Lungscreen (Coreline Soft) | *Failure of semi-automatic segmentation* (clinical practice): all nodules | 669/4990 (13.4%) |
| Hwang *et al.* 2021,[50] Republic of Korea | K-LUCAS (Korea)<br>10,424 LDCT taken between April 2017 and December 2018<br>10,080 nodules: 9465 (93.9%) solid; 157 (1.6%) part-solid; 458 (4.5%) pure ground glass<br>Non-enhanced CT, slice thickness < 1.5 mm | AVIEW Lungscreen (Coreline Soft) | *Failure of semi-automatic segmentation* (clinical practice): all nodules; solid nodules; part-solid nodules; ground-glass nodules | 874/10,080 (8.7%); 688/9465 (7.3%); 31/157 (19.7%); 155/458 (33.8%) |
| Hwang *et al.* 2021,[52] Republic of Korea | K-LUCAS (Korea)<br>3353 LDCT taken between April 2017 and December 2017<br>Non-enhanced CT, slice thickness < 1.5 mm | AVIEW Lungscreen (Coreline Soft) | *Failure of semi-automatic segmentation*: 20 radiologists from 14 institutions in clinical practice; central review (1 radiologist, retrospective reading) | 497/3452 (14.4%); range 0–57.0% (coefficient of variation 1.28); 1.1% (107/9389) |
| Singh *et al.* 2021,[56] USA | NLST data set (USA)<br>150 LDCT; first 125 patients with subsolid nodules; first 25 patients with no nodules<br>Non-enhanced CT, slice thickness: 1.2–2 mm | ClearRead CT (Riverain Technologies) | *Software processing failure* due to artifacts and/or thick slices (retrospective MRMC study) | 27/150 (18.0%) |
| Jacobs *et al.* 2021,[64] Denmark, the Netherlands | NLST data set (USA)<br>160 LDCT selected by Lung-RADS category; 40 Lung-RADS 1 or 2; 40 Lung-RADS 3; 40 Lung-RADS 4A; 40 Lung-RADS 4B<br>Non-enhanced CT, slice thickness: 1.0–3.2 mm | Veolity (MeVis) | *Need to manually tune segmentation parameters*<br>*Manual diameter measurement deemed necessary*; retrospective MCMR study | 28% of nodule segmentations<br>3/160 (1.9%) nodules (one reader); 2/160 (1.3%) nodules (two readers); 0/160 nodules (four readers) |
| Hall *et al.* 2022,[27] UK | LSUT study (UK)<br>All 770 LDCT with a lung health check appointment between November 2015 and July 2017; 158 with ≥ 1 nodule (≥ 5 mm or ≥ 80 $mm^3$)<br>Non-enhanced CT, slice thickness: 0.5–1.0 mm | Veolity (MeVis) | *Issues with the CADe software* (no CADe interpretation, CADe processing failure): retrospective MRMC study | Reader 1: 9/770 (1.2%); reader 2: 18/770 (2.3%) |
| ***Surveillance population with applicability concerns (two studies)*** | | | | |
| Cohen *et al.* 2017,[62] Republic of Korea | One hospital in Seoul (Korea)<br>73 patients with preoperative CT scans for subsolid nodules taken between July 2014 and May 2015; 73 subsolid nodules<br>Non-enhanced CT, slice thickness 0.625 mm<br>Reconstructed with FBP and MBIR, respectively | Veolity (MeVis) | *Failure of semi-automatic segmentation* (MRMC study): subsolid nodules – FBP; subsolid nodules – MBIR<br>*Manual modifications of nodule segmentation required* (MRMC study): subsolid nodules – FBP; subsolid nodules – MBIR | 7/73 (9.6%); 5/73 (6.8%)<br>27/73 (37.0%) for reader 1; 43/73 (58.9%) for reader 2<br>(median 35/73, 47.9%)<br>21/73 (28.8%) for reader 1; 39/73 (53.4%) for reader 2 (median 30/73, 41.1%).<br>FBP vs. MBIR ($p = 0.58$) |
| Kim *et al.* 2018,[63] Republic of Korea | One hospital in Seoul (Korea)<br>89 patients with preoperative CT scans for subsolid nodules taken between November 2014 and July 2016; 109 subsolid nodules<br>Non-enhanced CT, slice thickness 0.625 mm | Veolity (MeVis) | *Failure of semi-automatic segmentation* (MRMC study): subsolid nodules | 7/109 (6.4%) |

**TABLE 7** Technical failure rate of AI-based software for lung nodule detection and analysis, by target population and technology (12 studies) (*continued*)

| Reference and country | Population/nodule characteristics/slice thickness | Technology | Details of technical failure | Failure rate |
|---|---|---|---|---|
| ***Mixed population (four studies)*** | | | | |
| Röhrich *et al.* 2023,[31] Austria | One hospital in Austria in 2018<br>First 100 patients with lung pathologies (22 unique, verified diagnoses, but none with lung nodules), first 8 patients without pathological lung findings<br>Slice thickness: 1 mm | contextflow SEARCH Lung CT (contextflow) | 'Technical difficulties' (not further specified), retrospective MRMC study | 2/216 (0.9%) |
| Hempel *et al.* 2022,[34] the Netherlands | One hospital in the Netherlands<br>50 chest CT scans taken between July and September 2013 with ≤ 5 incidentally detected nodules ($n$ = 45 : 35 with and 10 without prior imaging) or no nodules ($n$ = 5) on initial radiology report<br>Slice thickness: 2.00 mm ($n$ = 73) and 3.00 mm ($n$ = 12) | Veye Chest (Aidence) | 'Volumetry not deemed reliable' (retrospective MRMC study): relevant nodules that contributed to the reader's management decision | Reader 1: 1/41 (2.4%); reader 2: 2/44 (4.5%) |
| Martins Jarnalo *et al.* 2021,[66] the Netherlands | One hospital in the Netherlands<br>Random 145 chest CT scans performed for various indications between December 2018 and May 2020<br>91 nodules: 16 subsolid nodules, 73 solid nodules, 2 mixture of solid/subsolid<br>Slice thickness: 1 or 3 mm | Veye Chest (Aidence) | *Failure of semi-automatic segmentation* (retrospective study): all 80 nodules correctly detected by stand-alone software | 3/80 (3.8%) |
| Murchison *et al.* 2022,[33] UK | One hospital in Edinburgh (UK)<br>337 scans of 314 current smokers, ex-smokers and/or those with radiological emphysema between 55 and 74 years taken between January 2008 and December 2009<br>(1) 178 without reported nodules; (2) 95 with 1–10 reported nodules; 23 CT images from the same patients with (3) baseline CT scan and (4) follow-up CT scan; (5) 18 with subsolid nodule<br>Slice thickness 1.0–2.5 mm | Veye Chest (Aidence) | *Failure of semi-automatic segmentation* (retrospective MRMC study): 428 nodules (3–30 mm) from groups 1, 2, 3 and 5 | 21/428 (4.9%) |

FBP, filtered back projection; K-LUCAS, Korean Lung Cancer Screening; LDCT, low-dose computed tomography; MBIR, model-based iterative reconstruction; MRMC, multi-reader multi-case study; NLST, National Lung Cancer Screening.

in screening practice and also compared them with retrospective central review of the same CT images by one experienced chest radiologist. Segmentation failure ranged from 0% to 57.0% (coefficient of variation 1.28) among the 20 original pilot programme radiologists. The frequency of segmentation failure was significantly higher in the original institutional reading (14.4%) than in retrospective central review (1.1%) ($p$ < 0.001), suggesting that segmentation failures in the institutional (clinical practice) reading were mostly rejections of segmentation results by radiologists, rather than the inability of the system to segment the nodule.

*Surveillance population with applicability concerns: Veolity (MeVis) (two studies)*

Kim *et al.*[63] included 89 patients with subsolid nodules who underwent preoperative non-enhanced CT and subsequent surgical resection. Veolity version 1.2 (MeVis) was used in concurrent mode by two experienced radiologists. The segmentation success rate of the software in 109 subsolid nodules was 93.6% (102/109).

The study by Cohen *et al.*[62] comprised 73 patients in whom preoperative CT scans for subsolid nodules were reconstructed on a single CT system and compared the effects of MBIR and FBP algorithms on software (Veolity, MeVis) semiautomatic measurements. Adequate nodule segmentation was obtained in 66 out of 73 (90.4%) images with FBP and in 68 out of 73 (93.2%) of images with MBIR. All seven of the inadequate segmentations were graded as 'insufficient segmentations' for the following reasons: inclusion of a vessel in segmentation (*n* = 2), inclusion of a significant part of the chest wall (*n* = 2), inaccurate segmentation of the ground-glass component (*n* = 1), a combination of those reasons (*n* = 2), inaccurate ground-glass segmentation and chest wall inclusion (*n* = 1) and inaccurate ground-glass segmentation and inclusion of a solid component (*n* = 1). Using FBP, manual modifications were required in 27 cases for reader 1 and 43 cases for reader 2 (median 35 cases). Using MBIR, reader 1 performed manual modifications in 21 cases and reader 2 performed manual modifications in 39 (median 30 cases). The number of manual modifications was similar with FBP and MBIR (*p* = 0.58).

*Mixed population: Veye Chest (Aidence) (three studies)*

The study by Murchison *et al.*[33] included a routine cohort of current or ex-smokers and/or those with radiological evidence of pulmonary emphysema between 55 and 74 years (to mimic a screening population) who underwent chest CT for non-screening purposes at a single centre in Edinburgh (UK) (337 chest CT images from 314 subjects). The Veye Chest software from Aidence was able to successfully segment 95% of the total 428 nodules between 3 and 30 mm.

Martins Jarnalo *et al.*[66] randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital. The study found that Veye Chest (Aidence) reported an unknown diameter for 3 out of 80 (3.8%) nodules between 4 and 30 mm.

Hempel *et al.*[34] selected 50 chest CT scans with incidentally detected nodules (35 with and 10 without prior imaging) or no nodules (*n* = 5) from one hospital in the Netherlands. For this MRMC study, two experienced radiologists independently assessed the CT images to determine the nodule management recommendation based on nodule type and size twice, once using a semiautomated volumetry tool (Vitrea Enterprise Solutions, Vital Images, Inc.) and once using Veye Chest (Aidence) for automatic diameter and volume measurement. With the semiautomated volumetry tool, reader 1 and reader 2 deemed 54.6% (35/64) and 44.4% (28/63) of volume measurements not reliable (and chose to report longest axial diameter instead), whereas with Veye Chest only 2.4% (1/41) and 4.5% (2/44) of volume measurements were deemed not reliable.

*Mixed population: contextflow SEARCH Lung CT (contextflow) (one study)*

From all patients who had CT images performed on one scanner model at a single hospital in Austria in 2018, Röhrich *et al.*[31] included the first 100 patients with lung pathologies (22 unique, clinically and/or histopathologically verified diagnoses, but none with lung nodules) as well as the first eight patients without pathological lung findings. Each CT image was read twice with AI-based software. Two of the in total 216 readings (0.9%) with concurrent software use (contextflow SEARCH Lung CT) had to be excluded due to 'technical difficulties' (no further details reported).

### Radiologist reading time (10 studies)

Ten studies[27,31,34,47,53,54,57,59,60,64] were identified that reported on the reading time of radiologists with and without software support. Three studies[27,54,64] included chest CT images from screening populations, one study[59] included a symptomatic population, and the remaining six studies[31,34,47,53,57,60] included mixed indications for the CT scans. The included studies compared the reading times between unaided readers and readers supported by six different technologies: AI-Rad Companion Chest CT (Siemens Healthineers) in stand-alone and concurrent mode, respectively;[47] ClearRead CT (Riverain Technologies) in concurrent[53,54,57] and assisted second read mode,[53] respectively; contextflow SEARCH Lung CT (contextflow) in concurrent mode;[31] InferRead CT Lung (Infervision) in concurrent mode;[59,60] Veolity (MeVis) in concurrent mode;[27] and Veye Chest (Aidence) in concurrent mode.[34] Nine[27,31,34,47,53,54,59,60,64] of the 10 identified studies reported reduced radiologist reading times by 11.3% to 78% with concurrent software use, whereas one study[57] found similar reading times when using software with vessel suppression function only (*Table 8*). Software assistance as second reader resulted in a significant increase in radiologist reading times by 26% in one study.[53]

**TABLE 8** Effect of software use on radiologist reading time, by target population and technology (10 studies)

| Reference and country | Population | Technology | Index test | Comparator test | Reader task | Effect of software use on reading time compared to unaided reading |
|---|---|---|---|---|---|---|
| ***Symptomatic population (one study)*** | | | | | | |
| Kozuka *et al.* 2020,[59] Japan | 120 chest CT images from cases of suspected lung cancer in patients, one hospital in Japan | InferRead CT Lung (Infervision) | MRMC, two less experienced radiologists, concurrent mode | MRMC, same as 'index test', unaided | To detect any nodules ≥ 3 mm | Concurrent mode: (↓) (−11.3%) |
| ***Screening population (three studies)*** | | | | | | |
| Lo *et al.* 2018,[54] USA | 324 LDCT from the NLST data set and two hospitals (USA), 216 with no actionable nodules, 108 with actionable nodules | ClearRead CT (Riverain Technologies) | MRMC, 12 general radiologists certified by the American Board of Radiology (6–26 years of experience), concurrent mode | MRMC, same as 'index test', unaided | To detect any actionable nodules (5–44 mm) | Concurrent mode: ↓ (−26%) |
| Jacobs *et al.* 2021,[64] Denmark, Netherlands | NLST data set (USA): 160 CT images (40 per Lung-RADS category) | Veolity (MeVis) | MRMC, three experienced chest radiologists and four radiology residents, concurrent mode | MRMC, same as 'index test', unaided | To detect nodules ≥ 3 mm and classify Lung-RADS category of the risk-dominant nodule | Concurrent mode: ↓ all readers (−46%), ↓ 3 experienced chest radiologists (−51%), ↓ 4 radiology residents (−37%) |
| Hall *et al.* 2022,[27] UK | All 770 LDCT from LSUT, London (UK) | Veolity (MeVis) | MRMC, two radiographers without prior experience in thoracic CT, concurrent mode | Clinical practice, LSUT study radiologists, unaided | To detect clinically significant nodules ≥ 5 mm and common incidental findings, to make patient management recommendation based on nodule type and size | Concurrent mode: ↓ radiographer 1 vs. pooled radiologists (−70%), ↓ radiographer 2 vs. pooled radiologists (−50%) |
| ***Mixed population (six studies)*** | | | | | | |
| Abadia *et al.* 2021,[47] USA | Random 103 patients with ≥ 1 lung condition and ≥ 1 lung nodule; 40 patients with ≥ 1 lung condition and no lung nodules from a single US hospital | AI-Rad Companion Chest CT (Siemens Healthineers), Prototype | Stand-alone mode. MRMC, one expert thoracic radiologist (15 years of experience) reading a random 20/103 CT images with nodules, concurrent mode | MRMC, same as 'index test', unaided; reading all 143 CT images | To detect nodules and measure size of the five largest nodules | Concurrent mode: ↓ (−78%) |

continued

**TABLE 8** Effect of software use on radiologist reading time, by target population and technology (10 studies) *(continued)*

| Reference and country | Population | Technology | Index test | Comparator test | Reader task | Effect of software use on reading time compared to unaided reading |
|---|---|---|---|---|---|---|
| Hsu *et al.* 2021,[53] Taiwan | 150 consecutive cases with lung nodules ≤ 1 cm or no nodules on chest CT performed at a single hospital in Taiwan; 93 standard dose from clinical routine, 57 LDCT from screening | ClearRead CT (Riverain Technologies) with vessel suppression and nodule detection functions | MRMC<br>'Junior group': six radiology residents, > 6 month of chest CT experience<br>'Senior group': 6 experienced chest radiologists, 5, 10 and 25 years of experience, respectively<br>Concurrent mode; 2nd-read mode | MRMC, same as 'index test', unaided | To detect any nodule (3–10 mm) | Concurrent mode:<br>↓ for all readers (−21%), ↓ for radiology residents (−23%), ↓ for experienced chest radiologists (−16%)<br>Assisted 2nd-read mode:<br>↑ for all readers (+ 26%), ↑ for radiology residents (+ 28%), ↑ for experienced chest radiologists (+ 24%) |
| Takaishi *et al.* 2021,[57] Japan | 61 thoracic or thoracic-abdominal unenhanced CT images conducted at a single hospital in Japan during September 2019; mixed indication | ClearRead CT (Riverain Technologies) with vessel suppression function only | MRMC<br>Six radiologists with 2–8 years of experience, Concurrent mode (vessel-suppressed CT images) | MRMC, same as 'index test', unaided (standard CT images) | To detect nodules ≥ 4 mm in maximum diameter | Concurrent mode:<br>= All readers (+ 9.5%), = reader A, ↑ reader B, ↓ reader C |
| Röhrich *et al.* 2023,[31] Austria | 108 CT images from one hospital in Austria; first 100 patients with lung pathologies (no lung nodules), first 8 patients without pathological lung findings | Contextflow SEARCH Lung CT (contextflow) | MRMC<br>Six radiology residents (mean 2.1 ± 0.7 years of experience), 4 attending general radiologists (mean 12 ± 1.8 years of experience)<br>Each image read by one radiology resident and one attending general radiologist, concurrent mode | MRMC<br>Same readers as 'index test' but each image only read once by each reader (with or without software), unaided | To interpret the CT images (diagnosis of lung pathologies) | Concurrent use:<br>↓ (−31.3%), (↓) radiology residents, (↓) attending general radiologists |
| Liu *et al.* 2019,[60] China | 123 (batch 1) and 148 (batch 2) chest CT images (screening and inpatient) from > 10 hospitals in China | InferRead CT Lung (Infervision) | MRMC<br>Two thoracic radiologists with approximately 10 years' experience, concurrent mode | MRMC, same as 'index test', unaided | To detect any nodules (size NR) | Concurrent mode: (↓) for both readers (33–66%) |

**TABLE 8** Effect of software use on radiologist reading time, by target population and technology (10 studies) (*continued*)

| Reference and country | Population | Technology | Index test | Comparator test | Reader task | Effect of software use on reading time compared to unaided reading |
|---|---|---|---|---|---|---|
| Hempel *et al*. 2022,[34]<br>Netherlands | 50 patients with ≤ 5 incidentally detected nodules (*n* = 45) or no nodules (*n* = 5) on initial radiology report with (*n* = 35) and without (*n* = 10) prior CT imaging from one Dutch hospital | Veye Chest (Aidence) | MRMC<br>One chest radiologist with 15 years of experience and 1 general radiologist with 13 years of experience, concurrent mode | MRMC, same as 'index test', unaided | To determine the nodule management recommendation and report relevant pulmonary nodules that contributed to management decision | Concurrent mode: ↓ for both readers (−33.4% and −42.6%)<br>Subanalysis for patients where an equal number of nodules was reported during aided and unaided reading sessions: (↓) for both readers (−38.0% and −30.3%) |

LDCT, low-dose CT images; MRMC, multi-reader multi-case study; NR, not reported.

**Notes**

↑ Significant increase; (↑) increase but no *p*-value or 95% CI reported.
= No significant change; (=) no change but no *p*-value or 95% CI reported.
↓ Significant decrease; (↓) decrease but no *p*-value reported.

*Symptomatic population: InferRead CT Lung (Infervision) (one study)*

Kozuka *et al.*[59] randomly selected 120 chest CT images from cases of suspected lung cancer in patients at a single hospital in Japan. In a MRMC study, two less experienced radiologists independently read the CT images first without software and then (after at least 14 days) with concurrent use of the software InferRead CT Lung (Infervision) to detect any nodules ≥ 3 mm. The total reading time decreased by 10.4% in reader A and by 11.9% in reader B (no level of significance reported). The total mean reading time of the average reader decreased by 11.3% with software use, from 373 to 331 minutes, reducing the mean reading time for one case from 3.1 minutes without software to 2.8 minutes with software (no level of significance reported).

*Screening population: Veolity (MeVis) (two studies)*

The study by Jacobs *et al.*[64] comprised a nodule-enriched screening population. Seven observers read all 160 CT images twice: once in the dedicated CT lung screening viewer including Veolity Lung CAD (MeVis) and once in the standard viewer without software support. Pooling all results, the median reading time of 86 seconds (IQR 51–141 seconds) when using the dedicated viewer was shorter than the median reading time of 160 seconds (IQR 96–245 seconds) when using the standard viewer ($p$ < 0.001).

The pooled median reading times of the three experienced chest radiologists reduced from 214 seconds (IQR 155–307 seconds) without software support to 105 seconds (IQR 61–158 seconds) with software support ($p$ < 0.0001). In the four less experienced radiology residents, the pooled reading time decreased significantly from a median of 118 seconds (IQR 78–182 seconds) in unaided readers to a median of 74 seconds with software support (IQR 46–128 seconds) ($p$ < 0.0001).

The MRMC study by Hall *et al.* comprised all 770 patients who received LDCT for lung cancer screening as part of the LSUT study.[27] Two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCT images with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm) and common incidental findings, including patient management recommendations. Self-reported reading times of each software-assisted radiographer were compared against the reading times of the pooled study radiologists who read the same CT images in clinical practice without software support. Reading times were available for 753 (97.8%) of radiologist reports, 738 (95.8%) of reports by radiologist 1 and 754 (97.9%) of reports by radiologist 2. Unaided radiologists recorded significantly longer and more variable reading times than either software-supported radiographer, with median reading times of 10 minutes (IQR 5–15 minutes) for the pooled radiologists versus 3 minutes (IQR 2–5 minutes) for radiographer 1 and 5 minutes (IQR 4–8 minutes) for radiographer 2 ($p$ < 0.001 for both comparisons).

*Screening population: ClearRead CT (Riverain Technologies) (one study)*

The MRMC study by Lo *et al.*[54] comprised a nodule-enriched screening population. Twelve general radiologists independently read the LDCT images first unaided and then with the concurrent use of ClearRead CT (Riverain Technologies) to detect any actionable nodules (5–44 mm). The radiologist interpretation time decreased from 132.3 seconds per case in the unaided reading session to 98.0 seconds per case with concurrent software use ($p$ < 0.01). The study showed that concurrent software use resulted in a significant (> 25%) decrease in interpretation time (mean 34.3 seconds, 95% CI 15.2 to 53.5 seconds) in a nodule-enriched data set.

*Mixed population: AI-Rad Companion Chest CT (Siemens Healthineers) (one study)*

Abadia *et al.*[47] included 103 patients with at least one lung condition and one suspicious lung nodule on radiology report and 40 patients with one lung condition and no lung nodule on radiology report. In a MRMC study, an expert thoracic radiologist read all 143 CT images without software support to detect nodules and to measure nodule size of the five largest nodules ≥ 4 mm. A month after initial assessment, the radiologist re-evaluated 20 positive cases at random with the assistance of an AI-Rad Companion Chest CT prototype. The average amount of time (minute:second) spent for analysis per image was 2:17 ± 0:29 for the stand-alone software and 2:44 ± 0:54 for the unaided expert. With concurrent software use, the expert saved on average 1:45 minutes per patient, significantly reducing the mean assessment time to 35.7 seconds per case ($p$ < 0.0001). Assuming continuous work, the unaided expert would have been able to assess ≈ 26 cases for lung nodules per hour, whereas, with the help of AI-Rad, the radiologist could assess 101 cases for nodules per hour.

*Mixed population: InferRead CT Lung (Infervision) (one study)*

In the study by Liu *et al.*,[60] chest CT scans (screening and inpatient) performed at multiple hospitals in China were retrospectively collected with convenience sampling. The total data set comprised 12,574 CT scans, of which 1,129 from more than 10 hospitals were included in the test set. In a MRMC study of a subset of 123 (batch 1) and 148 (batch 2) CT images, two thoracic radiologists independently first read the scans alone without using software and then performed reading with concurrent software use (InferRead CT Lung, Infervision) after a 1-week washout period to detect any nodules. The reading time was limited to approximately 20 minutes per scan (a typical reading period for radiologists at a top-tier hospital). Both radiologists experienced shorter reading time with concurrent software use, with a reduction from approximately 15 minutes per patient to approximately 5–10 minutes per patient (no level of significance reported).

*Mixed population: ClearRead CT (Riverain Technologies) (two studies)*

The study by Hsu *et al.*[53] retrospectively included 150 consecutive cases with lung nodules ≤ 1cm or no nodules on chest CT performed at a single hospital in Taiwan. Of these, 93 were standard-dose CT images from clinical routine and 57 were LDCT scans from lung cancer screening. The reader study with the request to detect any nodule (3–10 mm) included a ‘junior group’ (three residents in radiology, 1–2 years of CT experience and at least 6 months of chest CT experience) and a ‘senior group’ (three experienced chest radiologists with 5, 10 and 25 years of experience, respectively). In 2nd-read mode, readers first read the CT images without software and then combined the displays of the software results (ClearRead CT, Riverain Technologies, with vessel suppression and nodule detection functions) to make the final decision. In concurrent-read mode, the software results were simultaneously displayed to readers during the reading.

For all readers, the mean reading time per case was 2 minutes 36 seconds (range 100–227 seconds) for unaided readers, 3 minutes 17 seconds (range 118–278 seconds) in the 2nd-read mode, and 2 minutes 4 seconds (range: 82–171 seconds) in the concurrent-read mode. The reading time of all readers was significantly shorter with the concurrent-read mode than with the manual review mode (mean difference 32 seconds, −21%; $p$ < 0.001) and the assisted 2nd-read mode (mean difference 73 seconds; $p$ < 0.001). Similar results were found for both junior and senior readers: mean reading time per case for junior radiologists was 183 seconds for unaided readers, 235 seconds for 2nd-read mode and 141 seconds for concurrent mode ($p$ < 0.001 for all). Mean reading time per case for senior radiologists was 128 seconds for unaided readers, 159 seconds for 2nd-read mode and 107 seconds for concurrent mode ($p$ < 0.001 for all).

Takaishi *et al.*[57] included 61 thoracic or thoracic-abdominal unenhanced CT images conducted at a single hospital in Japan for various reasons. The MRMC study comprised three radiologists who either read standard CT images alone or both vessel-suppressed CT (ClearRead CT, Riverain) and standard CT images randomly to identify pulmonary nodules ≥ 4 mm in maximum diameter. The mean reading time increased significantly from 16.9 seconds without software use to 32.3 seconds with software use ($p$ < 0.01) in reader B, decreased significantly from 39.3 seconds without software use to 33.6 seconds with software use in reader C ($p$ = 0.09) and was unchanged (31.5 vs. 31.2 seconds) in reader A. The average reading time of all three radiologists was slightly longer with software use (29.2 seconds vs. 32.3 seconds, + 9.5%, $p$ = 0.11).

*Mixed population: contextflow SEARCH Lung CT (contextflow) (one study)*

From all patients who had CT images performed on one scanner model at a single hospital in Austria in 2018, Röhrich *et al.*[31] included the first 100 patients with lung pathologies as well as the first eight patients without pathological lung findings. The 108 distinct cases were distributed to eight participants taking part in a MRMC study, balancing out diseases between sets, where possible. Each participant interpreted 54 CT images (27 without software support and another 27 with concurrent use of contextflow SEARCH Lung CT), resulting in each CT image being read four times (two times with and without software, respectively).

The reduction in time taken per case with software support was more distinct for cases where the participants looked for other information than for cases where they did not (110 vs. 39 seconds saved; $p = 0.002$). Both the radiology residents and attending radiologists showed a decrease in reading time with concurrent software use, and there was a tendency towards a stronger decrease in reading time for senior radiologists (27% vs. 35%; $p = 0.078$). The modelled overall time used per case, controlling for individual participants, experience level and whether they looked for information, was reduced by 31.3% when using the software ($p < 0.001$).

*Mixed population: Veye Chest (one study)*

Hempel *et al.*[34] selected 50 chest CT scans with incidentally detected nodules (35 with and 10 without prior imaging) or no nodules ($n = 5$) from one hospital in the Netherlands. For this MRMC study, two experienced radiologists independently assessed the CT images to determine the nodule management recommendation grade based on the 2015 BTS guidelines[12] (A, discharge; B, CT at 3 months; C, Brock score; D, diagnostic work-up) twice, first unaided and then with concurrent use of Veye Chest software (Aidence). For both readers, the reading time was significantly reduced by 33.4% and 42.6%, respectively ($p < 0.001$ for both) with concurrent software use. To investigate if the reduced reading times could be attributed to the fact that the readers reported fewer actionable nodules with software use, a subgroup analysis of patients where an equal number of nodules was reported during both sessions was performed that found reading time reductions by 38.0% for reader 1 and 30.3% for reader 2.

### *Radiology report turnaround time (no study)*

No study was identified that assessed the radiology report turnaround with and without AI-based software use for the detection and analysis of lung nodules.

### *Acceptability and experience of using the software (three studies)*

Three studies were identified that assessed readers' acceptability and experience of using AI-based software for the detection and analysis of lung nodules.[27,47,66] One study was performed in a screening population[27] and the other two were in mixed populations.[47,66]

*Screening population: Veolity (MeVis) (one study)*

This substudy of the LSUT trial performed in London (UK) comprised all 770 patients who received LDCT for lung cancer screening.[27] In a reader study, two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCT images with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm) and common incidental findings, including patient management recommendations. Reader 1 and reader 2 deferred 6.5% (48/733) and 10.8% (82/760) of completed CT scans for discussion with a radiologist ($p = 0.015$).

*Mixed population: AI-Rad Companion Chest CT (Siemens Healthineers) (one study)*

Abadia *et al.*[47] included 103 patients with at least one lung condition and one suspicious lung nodule on radiology report and 40 patients with one lung condition and no lung nodule on radiology report. In a MRMC study, an expert thoracic radiologist read all 143 CT images without software support to detected nodules and measure nodule size of the five largest nodules ≥ 4 mm. One month after initial assessment, the radiologist re-evaluated 20 positive cases at random with the assistance of an AI-Rad Companion Chest CT prototype. The radiologist reported increased confidence in lung nodule detection for all 20 cases (100%).

*Mixed population: Veye Chest (Aidence) (one study)*

Martins Jarnalo *et al.*[66] randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital. The authors reported in the discussion that the single system threshold setting for nodule detection of the Veye Chest software from Aidence for various uses (e.g. follow-up vs. screening) has been found to be a limitation, and that it would be a useful improvement to be able to set different thresholds.

### *Other non-prespecified outcomes*

One study[32] reported on the simulated radiologist workload reduction when stand-alone AI-based software would be used as prescreen to rule out CT images with no or only benign nodules. This outcome was not prespecified in the protocol; results are reported in *Appendix 6*.

### *Subquestions 1 to 4: potential factors influencing practical implications*

No data were available to perform subgroup analyses based on contrast use, radiation dose, nodule type, patient ethnicity, radiologist speciality or reasons for CT scan (incidental population).

## Impact on patient management

### *Characteristics of detected nodules*

Most useful are studies that report characteristics of detected and missed nodules in readers assessing the same CT images with and without concurrent software use. These comparative studies will be prioritised in the following sections, with a focus on changes in detected and missed nodule characteristics due to software use. Non-comparative results are reported in *Appendix 6*, *Table 64*, and text.

#### All detected nodules (true positive and false positive) (six studies)

Six studies[34,47,50–52,66] were identified that reported on the characteristics of all detected nodules (true positives and false positives). Three studies were performed in consecutive screening populations,[50–52] and the remaining three studies comprised mixed populations.[34,47,66] Only one MRMC study[34] compared the characteristics of all nodules detected in the same CT images by readers with and without concurrent software use, respectively. With concurrent software use, the two readers reported less actionable nodules, and the proportion of solid nodules was lower than with unaided reading (87.1% vs. 90.6%, no level of significance reported).[34] A second study[51] used an unpaired design and reported nodule characteristics before and after software implementation in prospective screening practice. By contrast, this study observed a significantly larger ($p < 0.001$) number of nodules detected per participant and higher proportion of solid nodules with software use. No significant difference ($p > 0.05$) was observed in nodule size when nodules were measured on transverse planes. Further details on the findings of individual studies can be found in *Appendix 5*, *Tables 49–51* and text.

#### True-positive nodules (seven studies)

Seven studies[32,51,56,59–61,66] reported characteristics of correctly detected nodules. Four studies[32,51,56,61] were performed in screening populations, in one study[59] the indication for the chest CT scan was lung cancer suspicion, and in the remaining two studies[60,66] the indication for the chest CT scan was mixed. Of these, two studies[59,61] compared the characteristics of true-positive nodules in readers assessing the same CT images with and without software use (InferRead CT Lung, Infervision). Additional true-positive nodules detected with software use were 56–57% solid, due to larger improvements in the detection of subsolid nodules. This resulted in a lower proportion of solid nodules and higher proportions of part-solid and ground-glass nodules with software use. Twenty-two per cent of additional true-positive nodules were ≥ 6 mm.[59] Further details of the findings of individual studies can be found in *Appendix 5*, *Tables 52–54* and text.

#### Additional true-positive nodules detected by software compared with unaided reading (one study)

*Incidental population: AI-Rad Companion Chest (Siemens Healthineers) (one study)*

The study by Rückel *et al.*[49] comprised 105 consecutive patients who received a whole-body CT scan in the emergency department of a single German hospital. Retrospective reading by stand-alone software (AI-Rad Companion Chest CT prototype, Siemens Healthineers) detected three additional true-positive nodules compared with the original radiologist report (17% of CT scans have been originally reported by a board-certified radiologist alone, the other 83% CT scans have been commonly reported by a radiology resident and a board-certified radiologist). All three additional nodules detected measured at least 6 mm, with the largest nodule being 8 mm.

### False-positive nodules (four studies)

Four studies reported on characteristics of false-positive nodules detected by stand-alone software in a random screening population,[48] an incidental population[49] and mixed populations,[47,66] respectively. Findings of the non-comparative evidence can be found in *Appendix 5*, *Table 55* and text. No study compared characteristics of false-positive nodules between readers with and without concurrent software use.

### False-negative (missed) nodules (nine studies)

Nine studies reported characteristics such as nodule size and type of missed nodules: four studies[27,51,56,61] were performed in screening populations, one study[59] was performed in a symptomatic population, and the other four studies[47,58,60,66] were performed in populations with mixed indication for the chest CT scan. Of these, two studies[59,61] compared the characteristics of missed nodules in readers assessing the same CT images with and without concurrent software use (InferRead CT Lung, Infervision). Software use decreased the number of missed nodules in both studies. Relative reductions were larger for part-solid and ground-glass nodules than for solid nodules, with the result that the nodules missed with software use had a higher proportion of solid nodules and a lower proportion of subsolid nodules than nodules missed by unaided readers. Further details of the findings of individual studies can be found in *Appendix 5*, *Tables 56* and *57* and text.

### *Proportion of detected nodules that are malignant (three studies)*

Three studies[27,50,51] performed in consecutive screening populations reported on the proportion of detected nodules that were diagnosed as lung cancer. The two comparative studies found that the proportion of detected actionable nodules that were malignant was 6.6% and 21.3%, respectively, without software use and 5.2% and 16.7–19.4%, respectively, with software use.[27,51] Further details of the findings of individual studies can be found in *Appendix 5*, *Table 58* and text.

### *Impact of test result on clinical decision-making (six studies)*

Six comparative studies[27,55,56,63,64,67] were identified that reported the impact of software use on clinical decision-making. Four studies[27,56,64,67] were performed in screening populations, one study[63] was performed in a surveillance population with applicability concerns, and in the remaining study[55] the indication for the chest CT scan was not reported. Four studies consistently reported that with software use, readers tended to upstage rather than downstage Lungs-RADS[64,67] or Fleischner risk categories.[34,63] Further details of the findings of individual studies can be found in *Appendix 5*, *Tables 59–62* and text.

### *Number of people having computed tomography surveillance (five studies)*

Five studies reported on the number of people referred for CT surveillance ('intermediate nodules'),[27] people followed up as they had nodules suspected to be benign,[59] and the number of people classed as Lungs-RADS categories 3 or 4A[51,52,64] or 'intermediate' according to the NELSON criteria.[52] Four studies were performed in consecutive[27,51,52] or nodule-enriched screening populations,[64] and one study was performed in a random symptomatic population.[59] Of these, a MRMC study[64] and a before-and-after study[51] reported the proportion of people with Lungs-RADS categories 3 and 4A in readers with and without concurrent software use. Both studies found increased proportions of people classed as Lung-RADS 3 or 4A with software use. Further details on the findings of individual studies can be found in *Appendix 5*.

### *Number of computed tomography scans taken as part of computed tomography surveillance (no study)*

No study was identified that reported on the number of CT scans that were taken as part of CT surveillance.

### *Number of people having a biopsy or excision (five studies)*

Five studies reported on the number of people directly referred to MDT because of 'suspicious nodules',[27] of people with lung cancer diagnosed or followed up as they had nodules suspected of lung cancer,[59] and the number of people positive on the narrow definition using Lungs-RADS (i.e. category 4B or 4X by Lung-RADS)[51,52,64] or 'positive' according to NELSON criteria.[52] Four studies were performed in consecutive[27,51,52] or nodule-enriched screening populations,[64] and one study was performed in a random symptomatic population.[59] Of these, a MRMC study[64] and a before-and-after study[51] reported the proportion of people with Lungs-RADS categories 4B or 4B and 4X in readers with and without concurrent software use. The studies found similar or slightly higher proportions of people classed as Lung-RADS 4B/4X with software use. Further details on the findings of individual studies can be found in *Appendix 5*.

### *Stage of cancer at detection (no study)*

No study was identified that reported on the stage of lung cancer at detection.

### *Time to diagnosis (one study)*

One study[67] was identified that mentioned the potential effect of software use on the time to diagnosis in a nodule- and cancer-enriched screening population (200 baseline LDCT), selected from the US-based NLST data set. This MRMC study evaluated the effects of using the software VUNO Med-Lung CT AI (VUNO) on Lung-RADS categorisation. Five readers with varying levels of experience assessed the LDCT images with and without concurrent software use. For the 31 cancer-positive cases in the data set, substantial management discrepancies between the 310 reader pairs (Lung-RADS category 1/2 vs. 4A/B) were reduced by half (32/310 vs. 16/310) and pooled sensitivity for lung cancer significantly improved (85.2% vs. 91.6%; $p = 0.004$) with software use. This could eventually lead to an earlier diagnosis of lung cancer if confirmed in prospective studies in clinical practice.

### *Other non-prespecified outcomes*

Other outcomes not prespecified in the protocol are reported in *Appendix 6*. Three studies[50–52] based on consecutive participants from the K-LUCAS project (with possibly overlapping populations) reported on the positivity rate (proportion of people with Lung-RADS category ≥ 3) of LDCT images taken and assessed in screening practice with and without the use of the AVIEW Lungscreen software (Coreline Soft).

### *Subquestions 1 to 4: potential factors influencing impact on patient management*

No data were available to enable subgroup analyses based on contrast use, radiation dose, nodule type, patient ethnicity, radiologist speciality or reasons for CT scan (incidental population).

## Ongoing and/or unpublished studies

We identified seven relevant ongoing and/or unpublished studies from clinical trial registers and/or company submissions. The characteristics of ongoing studies are described in *Report Supplementary Material 1*, *Table 4*.

# Chapter 4 Systematic review of clinical effectiveness (key question 2): methods and results

## Methods

### *Identification and selection of studies*

#### Search strategy

The same search strategy as described in the methods for test accuracy was used (see *Identification and selection of studies*).

#### Study eligibility criteria

The study eligibility criteria were as follows:

***Population***

See *Study eligibility criteria*.

***Target condition***

Lung cancer.

***Intervention***

See *Study eligibility criteria*.

***Comparator***

Computed tomography scan review by a radiologist or another healthcare professional without AI-based software for automated detection and analysis of lung nodules (using diameter or volume to measure nodule size).

Where data permit, the following subgroups may be considered: general radiologist/other healthcare professional without software support; radiologist/other healthcare professional with thoracic speciality without software support.

***Outcomes***

- Morbidity (including any adverse events caused by assessment or treatment).
- Mortality.
- Health-related quality of life.
- Patients' acceptance of use of the software.

***Study design***

- Randomised controlled trials.
- Quasi-randomised trials.
- Cohort studies (retrospective/prospective).
- Before-and-after studies.
- Historical controlled studies.
- Qualitative studies (for patient acceptance of use of the software).

### *Publication type*

- Peer-reviewed papers.
- Conference abstracts and manufacturer data will be included. Only outcome data that have not been reported in peer-reviewed full-text papers will be extracted and reported.

### *Language*

English.

The same exclusion criteria as described in *Study eligibility criteria* were used.

### Study screening and selection

Two reviewers (JG/AA/SJ) independently screened the titles and abstracts of records identified by the searches and documents submitted by the companies through NICE. Any disagreements were resolved through discussion, or retrieval of the full publication. Potentially relevant publications were obtained and assessed independently by two reviewers (JG/AA/SJ). Disagreements were resolved through consensus, with the inclusion of a third reviewer (CS, YFC) if required. Records excluded at full-text stage were documented, including the reasons for their exclusion (see *Report Supplementary Material 1*, *Tables 1 and 2*).

## Results

No studies on intermediate outcomes (e.g. potential benefits from earlier nodule detection and shorter time to diagnosis; potential harms of increased surveillance to patients with benign nodules) and final health outcomes were identified (see *Appendix 1*, *Figure 17*). Consequently, the potential impact of AI-assisted nodule detection and analysis on final health outcomes was modelled using a linked evidence approach through a decision-analytic model and simulations using evidence from the systematic review of test accuracy evidence and additional types of evidence collected as described in *De novo cost-effectiveness analysis (full model): methods*. *Figure 6* illustrates the linked evidence approach.

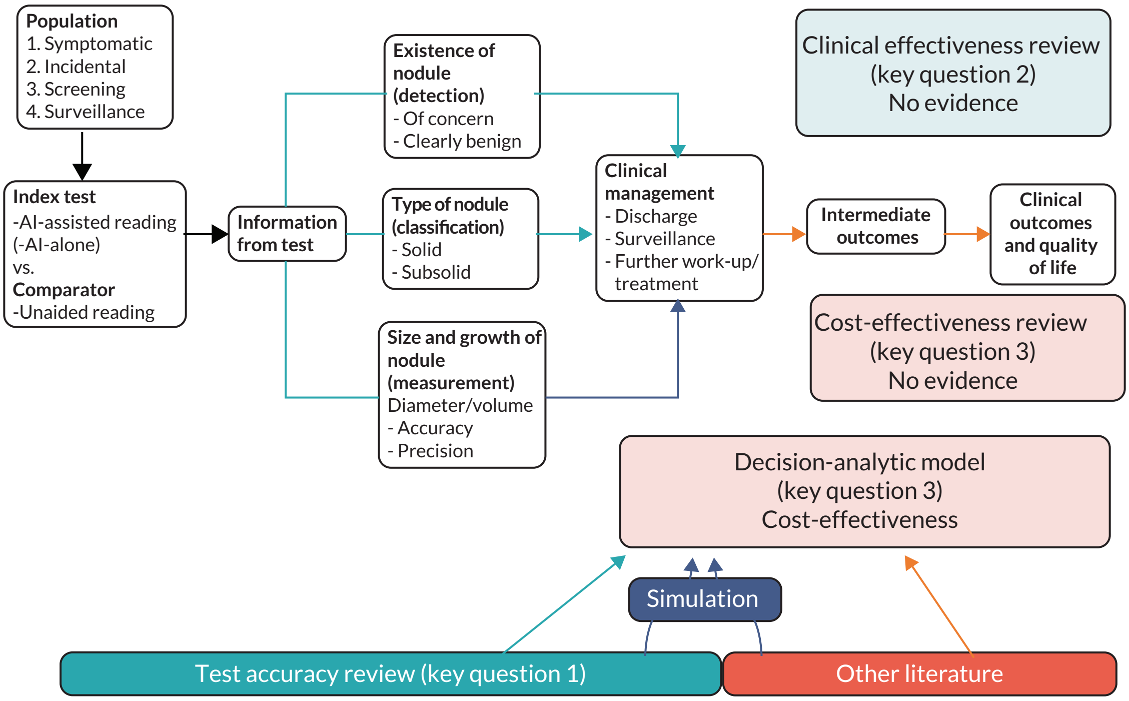


**FIGURE 6** An illustration of linked evidence approach adopted for this diagnostic assessment.

# **Chapter 5** Systematic review of cost-effectiveness (key question 3): methods and results

The majority of published model-based economic analyses related to nodule detection have considered the costs and benefits (and harms) of different strategies to screen for lung cancer in people who are at increased risk. However, the cost-effectiveness of nodule management strategies has not been assessed in detail,[71] especially with using AI software. Algorithms designed for nodule assessment and management use information to predict malignancy and may influence screening outcomes.[71]

This systematic review aimed to assess the cost-effectiveness of software for automated lung nodule detection and analysis from CT images, compared to unassisted CT analysis, in individuals undergoing chest CT scans for suspected lung cancer symptoms, unrelated purposes, nodule surveillance, or lung cancer screening.

## Methods for systematic review of cost-effectiveness

### *Identification and selection of studies*

#### Search strategy

The searches carried out for the systematic review of test accuracy and clinical effectiveness (see *Search strategy*) were centred around the concepts of AI, lung nodules/cancer and CT or screening, without any restrictions in terms of study type filters. They could therefore be expected to also retrieve any studies relating to cost-effectiveness of using AI-based software in lung nodule/cancer CT imaging.

Given the likely scarcity of economic evaluations on AI-based software for nodule detection in this context, broader searches on lung nodule/cancer imaging or screening (excluding AI and CT terms) were conducted to gather data on model structures, costs, and utility values for the economic model. Search filters for economic evaluations and/or cost or HRQoL studies were applied where relevant.

Sources were MEDLINE All (via Ovid); EMBASE (via Ovid); NHS Economic Evaluation Database (NHS EED) (Centre for Reviews and Dissemination); HTA database (Centre for Reviews and Dissemination); International HTA database (INAHTA); Cost-Effectiveness Analysis registry (Tufts Medical Center); EconPapers [Research Papers in Economics (RePEc)]; ScHARRHUD; targeted web searches (Google); selected organisations and conferences of interest (NICE, CADTH, ISPOR, HTAi, International Health Economics Association and Radiological Society of North America Annual Meetings) and reference lists of selected highly relevant papers. Full search strategies can be found in *Appendix 3*.

#### Study eligibility criteria

Studies that satisfied the following criteria were included.

##### *Population*

See *Study eligibility criteria*.

##### *Target condition*

Lung cancer.

##### *Intervention*

See *Study eligibility criteria*.

##### *Comparator*

Computed tomography scan review by a radiologist or another healthcare professional without AI-based software for automated detection and analysis of lung nodules (using diameter or volume to measure nodule size).

We also considered the following subgroups: general radiologist/other healthcare professional without software support; radiologist/other healthcare professional with thoracic speciality without software support.

#### *Outcomes*

Cost-effectiveness (e.g. incremental costs, incremental benefits, ICER, QALYs).

#### *Study design*

Full economic evaluations (including cost-effectiveness analysis, cost–utility analysis and cost–benefit analysis). Cost-minimisation analysis, cost–consequences/outcome description, costs analysis (UK only) and cost description (UK only) were taken into account when full economic evaluations were lacking.

#### *Publication type*

Peer-reviewed papers.

Abstracts and manufacturer data were included, but only outcome data that have not been reported in peer-reviewed full-text papers were extracted and reported.

#### *Language*

English.

Exclusion criteria are the same as described in *Study eligibility criteria*.

### Study screening and selection

All records retrieved were screened independently by two reviewers (PA/HG) at title/abstract stage, from which potentially relevant records were further examined at full-text stage. Any disagreements between the reviewers were resolved by a discussion, or recourse to a third reviewer (AA or JM) if an agreement could not be reached.

## *Extraction and study quality*

### Data extraction strategy

Information was extracted by two reviewers (PA/HG) independently, using a pre-piloted data extraction form for the full economic evaluation studies. The data extraction form was developed to summarise the main characteristics of the studies and to capture useful information for the economic model. From each paper included in the systematic review, we extracted information about study details (title, author, country, study setting  and year of study), baseline characteristics (population, intervention, comparator and outcomes), methods (study perspective, time horizon, discount rate, measure of effectiveness, assumptions and analytical methods), results (study parameters, base-case and sensitivity analysis results), discussion (study findings, limitations of the models and generalisability), other (source of funding and conflicts of interests), overall reviewer comments and conclusion (author's and reviewer's). Each reviewer cross-checked each other's extractions, with any discrepancies resolved by discussion, or recourse to a third reviewer (AA or JM) if an agreement could not be reached.

### Assessment of study methodological quality

The quality of any full economic evaluation studies was assessed using The Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist.[72] Any studies using an economic model were further assessed against the framework for the quality assessment of decision analytic modelling developed by Philips *et al.*[73]

## *Methods of analysis/synthesis*

Due to the nature of economic analyses (different aims/objectives, study designs, populations, and methods), the findings from individual studies were compared narratively, and recommendations for future economic analyses were discussed.

## Results of systematic review of cost-effectiveness

### *Results of literature search*

The literature search identified 1988 records through electronic database searches and other sources. After duplicates were removed, 1299 studies were screened for inclusion based on title and abstract. Fifteen studies were considered potentially relevant and were reviewed at full-text stage. All studies were excluded at the full-text stage as they compared different strategies beyond the scope of this assessment.

No studies met the eligibility criteria; however, two included relevant interventions or comparators (AI technologies not covered in this review). Therefore, we summarized these studies. [74,75] (see *Report Supplementary Material 3*)

Given that we have not identified any relevant studies for the systematic review, we did not undertake any formal data extraction or quality appraisal. However, we retained studies that might have contained relevant information that could be used to populate the model. Where there was more than one source of information/input, we provided justification for selecting specific input(s).

# Chapter 6 Preliminary model: methods and results

The External Assessment Group (EAG) used two separate modelling approaches. This chapter describes a simpler approach for evaluating the cost-effectiveness of AI-assisted detection of actionable lung nodules in the screening population. This method directly incorporates test accuracy data (sensitivity and specificity) as key model inputs. However, due to insufficient test accuracy data—such as studies reporting only sensitivity without specificity—a similar analysis could not be conducted for the symptomatic and incidental populations.

## Developing the model structure

We developed a decision tree to assess the cost-effectiveness of image analysis assisted by software with AI-derived algorithms for the detection of people with actionable lung nodules from CT images compared with unassisted CT image analysis in CT scans for lung cancer screening. The model structure is presented in *Figure 7*.

With clinical input, we developed our model in TreeAge Pro (TreeAge Software Inc., Williamstown, MA, USA) to represent the BTS-recommended clinical pathway for screening actionable lung nodules. An actionable nodule is defined as one that, if identified, would require further investigation, surveillance, or definitive diagnostic work-up per BTS guidelines. Key criteria for actionable nodules include size (≥ 5 mm) and the absence of features strongly suggestive of benignity. Other factors, such as nodule type (solid or subsolid), location, and morphology (e.g., shape and boundary), are also considered.

The decision tree model structure consists of identifying actionable nodules and then stratifying their 'observed' sizes (5 mm to < 8 mm, or ≥ 8 mm), which are associated with both subsequent nodule management pathways and cancer risks. The branches to the right of the decision node (square symbol) represent the strategies being compared. People being screened by strategy may have an actionable nodule(s) or no actionable(s), which is characterised by the prevalence to the right of the chance node (first circle symbol emanating from the prevalence). Based on the test result and whether people have an actionable lung nodule they will be categorised as 'actionable nodule detected' (true positive), 'actionable nodule missed' (false negative), 'actionable nodule reported' (false positive) or 'no actionable nodule reported' (true negative), and these results are based on test sensitivity and specificity. If lung nodules are observed (detected or reported) we assumed that they would have been categorised/measured at 5 mm to < 8 mm or ≥ 8 mm. The pathways combine the probabilities/conditional probabilities following a particular path and the associated costs and benefits that are captured at the end node (triangle).

The decision tree was modelled from the presence of actionable lung nodule, followed by the conditional probability of an individual with/without actionable nodule(s) testing either positive or negative, respectively. However, in clinical practice, the test result is obtained before the presence or absence of an actionable nodule is confirmed. Modelling the test result first followed by the presence of actionable nodule or vice versa makes no mathematical difference in terms of the expected values calculated.[76] We considered this illustrative structure appropriate as it depicts the clinical pathways to allow for the detection of actionable lung nodules and allows for the economic analysis of the costs and benefits associated with the two screening strategies being compared.

## Strategies

The model compares AI-assisted radiologist reading with unaided radiologist reading.

### *Artificial intelligence-assisted radiologist reading*

In this strategy, the software uses algorithms that have been produced using AI. AI is used to assist the radiologist or other healthcare professionals to identify lung nodules and measure their sizes, with or without additional features, such as classifying the type of the nodules.

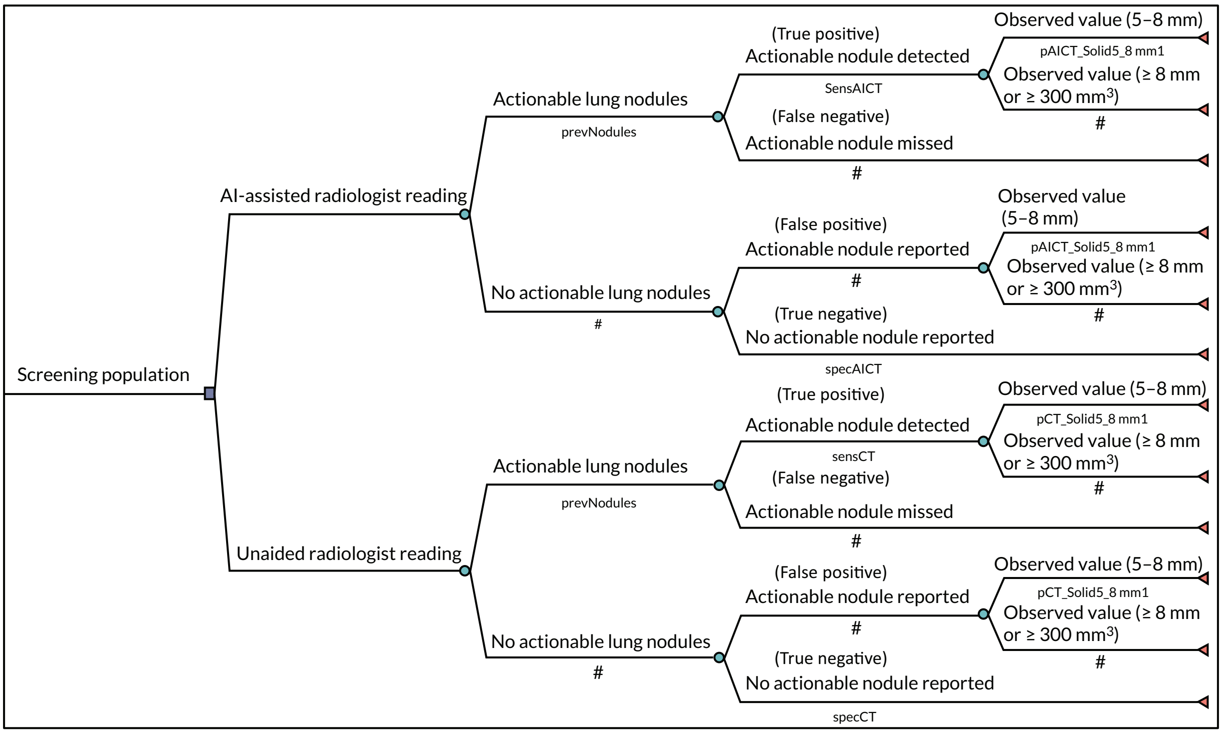


**FIGURE 7** Illustrative model structure for the detection of actionable lung nodules.

### *Unaided radiologist reading*

The strategy referred to as 'unaided radiologist reading' represents usual care/routine practice. Thus, it refers to the clinical pathway people would follow if undergoing a CT scan that includes part or all of the chest. Typically, all CT scans will be reviewed by a radiologist or a trained healthcare professional to identify lung nodules, their type and morphology and to measure the size of the lung nodule if one is present.

## Information required for the model

The model was populated with evidence identified from our test accuracy review and supplemented with information from secondary sources identified from additional searches. One major caveat in the use of evidence from our test accuracy review to inform this model arose from the mismatch between outcomes reported in the test accuracy studies, such as the sensitivity and specificity per actionable nodule detection as opposed to detection of a person with an actionable lung nodule.

### *Prevalence of actionable nodules*

The model required information about the prevalence of actionable nodules in each of our populations of interest. However, information was available only for the screening population. The prevalence of actionable nodules used in the model was 0.206 (95% CI 0.1786 to 0.2357), obtained from the UK LSUT, which was the largest UK study reporting this information.[77]

### *Test accuracy*

The model required information about the performance of AI-assisted radiologist reading and unaided radiologist reading to identify actionable lung nodules by population of interest. Comparative sensitivity and specificity were available from only one study conducted in a screening population,[54] reported in *Nodule detection* and summarised in *Table 9*.

**TABLE 9** Test accuracy estimates for identifying actionable nodules by test strategy

| Parameter | Value | 95% CI | Source |
|---|---|---|---|
| ***Screening population*** | | | |
| *AI-assisted radiologist reading* | | | |
| Sensitivity | 72.50 | 69.20 to 75.80 | Lo *et al.*[54] |
| Specificity | 84.40 | 82.40 to 86.40 | |
| *Unaided radiologist reading* | | | |
| Sensitivity | 60.10 | 56.80 to 63.40 | Lo *et al.*[54] |
| Specificity | 89.90 | 87.90 to 91.90 | |

CI, confidence interval.

We defined:

- true positive as actionable lung nodule detected
- true negative as actionable lung nodule neither present nor reported
- false positive as findings reported as actionable nodules (e.g. non-nodular structure incorrected identified as actionable nodules) that are in fact not actionable lung nodules
- false negative as actionable nodules that were not identified using each strategy.

### *Resource use and costs*

The resource use and costs included are those that are directly incurred by the NHS and Personal Social Services (PSS). Costs were required for the radiologist time, CT scan and software technologies. All costs are presented in 2021–2 prices.

#### Computer software

Costs per scan/output were obtained from the companies. For this analysis in the screening population, we used costs for ClearRead CT (Riverain technologies) as this was the AI software used in the study by Lo *et al.*,[54] which provided test accuracy data. Further details about our criteria used in the economic analysis are reported in *Resource use and costs*.

#### Time taken to read the computed tomography scan and report findings

For the detection of actionable lung nodules, we assumed that the costs incurred included CT scan and radiologist's time for reading and reporting CT scan image with/without the use of AI software assistance. We assumed that the procedure would be undertaken by a radiologist but used a band 9 radiographer as a proxy for costing purposes.[78]

Our test accuracy review found that the time taken to read/report CT scans reduced with AI-assistance in most studies [see *Radiologist reading time (10 studies)*]. However, these studies were predominantly conducted under research conditions, and there is uncertainty about how AI assistance may impact on read/reporting time in real clinical practice. Here, we used the median time of 10 minutes required for unaided radiologists to read and report a CT scan image reported in the UK LSUT[27] (assumed as mean value as the IQR of 5–15 was symmetrical around the median) and assumed that the time would be shorter for AI-assisted readers. In *Table 10* we present the time taken with AI-assisted and unaided reading by population. The longer times taken for reading and reporting a CT scan image for symptomatic and incidental population were based on clinical expert opinion, which suggests that more time may be needed to report other non-nodular findings in these patients, and the reading task is more susceptible to interruption than is analysing lung cancer screening images, which tends to be undertaken in batches during protected time. These alternative reporting times were not used here but were used in full model for the respective populations, described in *De novo cost-effectiveness analysis (full model): methods*.

*Table 11* provides a summary of the cost inputs used in the model.

## *Outcomes*

The outcome used in this analysis was correct identification of a person with an actionable nodule.

### Cost per correct identification of a person with actionable nodules

For this outcome, we assigned the value of 1 for people correctly identified with an actionable nodule (≥ 5 mm and no clear benign features) and 0 for all others. This was tallied to give the denominator for the ICER, expressed as cost per person with an actionable lung nodule detected.

## *Analysis*

The economic analysis was undertaken from the perspective of the NHS and PSS. A deterministic analysis was undertaken for the base case.

We undertook sensitivity and scenarios analyses. One-way sensitivity analysis was conducted to determine which input parameters were drivers of the economic analysis. Key input parameters were varied using the upper and lower values, and the findings showing the range of resulting ICERs for AI-assisted radiologist reading compared with unaided radiologist reading were presented in a tornado diagram.

**TABLE 10** Resource use associated with reporting CT scans

| | Population of interest | | |
|---|---|---|---|
| **Resource** | **Symptomatic** | **Incidental** | **Screening** |
| Radiologist time to report CT scan (AI-assisted) | 12 minutes | | 8 minutes |
| Radiologist time to report CT scan (unaided) | 15 minutes | | 10 minutes (Hall *et al.*[27]) |
| Type of CT scan at baseline | CT scan with contrast | | CT scan without contrast |
| Type of CT scan during surveillance, if required | CT scan without contrast | | |

**Note**
The 10 minutes for radiologist to report CT scan (unaided) for the screening population was based on Hall *et al.*[27] The estimated lengths of time for other readers/populations were based on expert advice as described in the main text.

**TABLE 11** Costs inputs used in the model

| Parameter | Value (£) | Source |
|---|---|---|
| ***Technologies (brand)*** | | |
| ClearRead CT (Riverain technologies) | 2.00 per scan/output | Supplied by the company |
| Radiologist consultation | 24.50 | PSSRU 2021 (cost per working hour (£147) for a band 9 radiographer as a proxy for a radiologist) (e.g. in the screening population, 10 minutes to report result) |
| Radiologist consultation (AI-assisted) | 19.60 | PSSRU 2021 (cost per working hour (£147) for a band 9 radiographer as a proxy for a radiologist) (e.g. in the screening population, 8 minutes to report result) |
| CT scan (single area, no contrast) | 106 | NHS reference schedule (RD20A – computerised tomography scan of one area, without contrast, 19 years and over) |
| CT scan (single area, pre- and post-contrast) | 145 | National schedule of NHS costs 2020–1 (RD22Z – CT scan of one area, with pre- and post-contrast) |

PSSRU, Personal Social Services Research Unit.

### Scenario analyses

We undertook several scenario analyses around the following model inputs:

- Prevalence of actionable lung nodules.
- Time taken to report CT scans. Given the uncertainty around this input parameter, which was obtained from clinical expert opinion, we explored in scenario analyses increasing or decreasing the reporting time with AI assistance and keeping the time the same as for unaided reading.

## Results

### *Deterministic results*

We present the deterministic result based on the outcome cost per correct identification of a person with an actionable nodule. Results are based on assuming a hypothetical cohort of 1000 people undergoing a CT scan.

### Cost per correct identification of a person with an actionable nodule

*Table 12* presents the estimates of costs and additional people correctly identified with an actionable nodule with the use of AI-assisted radiologist reading compared with unaided radiologist reading in a screening population. These results show that AI-assisted radiologist reading (ClearRead CT) is approximately £2900 cheaper and expected to correctly identify an additional 25.5 people with actionable nodules per 1000 CT screens, thereby dominating the unaided reading strategy.

### *Sensitivity analysis results*

Deterministic sensitivity analysis was conducted by varying key model input parameters by their ranges or, when these were unavailable, by assuming ± 10% (cost of CT scan) and ± 50% (time taken to read and report results) to assess the impact on the ICER (cost per correct identification of people with an actionable lung nodule), with the results presented in the form of tornado diagrams. Findings of the sensitivity analysis for the preliminary model are presented in *Figure 8*.

Sensitivity analysis results showed that the time taken to read and report image analysis findings were the key drivers of cost-effectiveness for the comparison of AI-assisted radiologist reading with unaided radiologist reading for identifying actionable lung nodules. However, varying these inputs within these limits is unlikely to change the ICERs outside acceptable thresholds.

### *Scenario analysis results*

Based on the alternative sources of evidence or assumptions made on key parameters, these results (*Table 13*) were robust to changes made.

**TABLE 12** Deterministic results based on expected costs and expected identification of people with actionable lung nodules (screening population of 1000 people undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with actionable nodules correctly identified | Incremental number of people with actionable nodules correctly identified | ICER (£) per correct identification of an individual with actionable lung nodules |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (ClearRead CT) | 127,600 | – | 149.3 | – | – |
| Unaided radiologist reading | 130,500 | 2900 | 123.8 | −25.5 | Dominated |

**Note**
Exact results were obtained from TreeAge but were rounded by the authors and presented.

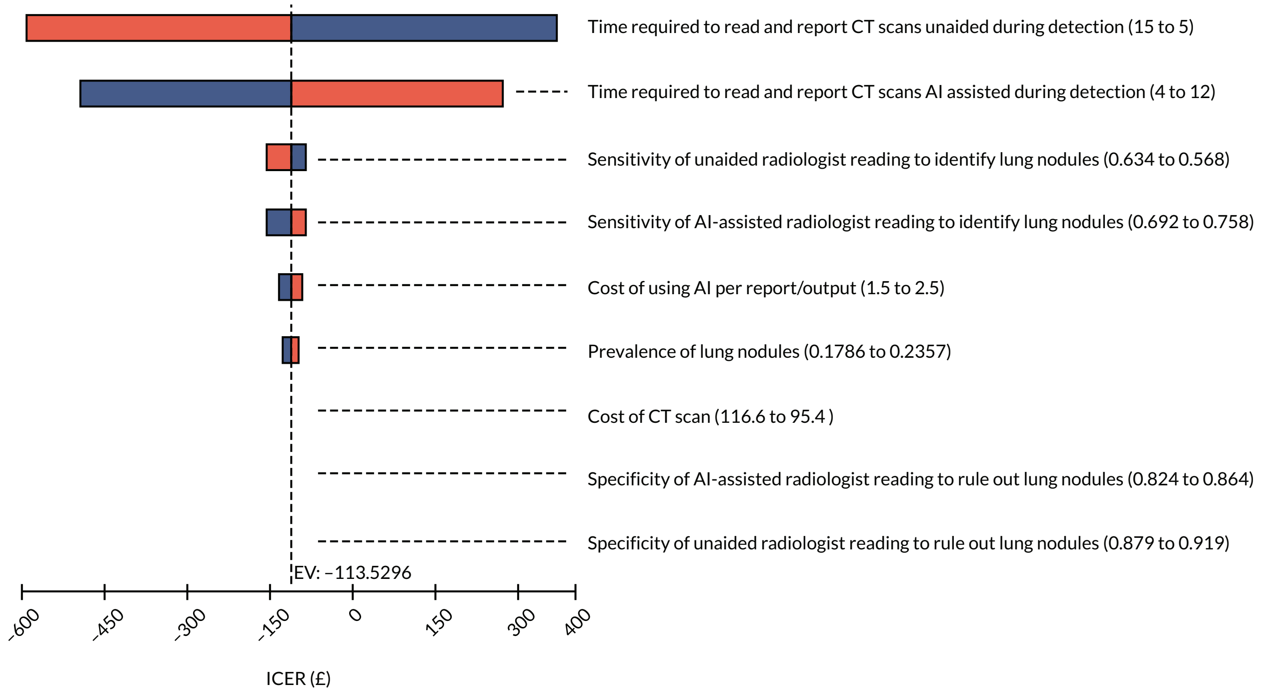


**FIGURE 8** Tornado diagram of the impact to the cost per actionable lung nodule correctly identified by changing individual parameters (screening population). Note: the ICERs shown were for AI-assisted radiologist reading compared with unaided radiologist reading.

**TABLE 13** Scenario analysis results based on cost per person with an actionable lung nodule correctly identified (screening population)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with actionable lung nodules | Incremental number of people with actionable lung nodules | ICER (£) per person with actionable lung nodules |
|---|---|---|---|---|---|
| ***Base-case*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 127,600 | – | 149.3 | – | – |
| Unaided radiologist reading | 130,500 | 2900 | 123.8 | –25.5 | Dominated |
| ***Prevalence of detecting actionable lung nodules from 0.206 to 0.2823 (estimate reported in another NELSON lung cancer screening trial)[4]*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 127,600 | – | 204.7 | – | – |
| Unaided radiologist reading | 130,500 | 2900 | 169.7 | –35 | Dominated |
| ***Time taken to read and report CT scans: assumed to be 10 minutes for both AI-assisted and unaided image analysis*** | | | | | |
| Unaided radiologist reading | 130,500 | – | 123.8 | – | – |
| AI-assisted radiologist reading (ClearRead CT) | 132,500 | 2000 | 149.3 | 25 | 78 |
| ***Time taken to read and report CT scans: assumed to be 10 minutes for AI-assisted and 8 minutes for unaided image analysis*** | | | | | |
| Unaided radiologist reading | 125,600 | – | 123.8 | – | – |
| AI-assisted radiologist reading (ClearRead CT) | 132,500 | 6900 | 149.3 | 25.5 | 270 |

### *Discussion*

The preliminary model provides a relatively straightforward approach to assessing the cost-effectiveness of AI-assisted detection and analysis of lung nodules for chest CT scan images. However, a major limitation of this simpler approach is that the test accuracy evidence related to the detection of actionable nodules is available only from per-nodule analysis, which is less suitable than test accuracy obtained from per-person analysis as the unit for decision analysis is individual persons, not nodules. In addition, this analysis only covers initial nodule detection and does not allow an evaluation of the impact of AI assistance on subsequent nodule management through analysis of surveillance CT scans. Consequently, we developed a more comprehensive decision-analytic structure, which started from the initial identification of any lung nodules, for which test accuracy data from per-person analysis were available from both screening and symptomatic populations. To bridge the gap between evidence on initial nodule detection and subsequent management according to BTS guidelines, as well as the link to health outcomes, the EAG conducted additional simulations. Further details are provided in the next chapter.

# Chapter 7 De novo cost-effectiveness analysis (full model): methods

## Developing the model structure

Given the limitations of the preliminary model mentioned in *Chapter 6, Discussion*, we developed a full economic model to assess the cost-effectiveness of using software with AI-derived algorithms for the automated detection and analysis of lung nodules from CT images compared with unassisted CT image analysis in people undergoing initial CT scans from symptomatic, incidental and screening populations. The main model structure was similar for all three populations, but the model parameters varied depending on the specific population where appropriate. Further details of the population are in *Prevalence of lung nodules*. For people undergoing CT surveillance for previously detected nodules, the surveillance component of the model can be used.

The decision model follows the illustrative pathways shown in *Figure 9*. After people undergo a CT scan that may identify lung nodules, the CT scan image is read by either human reader alone or human reader with software assistance. We used a two-stage approach to the decision model structure. The first stage consists of identifying lung nodules and their type and size in accordance with the BTS guidelines, and we used a decision tree structure. We considered this appropriate as it would capture all the short-term costs and events associated with identifying and analysing lung nodules. The branches of the decision tree represent the strategies under assessment and were populated with appropriate information (see *Information required for the model*). In the second stage, we continued/extended the decision tree structure for the evaluation to capture CT surveillance, the natural history of malignant lung nodules and treatment to capture CT surveillance, the growth of malignant nodules and the treatment of people with cancer.

## Strategies

The model compares AI-assisted radiologist reading with unaided radiologist reading.

### *Unaided radiologist reading*

The strategy referred to as 'unaided radiologist reading' represents usual care/routine practice. Thus, it refers to the clinical pathway people would follow if they underwent a CT scan including part or all of the chest. Typically, all CT scans will be reviewed by a radiologist or a trained healthcare professional to identify lung nodules, their type and morphology and measure the size of their lung nodule if present.

### *Artificial intelligence-assisted radiologist reading*

The alternative strategy is AI-assisted radiologist reading. In this strategy, the software uses algorithms that have been produced using AI. AI is used to assist the radiologist or the healthcare professional to identify lung nodules, as well as their morphology and size.

### *Pathway of people in the two strategies*

The pathway for both strategies is the same in the three populations (*Figure 10*). In people identified as having a lung nodule, the nodule will be further assessed for its type (e.g. solid or subsolid) as well as its size. In the model, we assumed that if at least one lung nodule is detected, the individual would have one primary lung nodule (usually the largest nodule according to the BTS guidelines;[12] also called 'risk dominant nodule'). The primary nodule would be measured by a radiologist (or other trained professionals) with/without the assistance of AI software and categorised as follows: solid (< 5 mm, 5 mm to < 8 mm, and ≥ 8 mm) or subsolid (< 5 mm and ≥ 5 mm). For people with a lung nodule that was missed on (reading of) CT scan, we assumed that these nodules could be undiagnosed as benign or malignant. People without a lung nodule who have been correctly identified as such are discharged.

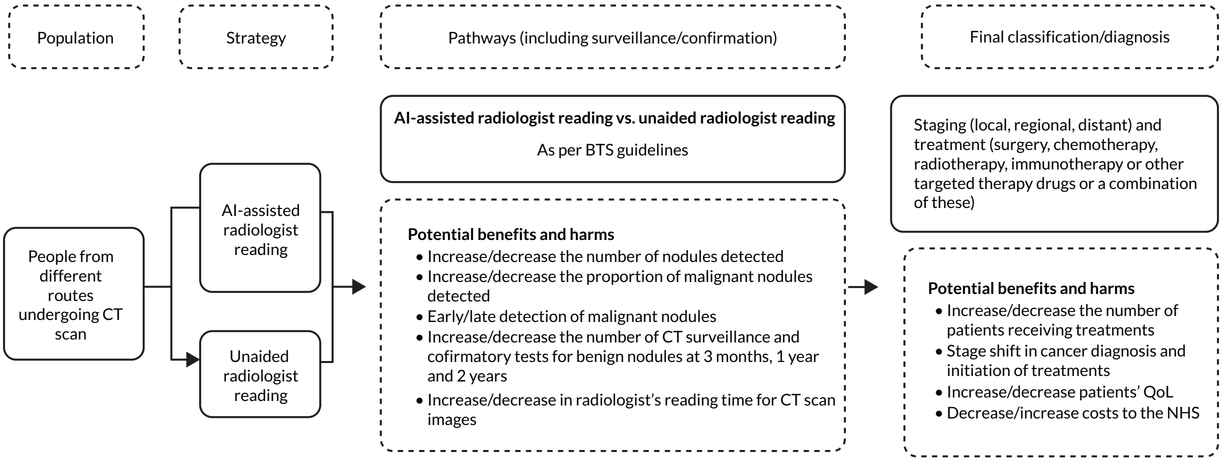

Population
Strategy
Pathways (including surveillance/confirmation)
Final classification/diagnosis
People from different routes undergoing CT scan
AI-assisted radiologist reading
Unaided radiologist reading
AI-assisted radiologist reading vs. unaided radiologist reading
As per BTS guidelines
Potential benefits and harms
• Increase/decrease the number of nodules detected
• Increase/decrease the proportion of malignant nodules detected
• Early/late detection of malignant nodules
• Increase/decrease the number of CT surveillance and cofirmatory tests for benign nodules at 3 months, 1 year and 2 years
• Increase/decrease in radiologist's reading time for CT scan images
Staging (local, regional, distant) and treatment (surgery, chemotherapy, radiotherapy, immunotherapy or other targeted therapy drugs or a combination of these)
Potential benefits and harms
• Increase/decrease the number of patients receiving treatments
• Stage shift in cancer diagnosis and initiation of treatments
• Increase/decrease patients' QoL
• Decrease/increase costs to the NHS

**FIGURE 9** Illustrative structure of the clinical pathways.

## Natural history

Our natural history model was developed to model the growth/disease progression of malignant disease, separately for solid nodules and subsolid nodules. We assumed that benign nodules did not grow following detection. The progression of lung cancer is characterised by its growth in malignant lung nodules. We assumed that the growth of tumours follows a Gompertz distribution and is conditional on VDT (i.e. the time required for the tumour to double its volume),[79] which is based on information obtained from Treskova *et al.*[71] Details of our nodule growth model and its development can be found in *Appendix 7*, *Table 65* and text.

## Information required for the model

The model was populated with information obtained from evidence identified from our test accuracy and cost-effectiveness reviews and supplemented with information from secondary sources identified from additional searches (see *Appendix 3*) as well as clinical expert opinion. One major challenge in using evidence from our test accuracy review to inform the decision-analytic model arose from the mismatch between outcomes reported in the test accuracy studies (e.g. sensitivity and specificity for detecting nodules of various sizes and types and concordance of measuring nodule size/volume) and data required to parametrise the model based on the specific BTS categorisation of the primary nodule (*Figure 1*). To translate the evidence reported in test accuracy studies into the BTS categorisation (< 5 mm, ≥ 5 and < 8 mm, and ≥ 8 mm for solid nodules; < 5 and ≥ 5 mm for subsolid nodules), which dictates subsequent clinical management (e.g. discharge, further CT surveillance, further clinical work-up and treatment), the EAG carried out simulations to bridge this disconnection in evidence. The rationale, approaches and assumptions of the simulation are described in the following section.

### *External Assessment Group simulation of measurement accuracy and precision*

Briefly, the simulations take the following initial inputs obtained from test accuracy review and additional evidence sources:

- Proportion of solid and subsolid nodules among identified primary nodules – this differs between populations of interest.
- The 'true' mean sizes of the primary nodules – these differ between populations of interest and between solid and subsolid nodules.
- The measurement precision (random errors in measurements, captured in measures of variation, such as standard deviations) – this may differ between unaided and AI-aided readings, with higher precision or better consistency being one of the purported advantages for AI-aided reading.
- The measurement accuracy (systematic error in measurements, e.g. consistently over- or underestimating the 'true' nodule size) – this may differ between unaided and AI-aided reading.

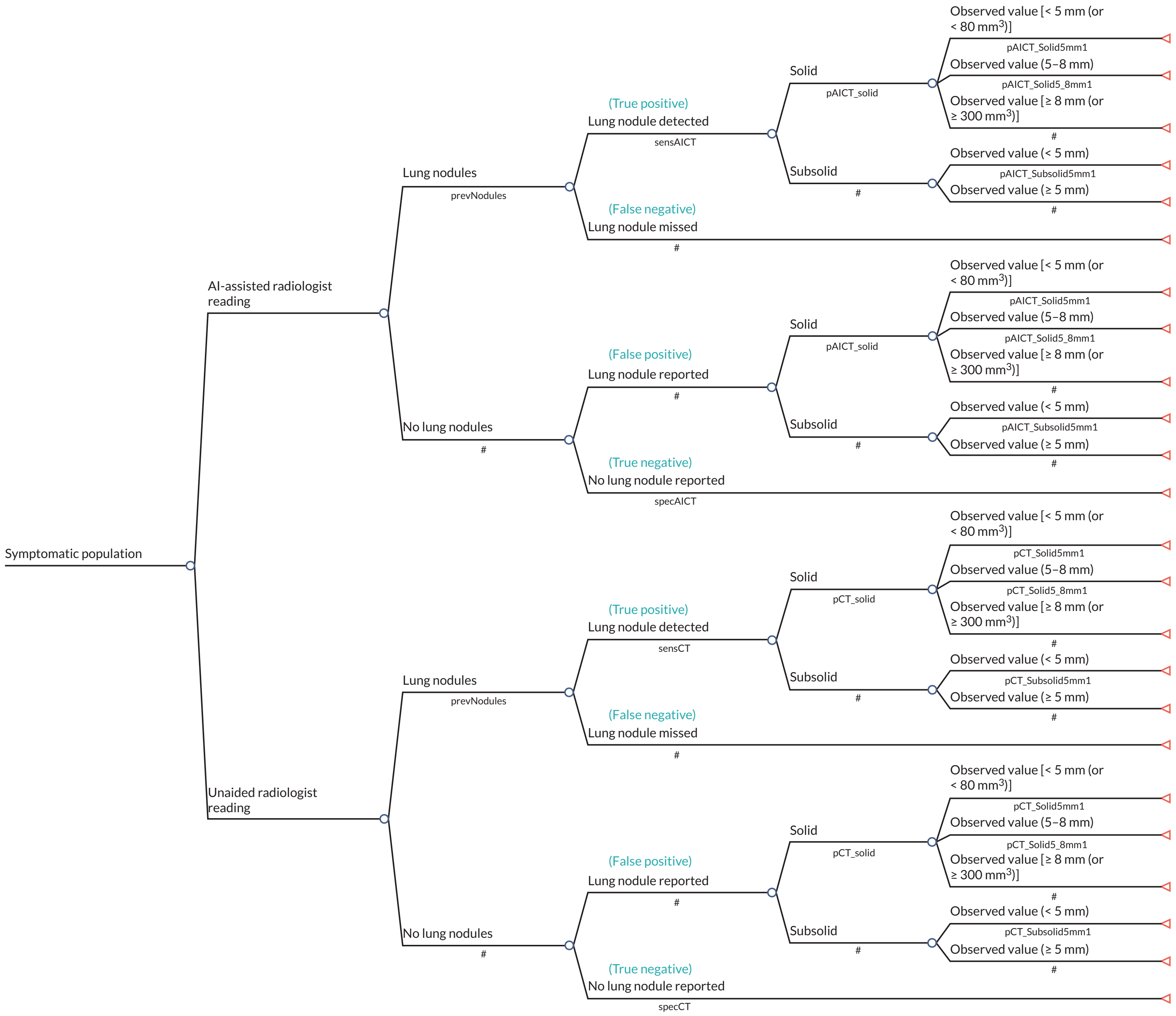


**FIGURE 10** Illustrative model structure for the detection of lung nodules.

The simulation models then generate distributions of (1) true nodule sizes, (2) nodule sizes based on AI reading alone, (3) nodule sizes based on AI-assisted radiologist reading, (4) nodule sizes based on unaided radiologist reading, separately for solid and subsolid nodules. By applying BTS categorisation, the proportion of nodules/patients falling into each BTS category based on 'true' nodule sizes, AI reading alone, AI-assisted reading and unaided reading can be estimated. A comparison of results between (1) and each of (2), (3) and (4) provides information concerning the miscategorisation of nodules arising from random and systematic measurement errors for AI reading alone, AI-assisted radiologist reading and unaided radiologist reading, respectively. Differences between AI-assisted reading and unaided reading, which is the main comparison of interest, can then be derived.

Detailed methods of the simulation are presented in *Appendix 8*, *Tables 66–69* and text, and in *Report Supplementary Material 4*.

For the decision-analytic model, information was required about the prevalence of lung nodules, the type of the lung nodules, the prevalence of lung cancer based on size and type of lung nodules, and the performance of AI-assisted radiologist reading and unaided radiologist reading for identifying and measuring lung nodules during the initial scan and subsequent surveillance, all by population of interest. *Figures 21–23* in *Appendix 1* provide an overview of the model parameters used and the sources of these data. Further information is detailed in the following sections.

### *Prevalence of lung nodules*

The model required data on the prevalence of lung nodules for three out of the four populations of interest. We assumed that the prevalence of lung nodules would vary across these populations. However, prevalence data was not needed for individuals undergoing surveillance, as they would, by definition, already have a previously detected nodule. *Table 14* presents the prevalence information. While individuals may have more than one lung nodule, we assumed that clinical management, following BTS guidelines, would be guided by a primary lung nodule.

### *Type of lung nodule*

The model also required information about the type of the primary lung nodule identified. In the model we categorised nodules as solid or subsolid, in line with the BTS guidelines.[12] Here we assumed that, if a nodule was identified, then it would be correctly categorised as solid or subsolid. We required the proportion of lung nodules by type and by reason for undergoing a CT scan. In *Table 15*, we report the proportions of each type of lung nodules for the symptomatic and screening populations. For the incidental population we used the same figures as for the screening population.

### *Prevalence of lung cancer based on size of lung nodule*

Following the measurement of the primary nodule and excluding/discharging people with nodules that had clear benign features (assumed 10% in each size band), the model required information about the prevalence of nodules that were malignant by size and by reason for undergoing CT scan (*Table 16*). The information was derived from the publication by Horeweg *et al.*[4] Their study is based on 7155 Dutch participants in the screening group of the NELSON trial. Lung

**TABLE 14** Prevalence of having at least one lung nodule by population of interest

| Population | Prevalence (95% CI) | Source | Justification |
|---|---|---|---|
| People with symptoms suggestive of lung cancer | 0.949 (0.8928 to 0.9763) | Kozuka *et al.* 2020[59] | Only study identified |
| Incidental (CT scan done for other reasons) | 0.13 (0.02 to 0.24)[a] | Callister *et al.* 2015[12] | Evidence review for 2015 BTS guidelines |
| Lung cancer screening | 0.509 (0.4868 to 0.5312) | Field *et al.* 2016[80] | Largest UK-based study that reported prevalence of any nodules |
| CT surveillance of a previously detected nodule[b] | Not applicable | – | – |

a Range.
b Not applicable because all of the people in the model would have an indeterminate lung nodule.

**TABLE 15** Proportion of detected risk-dominant nodules that are solid/subsolid

| Type of nodule | Proportion | Source |
|---|---|---|
| ***Radiologist read CT scan with software assistance and radiologist-read CT scan alone*** | | |
| *Symptomatic population* | | |
| Solid | 0.774 | Kozuka *et al.*,[59] table 1<br>518 solid nodules, 151 subsolid nodules |
| Subsolid | 0.226 | |
| *Screening population* | | |
| Solid | 0.939 | Hwang *et al.*,[50] table S3; 4357 solid nodules, 285 subsolid nodules |
| Subsolid | 0.061 | |

**Note**
The relative proportions are assumed to be the same for true positives (correctly identified nodules), false negatives (nodules missed by CT scan/reading) and false positives (non-nodular structures incorrected identified as nodules).

**TABLE 16** Prevalence of lung cancer in detected nodules, by population and nodule measurement

| Lung nodule baseline measurement | Population, prevalence and source | | | |
|---|---|---|---|---|
| | Symptomatic | Incidental | Screening | Surveillance |
| ***Solid*** | | | | |
| 5 to < 6 mm | Assumed same as screening | Assumed same as screening | 0.0089 (0.005, 0.016) (Horeweg *et al.* 2014[4]) | Assumed same as screening |
| 6 to 8 mm | Assumed same as screening | Assumed same as screening | 0.011 (Horeweg *et al.* 2014[4]) | Assumed same as screening |
| ≥ 8 mm | Assumed same as screening | Assumed same as screening | 0.094 (Horeweg *et al.* 2014[4]) | Assumed same as screening |
| ***Subsolid*** | | | | |
| ≥ 5 mm | Assumed same as screening | Assumed same as screening | 0.036 (Horeweg *et al.* 2014[4]) | Assumed same as screening |

cancer probability of screen-detected non-calcified nodules was reported by volume and volume-based diameter. Despite the lung cancer probability not being reported separately for solid and subsolid nodules, we chose this study as model input as the population was rated as most applicable to a UK screening population.

### *Test accuracy*

The model required information about the performance of radiologist-read CT scan with software assistance and radiologist-read CT scans to identify lung nodules by population. We used information about sensitivity and specificity as performance measures of these strategies for identifying any lung nodule. Sensitivity was defined as the probability of radiologist-read CT scan with/without software assistance to correctly identify an individual with a lung nodule (see *Lung nodules and lung cancer* for our definition of a lung nodule). Specificity was defined as the probability of the radiologist-read CT scan with/without software assistance to correctly identify individuals without a lung nodule. No attempt was made to derive the sensitivity and specificity of these strategies to identify people with malignant/benign nodules.

Three studies[53,59,61] were identified that reported these outcomes. Their characteristics, strengths and limitations are reported in *Table 17*. The study by Zhang *et al.*[61] was immediately discounted as it compared double reading with software use under laboratory conditions with double reading by different readers without software use in clinical practice.

From the remaining two studies,[53,59] we chose the study by Kozuka *et al.*[59] as the cost-effectiveness analysis input for the symptomatic population as this was the only identified study actually undertaken in patients suspected of having lung cancer. We also used Kozuka *et al.*[59] as the input for the incidental population as the readers were less experienced radiologists, who were judged to be similar to general radiologists in UK practice assessing CT images in accident and emergency. For the screening population, we decided to use the senior group (experienced chest radiologists) from the study by Hsu *et al.*[53] as this study reported separate accuracy results for the screening LDCT images, and the experience and speciality of the readers was most applicable to a UK screening programme. *Table 18* summarises test accuracy estimates for identifying any lung nodules for various populations included in the model.

**TABLE 17** Comparative studies reporting detection accuracy for any nodules that could be used as cost-effectiveness analysis model inputs and their advantages and disadvantages (three studies)

| Study | Study details | Sensitivity (per subject) | Specificity (per subject) | Advantages | Disadvantages |
|---|---|---|---|---|---|
| Hsu *et al.* 2021[53] | Mixed population: one hospital in Taiwan; 150 consecutive cases with lung nodules ≤ 1 cm or no nodules: 93 clinical routine; 57 screening population<br>Low dose (*n* = 57), standard dose (*n* = 93), no contrast, slice thickness 2.5 mm<br>MRMC study, ClearReadCT with vessel suppression and nodule detection: six chest radiologists – three less experienced (residents in radiology with > 6 months of chest CT experience) and three experienced chest radiologists (5, 10 and 25 years of experience)<br>Reference standard: consensus expert reading (two readers) | Per-nodule sensitivity (340 nodules)<br>[D] Mean 64% (95% CI 62% to 66%); [C] mean 80% (95% CI 81% to 85%) ($p < 0.001$)<br>Senior readers only<br>[D] Mean 74% (95% CI 72% to 77%); [C] mean 84% (95% CI 82% to 86%) ($p < 0.001$) | 52 patients without nodules<br>[D] Mean 80% (95% CI 78% to 81%); [C] mean 83% (95% CI 82% to 85%) ($p = 0.256$)<br>Senior readers only<br>[D] Mean 87% (95% CI 85% to 89%); [C] mean 88% (95% CI 87% to 90%) ($p = 0.729$) | Consecutive sampling; mixed population but separate data for screening population reported; MRMC study included six readers and reports accuracy separately for three experienced (senior) chest radiologists (high applicability for UK screening and symptomatic populations) | Taiwan, one hospital (not a UK or north-western European population, nodule prevalence might be different)<br>57 screening LDCT images (small sample size)<br>Lung nodules ≤ 1 cm only (inclusion of only small nodules might affect sensitivity); 2.5 mm slice thickness (UK ≤ 2 mm, might affect accuracy)<br>MRMC study (radiologist performance under laboratory conditions might be not representative of clinical practice)<br>No subject-level sensitivity reported, only per-nodule sensitivity (per-subject sensitivity might be higher)<br>Only reported mean sensitivity and mean specificity, no 2 × 2 data, no data for individual readers (no decimal places reported, cannot calculate exact estimates) |
| Kozuka *et al.* 2020[59] | Symptomatic population (suspected lung cancer): random 120 chest CT images from one hospital in Japan<br>Standard dose; no contrast; 1 mm slice thickness<br>MRMC study, InferRead CT Lung (Infervision); two less experienced radiologists (1 and 5 years of diagnostic experience); reference standard: consensus expert reading (three readers) | 111 subjects with nodules, pooled reader A + reader B<br>[D] 68.0% (151/222) (95% CI 61.4% to 74.1%); [C] 85.1% (189/222) (95% CI 79.8% to 89.5%) ($p < 0.001$) | Six subjects without nodules, pooled reader A + reader B<br>[D] 91.7% (11/12) (95% CI 61.5% to 99.8%); [C] 83.3% (10/12) (95% CI 51.6% to 97.9%) (no level of significance reported) | Only study on symptomatic population; random selection; 1 mm slice thickness (applicable to the UK); reported 2 × 2 data individually for reader A and reader B | Japan, one hospital (not a UK or north-western European population, so nodule prevalence might be different); 117 CT images included in analyses (small sample size); MRMC study (radiologist performance under laboratory conditions might be not representative of clinical practice); two less experienced radiologists (1 year and 5 years of experience) (applicability concerns to UK reading practice for symptomatic population)<br>Only six CT images without nodules (wide 95% CI for specificity; one additional FP case in one reader resulted in an apparently big difference in pooled point estimates) |

continued

**TABLE 17** Comparative studies reporting detection accuracy for any nodules that could be used as cost-effectiveness analysis model inputs and their advantages and disadvantages (three studies) (*continued*)

| Study | Study details | Sensitivity (per subject) | Specificity (per subject) | Advantages | Disadvantages |
|---|---|---|---|---|---|
| Zhang *et al.* 2021[61] | Screening population: 860 consecutive patients from one hospital in China (part of NELCIN-B3 project); low dose; no contrast; 0.625–1.0 mm; InferRead CT Lung (Infervision); one radiology resident with supervision of one experienced radiologist – with software (MRMC study); without software (clinical practice); reference standard: consensus expert reading (two readers) | [E] 43.3% (162/374); [C] 98.9% (370/374) (no level of significance reported) | [E] 100.0% (486/486); [C] 97.1% (472/486) (no level of significance reported) | Consecutive screening population; 860 patients included: 374 with nodules and 486 without nodules (quite big sample size) | China, one hospital (not a UK or north-western European population, nodule prevalence might be different); Different readers with and without software use:<br>[C] Performance of one resident and one radiologist only; [E] 14 different residents and 15 different radiologists<br>Unaided reading performed in clinical practice, whereas aided reading as part of MRCM study; not single reading, but reading by a radiology resident with supervision by experienced radiologist (applicability concerns to UK practice) |

[C] Concurrent AI; [D] unaided reading (MRMC study); [E] unaided reading (clinical practice); FP, false positive.

**TABLE 18** Test accuracy estimates to identify any lung nodule by reason for undergoing CT scan

| Parameter | Value | 95% CI | Source |
|---|---|---|---|
| ***People with symptoms suggestive of lung cancer*** | | | |
| *AI-assisted radiologist reading* | | | |
| Sensitivity | 85.14 | 79.80 to 89.50 | Kozuka *et al.* 2020[59] |
| Specificity | 83.33 | 51.60 to 97.90 | |
| *Unaided radiologist reading* | | | |
| Sensitivity | 68.02 | 61.40 to 74.10 | Kozuka *et al.* 2020[59] |
| Specificity | 91.67 | 61.55 to 99.88 | |
| ***Incidental (CT scan done for other reasons)*** | | | |
| *AI-assisted radiologist reading* | | | |
| Sensitivity | 85.14 | 79.80 to 89.50 | Kozuka *et al.* 2020[59] |
| Specificity | 83.33 | 51.60 to 97.90 | |
| *Unaided radiologist reading* | | | |
| Sensitivity | 68.02 | 61.40 to 74.10 | Kozuka *et al.* 2020[59] |
| Specificity | 91.67 | 61.55 to 99.88 | |
| ***Screening*** | | | |
| *AI-assisted radiologist reading* | | | |
| Sensitivity | 83.00 | 79.00 to 86.00 | Hsu *et al.* 2021[53] |
| Specificity | 88.00 | 85.00 to 91.00 | |
| *Unaided radiologist reading* | | | |
| Sensitivity | 73.00 | 69.00 to 77.00 | Hsu *et al.* 2021[53] |
| Specificity | 86.00 | 83.00 to 90.00 | |

We extracted information from individual studies identified from our test accuracy systematic review to populate 2 × 2 tables to derive study-specific test performance for both strategies. We used the following definitions:

- True positive: any lung nodule present.
- True negative: no lung nodule present.
- False positive (during detection of lung nodules): findings that are not lung nodules (non-nodular structure incorrectly identified as nodules).
- False negative: nodules that were not identified/missed using each strategy. We assumed that there would be lung nodules that were not identified at initial CT scan but later diagnosed. Here we assumed that these lung nodules were initially present but undetected and thus were not new lung nodules.

Additionally, we required information about the performance of these strategies during the surveillance of people with lung nodules to identify nodules that are/are not growing.

## *Effectiveness*

### Stage shift

In the model, we attempt to quantify the expected benefit with the use of AI assistance in terms of achieving an earlier diagnosis, as a person's prognosis is likely to be better if they are diagnosed at an earlier stage, hence improving their chances of long-term survival. The likely source of delay in diagnosis is 'watchful waiting', when people are referred to receive CT surveillance. During surveillance, people undergo imaging aimed at measuring the growth of lung nodules, which is characterised by its VDT. If the VDT is below a specified threshold at a specified time point, then lung nodules are likely to be malignant. People with lung nodules above this threshold may be referred to further surveillance or discharged.

## *Resource use and costs*

The resource use and costs included are those directly incurred by the NHS and PSS. Costs were required for the radiologist time, CT scan, software technologies, and treatment associated with lung cancer. All costs are presented in 2021–2 prices, and, after the first year, both costs and benefits were discounted at a rate of 3.5% per annum.  Additionally, identified costs  through literature review were adjusted to current prices, where necessary, using the Hospital and Community Health Services (HCHS) index from the Unit Costs of Health and Social Care 2021.

### Computer software

There is paucity of test accuracy and cost data for some of the technologies included in the final scope of this assessment. To avoid generating cost-effectiveness estimates for technologies for which no technology-specific data can be used in the model, we included only technologies that met both of the following criteria in our base case:

- The cost information for the technology should be supplied by the company or be publicly available.
- Test accuracy information related to the technology that could be used to inform at least one of the model input parameters (e.g. performance for identifying lung nodules or precision of lung nodule measurements) is available, either supplied by the company or accessible through publication.

In *Table 19*, we outline how each company's technology listed in the NICE scope performed against these criteria. Of the 13 relevant technologies identified by NICE, useful test accuracy information (e.g. sensitivity and specificity for identifying any lung nodules) was available for two companies; hence, these were considered in the economic analysis. For the screening and the incidental populations, we included the ClearRead CT (Riverain) technology in the economic analyses, and for the symptomatic population, we included InferRead CT Lung (Infervision) technology. It was noted that different costing structures were in place, so attempts were made to obtain/derive a per-scan cost.

For detection of lung nodules, we assumed that the costs incurred included CT scan, radiologist consultation and use of software assistance. We assumed that the procedure would be undertaken by a radiologist, taking 10 minutes, but used a band 9 radiographer as a proxy.

During surveillance of people with lung nodules or people suspected of having lung nodules, we assumed that additional costs would be incurred (visit to MDT, further CT scans and biopsy).

### Treatment costs

Total treatment costs by stage of disease were obtained from Bajre *et al.*[74] and were originally from Cancer Research UK 2014.[81] Total costs included retreatment costs and were reported in the price year 2014–5. These costs were obtained from the literature and uprated to current prices (2020–1) using the Hospital and Community Health Services (HCHS) index from *Unit Costs of Health and Social Care 2021*.[78] Cost inputs used in the model are reported in *Table 20*.

**TABLE 19** Technologies outlined in scope against our selection criteria for the base-case economic analysis

| | Criteria | | |
|---|---|---|---|
| Technology (company) | Cost information | Comparative data on nodule detection accuracy available | Software measurement accuracy or concordance with manual measurement data available |
| AI-Rad Companion (Siemens Healthineers) | Not available | No | Yes (concordance, mixed population[47]) |
| AVIEW LCS+ (Coreline Soft) | Not available | No | No |
| ClearRead CT (Riverain Technologies) | Yes | Yes | Yes (accuracy, screening population[56] and unclear indication[55])<br>Yes (concordance, mixed population[58]) |
| Contextflow SEARCH Lung CT (contextflow) | Yes | No | No |
| InferRead CT Lung (Infervision) | Yes | Yes | No |
| JLD-01K (JLK Inc.) | No | No | No |
| Lung AI (Arterys) | Not available | No | No |
| Lung Nodule AI (Fujifilm) | Not available | No | No |
| qCT-Lung (Qure.ai) | Not available | No | No |
| SenseCare-Lung Pro (SenseTime) | Not available | No | No |
| Veolity (MeVis) | Not available | No | Yes (concordance, surveillance population[63]) |
| Veye Lung Nodules (Aidence) | Yes | No | Yes (accuracy, mixed populations[33,66])<br>Yes (concordance, mixed population[33]) |
| VUNO Med-LungCT AI (VUNO) | Not available | No | No |

**TABLE 20** Costs inputs used in the model

| Parameter | Value (£) | Source |
|---|---|---|
| ***Technologies (brand)*** | | |
| ClearRead CT (Riverain) | 2.00 per scan/output | Supplied by the company |
| InferRead CT Lung (Infervision) | 3.34 per scan/output | Supplied by the company |
| Radiologist consultation | 24.50 | PSSRU 2021 (cost per working hour (£147) for a band 9 radiographer as a proxy for a radiologist) (e.g. in the screening population, 10 minutes to report result) |

**TABLE 20** Costs inputs used in the model (*continued*)

| Parameter | Value (£) | Source |
|---|---|---|
| Radiologist consultation (AI-assisted) | 19.60 | PSSRU 2021 (cost per working hour (£147) for a band 9 radiographer as a proxy for a radiologist) (e.g. in the screening population, 8 minutes to report result) |
| CT scan (single area, no contrast) | 106 | NHS reference schedule (RD20A – computerised tomography scan of one area, without contrast, 19 years and over) |
| CT scan (single area, pre and post contrast) | 143 | National schedule of NHS costs 2020–1 (RD22Z – CT scan of one area, with pre and post contrast) |
| MDT | 146 | National schedule of NHS costs 2020–1 (CDMT_OTH other cancer MDT meetings) |
| Guided needle biopsy | 1670 | NHS reference schedule (DZ71Z – minor thoracic procedure, guided needle biopsy) |
| PET scan | 1161 | RN01a – PET-CT of one area, 19 years and over |
| ***Treatment*** | | |
| Stage I | 18,705 | Bajre *et al.* 2017[74] |
| Stage II | 21,312 | |
| Stage III | 23,922 | |
| Stage IV | 14,909 | |

PET, positron emission tomography; PSSRU, Personal Social Services Research Unit.

### *Utility values*

The utility values used to derive the QALYs for people with lung cancer were mainly obtained from Bajre *et al.*[74] and were originally obtained from Naik *et al.*[82] Briefly, these authors collected health-related quality-of-life information using the EuroQol-5 Dimensions (EQ-5D) questionnaire from 1760 Canadian ambulatory cancer patients and reported utility values by stage at diagnosis. Among the participants with lung cancer ($n$ = 128), those with stage I, II, III and IV diagnoses had utility estimates of 0.81, 0.77, 0.76 and 0.76, respectively. For people without a lung nodule, we assigned a utility value of 0.855.[83]

In the base case, we assumed that there is a −0.063 disutility for people with a non-nodular structure incorrectly identified as a nodule (false positive during detection of a lung nodule). In the model, we assumed that these non-nodular structures will be discharged at the first CT surveillance (i.e. at 3 months or 1 year). In addition, we assumed that for people under CT surveillance with lung nodules that were later diagnosed as benign, there would be a disutility of −0.063 lasting until the person was discharged. People without lung nodules and those with benign nodules were assumed to have utility values representing age-/sex-adjusted UK-specific general population norms.[83]

We assumed a disutility of −0.2 associated with undergoing a biopsy with a duration of 3 months.

### *Mortality*

Two types of mortality were considered in the model: lung cancer death and death from other causes. Survival following treatment of lung cancer was obtained from secondary sources. General population mortality of people without lung cancer was obtained from the Office for National Statistics and an average of the mortality rate for male and female individuals was used in the model. We assumed that all-cause mortality would not differ between the two strategies or by reason for requiring CT scan. We included a 1.3 increased risk of death due to the smoking status of our population,[84] but we did not apply any increase to mortality for individuals with benign lung nodules.

### *Outcomes*

Three different outcome/effectiveness measures were used in the analysis: correct identification of actionable nodules, cancer correctly detected and treated, and QALYs.

### Cost per correct identification of actionable nodules

For this outcome, we assigned the value of 1 for people correctly identified with actionable nodules (≥ 5 mm and no clear features of being benign), and 0 for all others.

### Cost per cancer correctly detected and treated

No effectiveness information was required. We reserved the value of 1 for people with cancer correctly detected and then calculated the difference between strategies.

### Cost per quality-adjusted life-year

Four sets of QALY values were estimated for use in the model: first, for people who do not have any lung nodules; second, for people with benign lung nodules; third, for people treated for lung cancer; and fourth, for people who have undiagnosed lung cancer.

## *Model assumptions*

We made several assumptions to allow us to develop an executable model to undertake these analyses:

- People with lung nodules will have one primary lung nodule.
- Before detection a nodule grows, but after detection, a benign nodule does not continue growing.
- We assumed that lung nodules not identified at initial CT scan but later detected or diagnosed as cancer were initially present but undetected, and thus they were not new lung nodules or interval cancers.
- Due to the paucity of information for the incidental population, we assumed that the population is similar to the screening population and hence we used the same model input values for both population except for the prevalence of any lung nodules.
- Benign nodules were assumed to have grown up to the point of detection but not to have grown afterwards.
- For the AI-assisted reading strategy, we assumed that 95% of people with benign nodules would be discharged at the 1-year CT surveillance and 5% would be discharged at the 2-year CT surveillance. For the unaided reading strategy, we assumed that 95% of people would be discharged at the 2-year CT surveillance and 5% at the 1-year CT surveillance.
- Among false-negative cases at initial CT scan, we assumed that 0.04% would be malignant.[4]
- We assumed a utility decrement associated with undergoing a biopsy as −0.2.[85,86]
- There would be no cancers caused by radiation exposure.

## *Analysis*

The economic analysis was undertaken from the perspective of the NHS and PSS and in accordance with CHEERS.[72] The results of the analysis are presented in terms of an ICER, expressed as cost per correct identification of actionable nodules, cost per cancer detected and treated, and cost per QALY gained. Cost-effectiveness was assessed over a lifetime horizon, and all costs incurred and benefits accrued over the model time horizon were discounted at 3.5% per annum in line with recommended guidelines.[22] A deterministic analysis was undertaken for the base case of the primary and secondary outcome measures.

We undertook probabilistic sensitivity analysis to determine the joint uncertainty in model input parameters. We undertook the probabilistic sensitivity analysis based on the outcome of cost per QALY gained only. In the probabilistic sensitivity analysis, each chosen model parameter was assigned a distribution (e.g. beta, Dirichlet or gamma), reflecting the amount and pattern of its variation, and cost-effectiveness results were calculated by simultaneously selecting random values from each distribution. This process was repeated 10,000 times in a Monte Carlo simulation to give an indication of how variation in the model parameters leads to variation in the ICERs for a given strategy. The results of the simulation were plotted on an incremental cost-effectiveness plane, where each simulation/point represents the change/difference in costs divided by the difference/change in their benefits between strategies. We also calculated the probability that each strategy was the most cost-effective at different willingness-to-pay thresholds per QALY gained, with the results plotted on a Cost-Effectiveness Acceptability Curve (CEAC).

Additionally, we undertook several sensitivity and scenario analyses. One-way sensitivity analysis was conducted to determine which input parameters were drivers of the economic analysis. Key input parameters were varied using the upper and lower values, and the results were presented in a tornado diagram.

### *Scenario analyses*

Given the limited evidence available, we had to use information from different studies and sources, which often had some concerns related to risk of bias and applicability, to link evidence on diagnostic accuracy of AI-assisted reading compared with unaided reading of CT scans for identifying and analysing lung nodules to subsequent clinical processes and patient outcomes. Structuring this evidence on the clinical and economic outcomes in the form of a model is likely to introduce uncertainty, especially in several parameter inputs. We addressed this by undertaking scenario analyses for different values for each variable, and structures of the economic model. We identified three parameters that are likely to result in uncertainty around the cost-effectiveness. These parameters include:

- prevalence of lung nodules detected at baseline CT scans
- accuracy for identifying actionable nodules
- time taken to read CT scans.

#### Prevalence of lung nodules detected at baseline computed tomography scans

In the detection phase of the model, we explored using the prevalence of any lung nodules detectable on baseline CT scans from other sources to estimate the impact on the results for the screening and incidental populations. No alternative prevalence information was identified for the symptomatic population. *Table 21* shows the prevalence information that we used in scenario analysis.

#### Accuracy for identifying actionable nodules

The base case includes identifying people with any lung nodules (≥ 3 mm to 30 mm) and discharging people with lung nodules < 5 mm. In this scenario, we explore in the detection phase of the model the impact of identifying 'actionable' nodules and hence using sensitivity and specificity estimates to identify people with lung nodules ≥ 5 mm.

#### Time taken to read computed tomography scans

The time taken to read CT scans was reduced with AI assistance in most studies included in our review.[27,31,34,47,53,54,59,60,64] However, these studies were predominantly conducted under research conditions and there is uncertainty about how AI assistance may impact on read/reporting time in real clinical practice. In the base case, we assumed that the time required to read and report a CT scan image would be shortened from 10 minutes for unaided readers to 8 minutes for AI-assisted reading. In *Table 22*, we report the time taken (expert opinion), by population. In scenario analyses, we explored the possibility of varying this time for different strategies.

### *Areas beyond the scope of the assessment*

A quantitative evaluation of potential effects of using AI-derived software on workflow, changes in the interactions between health professionals and patients and between different health professionals and impact on workload and staffing is beyond the scope of the current assessment, except that where evidence is found on radiologist reading time and/or radiology turnaround time related to the use of the software this will be taken into account in the estimation of costs.

**TABLE 21** Scenario analyses by changing the prevalence of any lung nodules detected at baseline CT scans in a screening population and incidental population, respectively

| Screening population | | Incidental population | |
|---|---|---|---|
| Prevalence used in base model | Prevalence used in scenario analysis | Prevalence used in base model | Prevalence used in scenario analysis |
| 0.509 (Field *et al.* 2016[80]) | 0.33 (Callister *et al.* 2015[12]) | 0.13 (Callister *et al.* 2015[12]) | 0.380 (Lancaster *et al.* 2021[87]) |

**TABLE 22** Resource use associated with reading and reporting CT scans

| | Population of interest | | | |
|---|---|---|---|---|
| **Resource** | **Symptomatic** | **Incidental** | **Screening** | **Surveillance** |
| Radiologist time to report CT scan (AI assisted) | 12 minutes | | 8 minutes | 8 minutes |
| Radiologist time to report CT scan (unaided) | 15 minutes | | 10 minutes | 8 minutes |
| Type of CT scan at baseline | CT scan with contrast | | CT scan without contrast | |
| Type of CT scan during surveillance, if required | CT scan without contrast | | | |

# Chapter 8 De novo cost-effectiveness analysis (full model): results

## Base-case results

The full model comprising two stages provides a quantitative framework to link diagnostic accuracy using AI-assisted reading compared with unaided reading of CT scans for identifying any lung nodules, to determining those requiring further actions and then to tracking the growth of the lung nodules under further surveillance, to the short-term costs (costs associated with correct identification of actionable lung nodules) and benefits (number of lung cancers identified) and the long-term costs and health outcomes expressed in QALYs. We first present findings related to intermediate outcomes in *Table 23* and then summarise deterministic results for the following outcomes: cost per correct identification of actionable nodules, cost per cancers detected and treated, and cost per QALY. The results are based on assuming a hypothetical cohort of 1000 people undergoing a CT scan.

Findings are presented for the symptomatic population, the incidental population and the screening population. Additionally, we present sensitivity and scenario analyses results.

### *Symptomatic population*

Deterministic results are reported in *Tables 24–26* for the symptomatic population.

### Cost per correct identification of people with actionable nodules

*Table 24* presents the estimates of the costs and additional people correctly identified with an actionable nodule with the use of AI-assisted radiologist reading compared with unaided radiologist reading in a symptomatic population. These

**TABLE 23** Summary of intermediate outcomes from the full model

| | Symptomatic | | Incidental | | Screening | |
|---|---|---|---|---|---|---|
| **Results** | **AI assisted** | **Unaided** | **AI assisted** | **Unaided** | **AI assisted** | **Unaided** |
| Correct detection of any lung nodules | 808.0000 | 645.5000 | 110.7000 | 88.4000 | 422.5000 | 371.6000 |
| Correct detection of actionable nodules | 481.8000 | 333.4000 | 58.6000 | 42.5000 | 223.8000 | 178.7000 |
| Lung cancer detected at first presentation | 7.0100 | 6.5510 | 1.3985 | 1.0810 | 5.3351 | 4.5423 |
| Cancer detected at 3-month CT surveillance | 1.9230 | 3.6700 | 0.2181 | 0.3506 | 0.8326 | 1.4732 |
| Cancer detected at 1-year CT surveillance | 2.3120 | 1.2360 | 0.2233 | 0.1796 | 0.8523 | 0.7546 |
| Cancer detected at 2-year CT surveillance | 1.9060 | 0.758 | 0.1563 | 0.1227 | 0.5964 | 0.5158 |
| Cancer detected at 4-year CT surveillance | 2.3600 | 0.6140 | 0.1893 | 0.1105 | 0.7225 | 0.4642 |
| Cancers detected | 15.5120 | 12.8290 | 2.1850 | 1.8440 | 8.3420 | 7.7500 |
| Cancers missed (< 5 mm) | 2.2823 | 2.8212 | 0.3702 | 0.3673 | 1.4129 | 1.5433 |
| Cancers missed (no lung nodule detected) | 0.5641 | 4.992 | 0.0773 | 0.7069 | 0.3461 | 1.7816 |
| Cancers missed (slow-growing) | 4.1302 | 1.8466 | 0.5879 | 0.3023 | 2.2439 | 1.2701 |
| Cancers missed | 6.9770 | 9.6600 | 1.0353 | 1.3764 | 4.0029 | 4.5950 |
| Total cancers | 22.4890 | 22.4890 | 3.2203 | 3.2204 | 12.3450 | 12.3450 |

**TABLE 24** Deterministic results based on expected costs and expected correct identification of people with actionable lung nodules (symptomatic population of 1000 people undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with actionable nodules correctly identified | Incremental number of people with actionable nodules correctly identified | ICER (£) per correct identification of an individual with actionable lung nodules |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (InferRead CT Lung) | 138,740 | – | 481.8 | – | – |
| Unaided radiologist reading | 142,750 | 4010 | 333.4 | −148.4 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

**TABLE 25** Deterministic results based on expected costs and expected correctly identified people with lung cancer detected and treated (symptomatic population of 1000 people undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with cancer correctly detected and treated | Incremental number of people with cancer correctly detected and treated | ICER (£) per cancer correctly detected and treated |
|---|---|---|---|---|---|
| Unaided radiologist reading | 715,450 | – | 12.83 | – | – |
| AI-assisted radiologist reading (ClearRead CT) | 816,520 | 101,080 | 15.51 | 2.68 | 38,316 |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

**TABLE 26** Deterministic results based on expected costs and expected QALYs (symptomatic population of 1000 people undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| Unaided radiologist reading | 715,450 | – | 6349.89 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 816,520 | 101,080 | 6329.90 | −19.99 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

results show that AI-assisted radiologist reading (InferRead CT Lung) is approximately £4000 cheaper and expected to correctly identify an additional 148.4 people with actionable nodules, thereby dominating the unaided reading strategy.

## Cost per cancer correctly detected and treated

Results in *Table 25* show that the AI-assisted reading strategy is approximately £101,100 more costly and is expected to correctly identify and treat an additional 2.68 people with lung cancer, which equates to an ICER of approximately £38,300.

### Cost per quality-adjusted life-year

Results in *Table 26* show that unaided reading strategy dominates by being less costly and more effective than AI-assisted radiologist reading (InferRead CT Lung) when QALYs are considered.

### Sensitivity analysis

Deterministic sensitivity analysis results were conducted by varying key model input parameters by their ranges or when unavailable by assuming ± 50% for time required to read and report CT scan with/without AI software and ± 10% cost of CT scan of the base-case values to assess the impact on the ICER (cost per QALY), with the results presented in the form of tornado diagrams.

*Figure 11* shows the impact on the cost per QALY by varying inputs. Results show that the sensitivity of unaided reading and the times taken to read and report results (for both AI-assisted and unaided reading) are the most influential. However, within the limits used the results continued to show that unaided reading dominated AI-assisted radiologist reading.

### Scenario analyses

*Table 27* shows that AI-assisted reading remains dominated by unaided radiologist reading in most scenarios explored, except when no disutility associated with false-positive nodule detection and CT surveillance was assumed.

## *Incidental population*

### Cost per correct identification of a person with actionable lung nodules

*Table 28* presents the estimates of the costs and additional people correctly identified with an actionable nodule with the use of AI-assisted radiologist reading compared to unaided radiologist reading in an incidental population. These results show that AI-assisted radiologist reading (InferRead CT Lung) is approximately £4000 cheaper and expected to correctly identify an additional 16.1, resulting in the unaided reading strategy being dominated.

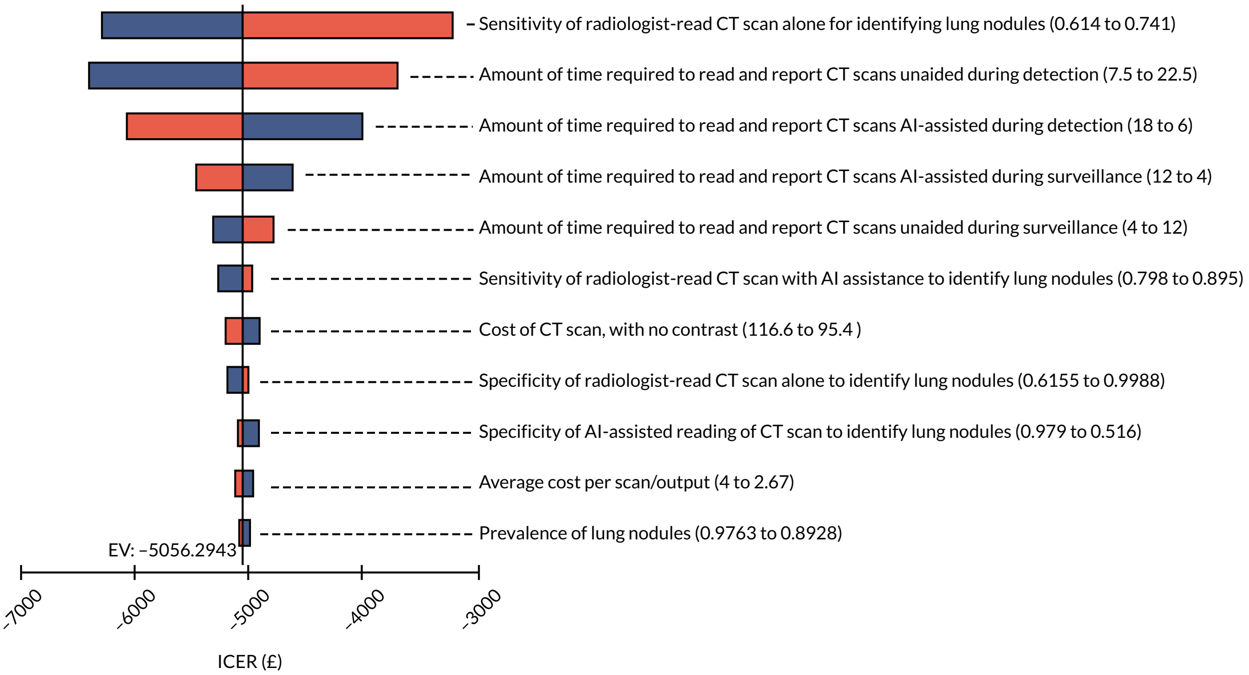


**FIGURE 11** Tornado diagram of the impact on the cost per QALY from changing individual parameters (symptomatic population). Note: the ICERs shown were for AI-assisted radiologist reading compared with unaided radiologist reading.

**TABLE 27** Scenario analysis results based on cost per QALY (symptomatic population)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| ***Base case*** | | | | | |
| Unaided radiologist reading | 715,450 | – | 6349.89 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 816,520 | 101,080 | 6329.90 | −19.99 | Dominated |
| ***Prevalence of detecting any lung nodules (0.9490–0.5000) (assumption)*** | | | | | |
| Unaided radiologist reading | 450,060 | – | 6416.06 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 508,780 | 58,780 | 6403.04 | −13.18 | Dominated |
| ***Time taken to read and report CT scans: assumed to take 12 minutes for AI-assisted and unaided*** | | | | | |
| Unaided radiologist reading | 704,700 | – | 6349.89 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 816,520 | 111,830 | 6329.90 | −19.99 | Dominated |
| ***Time taken to read and report CT scans: assumed to take 15 minutes for AI-assisted and 12 minutes unaided*** | | | | | |
| Unaided radiologist reading | 704,700 | – | 6349.89 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 826,890 | 122,190 | 6329.90 | −19.99 | Dominated |
| ***People with benign nodules discharged at 2-year CT surveillance (solid nodules) and 4-year CT surveillance (subsolid nodules) in both strategies*** | | | | | |
| Unaided radiologist reading | 717,470 | – | 6349.50 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 860,190 | 142,720 | 6320.50 | −29.00 | Dominated |
| ***No disutility associated with false-positive nodules during detection or disutility associated with undergoing CT surveillance*** | | | | | |
| Unaided radiologist reading | 715,450 | – | 6385.86 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 816,520 | 101,080 | 6393.81 | 7.95 | 12,709 |

**TABLE 28** Deterministic results based on expected costs and expected cases appropriately identified (incidental population of 1000 people undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with actionable nodules correctly identified | Incremental number of people with actionable nodules correctly identified | ICER (£) per correct identification of an individual with actionable lung nodules |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (InferRead CT Lung) | 138,740 | – | 58.6 | – | – |
| Unaided radiologist reading | 142,750 | 4010 | 42.5 | −16.1 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

Results in *Table 29* show that the AI-assisted reading strategy is approximately £2430 cheaper and is expected to correctly identify and treat an additional 0.34 people with lung cancer, resulting in its dominance over unaided radiologist reading.

## Cost per quality-adjusted life-year

Results in *Table 30* show that the unaided strategy is £2430 more costly and expected to yield an additional 2.44 QALYs in an incidental population undergoing CT scan, yielding an ICER of £996 per QALY.

## Sensitivity analysis

*Figure 12* shows the impact on the cost per QALY of varying model inputs. Results show that prevalence of lung nodules is the most influential driver. Higher prevalence of lung nodules is associated with more favourable cost-effectiveness for AI-assisted reading.

## Scenario analyses

*Table 31* shows that the cost-effectiveness of AI-assisted radiologist reading compared with unaided radiologist reading is highly uncertain for incidental population and may change between different scenarios.

### *Screening population*

Deterministic results are reported in *Table 32–34* for the screening population.

**TABLE 29** Deterministic results based on expected costs and expected cancer correctly detected and treated (incidental population of 1000 undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with cancer correctly detected and treated | Incremental number of people with cancer correctly detected and treated | ICER (£) per cancer correctly detected and treated |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (InferRead CT Lung) | 229,210 | – | 2.185 | – | – |
| Unaided radiologist reading | 231,640 | 2430 | 1.844 | −0.34 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

**TABLE 30** Deterministic results based on expected costs and expected QALYs (incidental population of 1000 undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (InferRead CT Lung) | 229,210 | – | 6571.19 | – | – |
| Unaided radiologist reading | 231,640 | 2430 | 6573.63 | 2.44 | 996 |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

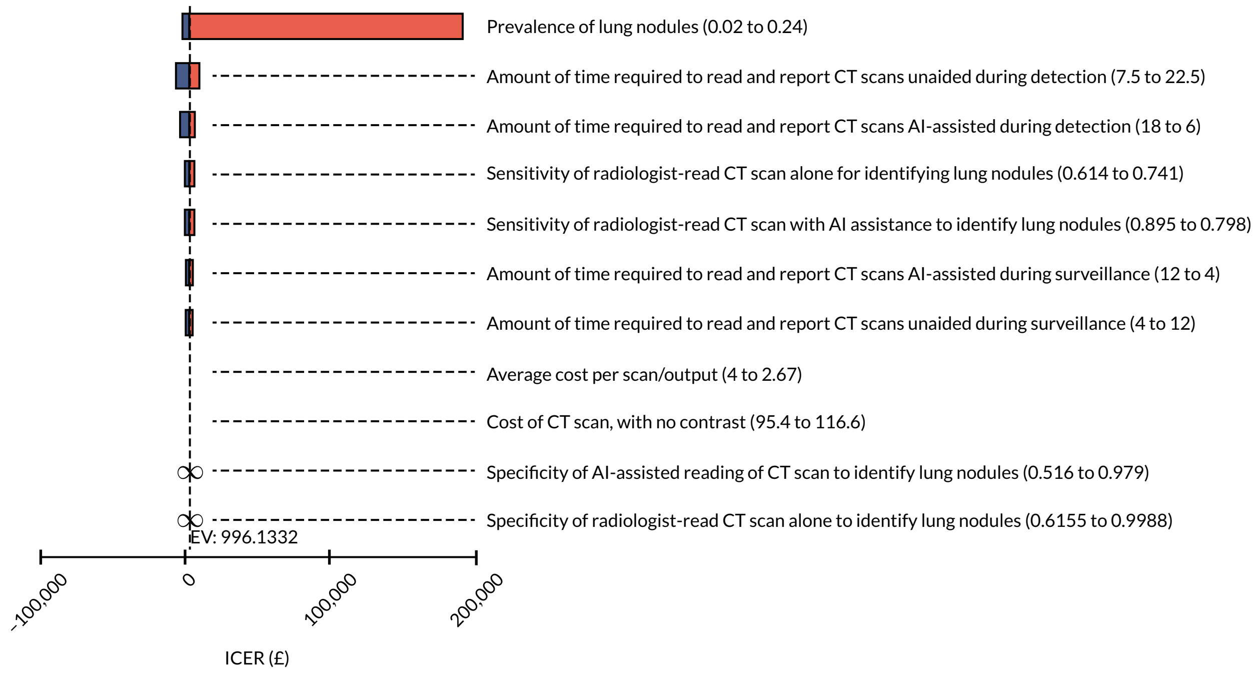


**FIGURE 12** Tornado diagram of the impact on the cost per QALY identified from changing individual parameters (incidental population). Note: the ICERs shown were for AI-assisted radiologist reading compared with unaided radiologist reading.

**TABLE 31** Scenario analysis results based on cost per QALY (incidental population)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| *Base case* | | | | | |
| AI-assisted radiologist reading (InferRead CT Lung) | 229,210 | – | 6571.19 | – | – |
| Unaided radiologist reading | 231,640 | 2430 | 6573.63 | 2.44 | 996 |
| *Prevalence of detecting any lung nodules (0.1300–0.3800)* | | | | | |
| AI-assisted radiologist reading (InferRead CT Lung) | 356,490 | – | 6541.56 | – | – |
| Unaided radiologist reading | 381,670 | 25,180 | 6538.59 | −29.6 | Dominated |
| *Time taken to read and report CT scans: assumed to take 12 minutes for AI-assisted and unaided* | | | | | |
| Unaided radiologist reading | 223,910 | – | 6573.63 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 229,210 | 5300 | 6571.19 | −2.44 | Dominated |
| *Time taken to read and report CT scans: assumed to take 15 minutes for AI-assisted and 12 minutes unaided* | | | | | |
| Unaided radiologist reading | 223,910 | – | 6573.63 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 236,580 | 12,670 | 6571.19 | −2.44 | Dominated |
| *People with benign nodules discharged at 2-year CT surveillance (solid nodules) and 4-year CT surveillance (subsolid nodules) in both strategies* | | | | | |
| Unaided radiologist reading | 231,900 | – | 6573.58 | – | – |
| AI-assisted radiologist reading (InferRead CT Lung) | 232,540 | 640 | 6570.46 | −3.11 | Dominated |
| *No disutility associated with false-positive nodules during detection or disutility associated with undergoing CT surveillance* | | | | | |
| AI-assisted radiologist reading (InferRead CT Lung) | 229,210 | – | 6583.58 | – | – |
| Unaided radiologist reading | 231,640 | 2430 | 6582.69 | −0.89 | Dominated |

**TABLE 32** Deterministic results based on expected costs and expected correct identification of people with actionable nodules (screening population of 1000 people undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with actionable nodules correctly identified | Incremental number of people with actionable nodules correctly identified | ICER (£) per correct identification of an individual with actionable lung nodules |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (ClearRead CT) | 127,600 | – | 223.8 | – | – |
| Unaided radiologist reading | 130,500 | 2900 | 178.7 | −45.1 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

**TABLE 33** Deterministic results based on expected costs and expected identification of people with cancer detected and treated (screening population of 1000 people undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected number of people with cancer correctly detected and treated | Incremental number of people with cancer correctly detected and treated | ICER (£) per cancer correctly detected and treated |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (ClearRead CT) | 400,410 | – | 8.342 | – | – |
| Unaided radiologist reading | 470,630 | 70,220 | 7.750 | −0.592 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

**TABLE 34** Deterministic results based on expected costs and expected QALYs (screening population of 1000 undergoing CT scan)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (ClearRead CT) | 400,410 | – | 6532.1 | – | – |
| Unaided radiologist reading | 470,630 | 70,220 | 6524.1 | −7.9549 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

## Cost per correct identification of a person with an actionable lung nodule

*Table 32* presents the estimates of the costs and additional people correctly identified with an actionable nodule with the use of AI-assisted radiologist reading compared with unaided radiologist reading in a screening population. These results show that AI-assisted radiologist reading (ClearRead CT) is expected to correctly identify an additional 45.1 people with actionable nodules. The use of AI-assistance software strategy is cheaper than unaided reading, resulting in the latter being dominated.

### Cost per cancer correctly detected and treated

Results in *Table 33* show that the AI-assisted reading strategy is cheaper and is expected to correctly identify and treat an additional 0.592 people with lung cancer resulting, thus dominating the unaided radiologist reading strategy.

### Cost per quality-adjusted life-year

Results in *Table 34* show that the AI-assisted radiologist reading strategy is cheaper and expected to yield 7.9549 more QALYs, thus dominating the unaided radiologist reading strategy.

### Sensitivity analysis

*Figure 13* shows the impact on the cost per QALY of varying model inputs. Results show that the amount of time required to read and report CT scan for unaided readers and AI-assisted readers is the most influential driver.

### Scenario analyses

*Table 35* shows that unaided radiologist reading remains dominated by AI-assisted reading under the various scenarios explored.

In addition to sensitivity and scenario analyses presented above, the EAG further carried out a probabilistic sensitivity analysis for each of the populations. The findings are presented in *Appendix 9*, *Figures 24–29*. Results suggest that unaided reading has very high probability of being cost-effective for the symptomatic population, while AI-assisted reading has very high probability of being cost-effective for the screening population. Uncertainty is much higher for the incidental population. The EAG recognised that there are additional uncertainties that might not have been fully captured in these analyses.

### *Surveillance population*

In addition to exploring the cost-effectiveness of AI-assisted image analysis in the symptomatic, incidental and screening populations, the EAG undertook a cost-effectiveness analysis in the surveillance population. This population represents people who have an actionable nodule detected and require CT surveillance. The population is of interest as a main advantage of AI-assisted image analysis lies in improved reliability of nodule size measurement, based on which VDT or nodule size growth is determined, and this in turn influences clinical decision-making after the follow-up scan.

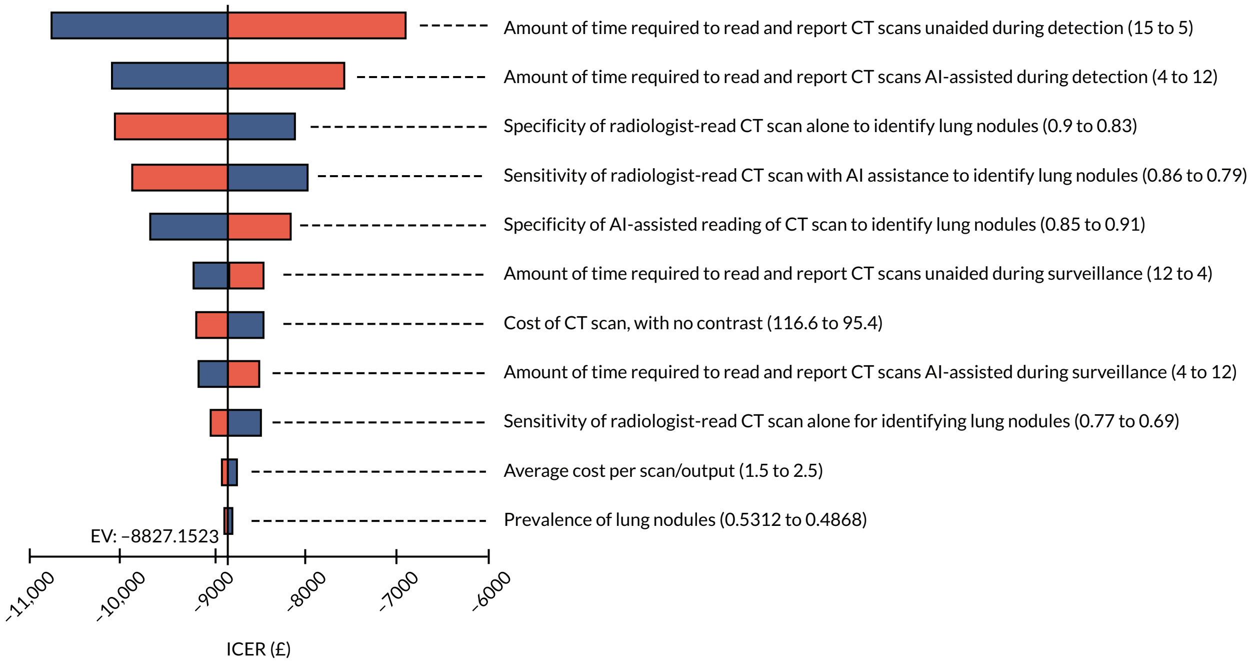


**FIGURE 13** Tornado diagram of the impact on the cost per QALY of changing individual parameters (screening population). Note: the ICERs shown were for AI-assisted radiologist reading compared with unaided radiologist reading.

**TABLE 35** Scenario analysis results based on cost per QALY (screening population)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| ***Base case*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 400,410 | – | 6532.1 | – | – |
| Unaided radiologist reading | 470,630 | 70,220 | 6524.1 | −7.95 | Dominated |
| ***Prevalence of detecting any lung nodules (0.509–0.330)*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 310,590 | – | 6552.28 | – | – |
| Unaided radiologist reading | 357,460 | 46,870 | 6546.68 | −5.60 | Dominated |
| ***Time taken to read and report CT scans: assumed to take 10 minutes for AI-assisted and unaided*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 405,350 | – | 6532.1 | – | – |
| Unaided radiologist reading | 470,630 | 65,280 | 6524.1 | −7.95 | Dominated |
| ***Time taken to read and report CT scans: assumed to take 12 minutes for AI-assisted and 10 minutes unaided*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 410,290 | – | 6532.08 | – | – |
| Unaided radiologist reading | 470,630 | 60,340 | 6524.12 | −7.95 | Dominated |
| ***People with benign nodules discharged at 2-year CT surveillance (solid nodules) and 4-year CT surveillance (subsolid nodules) in both strategies*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 412,620 | – | 6529.31 | – | – |
| Unaided radiologist reading | 471,660 | 59,040 | 6523.89 | −5.42 | Dominated |
| ***No disutility associated with false-positive nodule detection or disutility associated with undergoing CT surveillance*** | | | | | |
| AI-assisted radiologist reading (ClearRead CT) | 400,410 | – | 6548.21 | – | – |
| Unaided radiologist reading | 470,630 | 70,220 | 6547.32 | −0.89 | Dominated |

This analysis, therefore, focuses on, and isolates out, the potential impact of improved measurement reliability on health and economic outcomes following CT surveillance. It is worth noting that assessment of nodule growth relies on two (or more) measurements, and so the first (previous) CT scan also contributes to any potential benefits of a reading strategy that would be realised at the follow-up scan. Consequently, we retain the original characteristics of the surveillance population (e.g. whether they belong to a symptomatic or screening population at the initial scan) and assume that the same reading strategy is used at both scans.

## Cost per quality-adjusted life-year

The results in *Table 36* are reported for a screening population who are under surveillance. Here we assumed that this population excludes people with nodules that have clear benign features or people with lung nodules measuring < 5 mm on the initial scan. Information used to undertake these analyses was obtained from our simulation used to inform the cost-effectiveness analysis in the full model for the screening population. Within this screening population under surveillance, we obtained information about the number of people with benign nodules (and when they were discharged), the number of cancers detected (and when they were detected) and the number of cancers missed. Costs and QALYs yielded were affixed to these proportions. In this scenario, we assumed that people detected with cancer all have stage I disease. Additionally, we assumed that any person with a cancer missed by the surveillance will present later with stage I disease.

**TABLE 36** Deterministic results based on expected costs and QALYs (screening population of 1000 people undergoing CT surveillance)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (InferRead CT Lung) | 719,813 | – | 6365.01 | – | – |
| Unaided reading | 921,015 | 201,202 | 6323.07 | −41.94 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

These results show that the AI-assisted strategy is less costly and more effective, thus dominating the unaided strategy.

### Scenario analyses

We undertook further scenario analysis in which we assumed that people whose cancers were missed during surveillance would present with stage IV disease instead (*Table 37*). These results showed that the AI-assisted strategy continued to dominate unaided reading in this patient population.

## Discussion

### *Summary of key results*

Artificial intelligence assistance increases the number of lung nodules detected at first presentation. The number of extra nodules detected per 1000 persons screened is 162.5, 22.3 and 50.9 for symptomatic, incidental and screening populations, respectively. It also increases the number of actionable nodules detected. The number of extra actionable nodules detected per 1000 persons screened is 148.4, 16.1 and 45.1 for symptomatic, incidental and screening populations, respectively.

The majority of these additional nodules detected will be benign. There will be additional costs associated with investigating them, and potentially disutility experienced during the time that the nodule is under investigation and the possibility of malignancy remains. However, we assume that a proportion of these additional nodules will be malignant and therefore detected early because of the nodule's correct identification. For every 1000 persons screened, the number of additional cancers detected in this way by AI assistance would be 5.0, 0.6 and 1.6 for symptomatic, incidental and screening populations, respectively (3–4% of the additional actionable nodules detected).

All actionable nodules assessed as being between 5 mm and 8 mm undergo surveillance and are investigated only if the growth rate is above a certain threshold. It is possible for some malignant tumours to be missed if their measured growth rate is too low. Our modelling suggests that this is slightly more likely with AI assistance. Per 1000 persons screened, AI assistance would result in 2.3, 0.3 and 1.0 fewer cancers being detected during surveillance. The reason for this is likely to be our assumption that AI assistance, although improving measurement accuracy, also introduces a systematic overestimation of size. The way this is modelled implies that, when repeated measurements are taken to estimate VDTs, these will be systematically underestimated. However, the cancers missed this way will be slow-growing and therefore likely to be less aggressive, implying that the consequences of not detecting them will be less severe than those of missing cancers through failing to detect a nodule.

In terms of cost per QALY, use of AI was estimated to be cost-effective in the screening population, but not in the symptomatic or incidental population. For symptomatic, screening and incidental populations, use of AI reduced costs initially through reducing nodule detection costs, and detected more actionable nodules, resulting in AI dominating unaided readers for the outcome of actionable nodule detection. This translated to £38,316 per extra cancer detected for the symptomatic population, whereas AI dominated unaided readers for cancer detection in the incidental and screening populations, with lower costs and increased cancer detection. In the symptomatic population, the increased cancer detection does not translate into an overall QALY gain, and AI is more expensive when the cost of follow-up tests and CT surveillance is included, and so AI is dominated by unaided readers in the assessment of cost per QALY.

**TABLE 37** Scenario analysis assuming people with cancers missed during surveillance would present with stage IV disease (screening population of 1000 people undergoing CT surveillance)

| Strategy | Expected total costs (£) | Incremental costs (£) | Expected QALYs | Incremental QALYs | ICER (£) per QALY |
|---|---|---|---|---|---|
| AI-assisted radiologist reading (InferRead CT Lung) | 699,100 | – | 6345.62 | – | – |
| Unaided reading | 898,678 | 199,578 | 6302.17 | −43.46 | Dominated |

**Note**
Exact results were obtained from TreeAge but rounded by the authors and presented.

One scenario analysis, the removal of QALY decrement for false-positive results and CT surveillance, resulted in a cost per QALY of £12,709 for AI in comparison with unaided reading. This is below the £20,000 threshold, indicating that the QALY decrement and increased follow-up costs for false-positive results and CT surveillance is the reason why AI assistance is not cost-effective in the base case in the symptomatic population. The distress caused to the large number of individuals in whom benign nodules are found outweighs the health gains experienced by the few whose cancers would have been missed without AI assistance. No other sensitivity or scenario analysis significantly changed the results. In the incidental population there were higher QALYs overall for the unaided reader than for the AI-assisted strategy, so unaided reading had a cost per QALY of £996 compared with AI assistance, indicating that the addition of AI assistance is not cost-effective in this population. This result was sensitive to the prevalence of lung nodules in the population, with increased prevalence favourable towards AI, which was estimated to have greater sensitivity to detect these nodules in the model. Removal of the QALY decrement for false-positive results and CT surveillance in a scenario analysis resulted in AI dominating the unaided reader. This indicates that the cost per QALY is heavily influenced by the costs and QALY decrements of false-positive results and surveillance. In the screening population, AI was cost-effective and dominated unaided readers in cost per QALY, a result that was unaffected by sensitivity and scenario analyses. Many of the data inputs for the screening population differed from those for the other two populations, because there were different data sources and more data available, including from screening trials. The driving force behind AI assistance estimates being cost-effective for screening and not for the other two populations is in the estimated number of false-positive results and people undergoing CT surveillance. In the screening population there are fewer people experiencing these harms and costs when AI assistance is used than when unaided readers are used. In the symptomatic and incidental populations more people experience these harms and costs when AI assistance is used than when unaided readers are used. This can be seen in the differing impacts of removing the disutility associated with false-positive results and CT surveillance, which improves cost-effectiveness for the symptomatic and incidental populations and reduces cost-effectiveness estimates for the screening population. This is driven by differing data inputs; for example, the screening data suggested that AI was more specific, whereas the symptomatic and incidental data suggested that the unaided reader was more specific (see *Test accuracy*). Although there were more data for the screening population, there was a paucity of available data throughout.

Our modelling does include limitations, largely driven by the data available to populate it. It is possible that we have overestimated the proportion of additional nodules that are malignant, which would exaggerate the benefits of improved nodule detection with AI assistance. We have used the best sources in the literature we could find to inform the size distribution of actionable nodules at initial assessment, the measurement error with or without AI-assistance, and the growth rate of malignant nodules during surveillance. However, these are taken from studies in different populations, with their own limitations, which does affect the robustness of our results.

### Generalisability of results

A key limitation is the paucity of data, with major concerns regarding generalisability. For example, while our base-case analysis indicates that AI-assisted reading dominates unaided reading in the screening population, the test accuracy results suggesting that AI-assisted reading has both better sensitivity and specificity for detecting any lung nodules came from a single study conducted in Taiwan.[53] The results are not consistent with findings from other studies, which suggested that the specificity for AI-assisted reading tends to be worse than that for unaided reading. The risk of bias and applicability concerns commonly found in studies included in our systematic review and highlighted in *Methodological quality of the evidence* further limit the generalisability of findings of our cost-effectiveness analyses.

Only one study included in our test accuracy review was carried out specifically in an incidental population. Consequently, many model parameters were assumed for this population, and this may limit the validity and generalisability of findings particularly in relation to the incidental population.

Our cost-effectiveness analysis would be most generalisable to technologies (ClearRead CT and InferRead CT Lung) that have directly contributed to model parameter inputs related to test accuracy and costs. The generalisability of the findings to other technologies would depend on the demonstration of equivalent or more favourable evidence. However, it is worth reiterating that, overall, our findings are highly uncertain because of the paucity of evidence and other issues explicated in *Strengths and limitations of analysis*.

### *Strengths and limitations of analysis*

Our economic analysis has several strengths:

- As far as we are aware, it is the first full economic evaluation that has explicitly modelled nodule detection and management in accordance with the BTS guidelines, which is the current standard practice in the UK. Our economic evaluation is also likely to be the first to evaluate the cost-effectiveness of AI-assisted reading of chest CT scans compared with reading by unaided radiologists for the detection and analysis of lung nodules.
- Despite the complete absence of clinical effectiveness and cost-effectiveness evidence and the substantial gaps between data concerning the performance of different image analysis strategies and downstream clinical outcomes, our innovative approach of using simulations to inform decision-analytic models based on available data enabled us to conduct a full economic evaluation for the primary comparison of interest.
- The parameter inputs for our model are informed by our systematic review of test accuracy.
- Although the decision-analytic model that we created is likely to require further refinement and validation, and the findings are highly uncertain because of the paucity of data, it provides a useful framework that will allow further evaluation to be undertaken when more evidence emerges.

Although our simulations have enabled us to explore the potential impact of improved consistency in nodule measurement quantitatively in an explicit way, many simplifying assumptions are required during their implementation, with corresponding limitations. These are as follows:

- The starting point of the simulation is a population who all have a nodule detected (or detectable by reference standard). Hence, when we apply this to the economic model, the better nodule detection with AI assistance is not automatically captured in the simulation. This benefit is modelled separately in the decision tree, but this approach leads to a slight imbalance in the total number of nodules, which creates a small artificial difference in the number of cancer cases in the populations considered by the different readers. We correct for this through an adjustment of cancer prevalence among nodules not detected with unaided reading to ensure equal cancer prevalence between the populations subject to different detection strategies.
- We assume that all nodules presented are 3–30 mm in starting size and come from a log-normal distribution. These cut-off values are plausible, but it is possible that there are nodules slightly smaller or bigger than these thresholds. The log-normal distribution was the best at replicating the source information we had to describe median and IQR of the sizes; however, it probably does not perfectly capture the true distribution of starting sizes.
- We assume that only malignant nodules grow; however, it is possible that benign nodules do show some growth and may be falsely detected as cancerous.
- We do not account for the occurrence of new nodules or new cancers within the follow-up of the simulation. Potential issues related to overdiagnosis are not considered.[88]
- Each patient's solid nodule growth is assumed to follow a single Gompertz growth rate as reported in the literature. While each rate varies over time, it may not fully represent the full range of growth rates (e.g. account for periods of nodule dormancy).
- The subsolid nodules are assumed to follow a linear growth rate based on how their growth is reported in the literature. Part-solid and non-solid tumours were modelled separately and pooled in a ratio of 4 : 5. The linear growth

assumption, while following growth, is not capped and means that the nodules of some patients may grow much faster than would occur in real life.

- No mortality is factored into the simulation and so cases of severe or fast-moving disease that are not detected early on may have their QALY contribution overestimated.
- We assume that the measurement error is random and not correlated with any patient characteristics. The base case currently assumes that the error term for a patient with a benign nodule is the same across all of their measurements, meaning there is no possibility of falsely detected growth; however, we explore having an independent error term for each measurement in a scenario analysis.
- Despite using the reported standard deviations, which were generally small, it is likely that a small number of patients had a large measurement error that is unlikely to be representative of practice.
- We focus on the risk-dominant nodule (the largest single nodule) per patient and do not consider cases where there may be multiple nodules in different locations.
- It is assumed that all nodules identified as having clear features of benignancy are in fact benign.
- When categorising patients at later follow-up points, stable patients would usually fulfil the criteria for more than one of the stable categories (e.g. VDT > 600, stable of diameter, stable on volumetry), and it was not possible to generate a sequential order of allocation or distribution across these groups. These categories do, however, differ in the resulting follow-up. These differences should represent differences in methods and technology available at each site; however, no information was available.
- Assessment of malignancy in later follow-up was based on VDT, where the growth rate (and thus the VDT) was independent of the starting nodule size.

The large number of simplifying assumptions indicates a high level of structural and methodological uncertainty associated with the decision-analytic model that may have not been captured in sensitivity and scenario analyses presented in this report. While the EAG has made every effort to create and refine this modelling framework to enable the use of very sparse and heterogeneous evidence to evaluate the clinical effectiveness and cost-effectiveness of the technologies of interest, this work was undertaken within a fairly limited timeframe and therefore further validation and refinement of the model is likely to be needed. Current findings from the model should therefore be interpreted with great caution.

# Chapter 9 Assessment of factors relevant to the National Health Service and other parties

This technology assessment focuses on evaluating test accuracy, clinical effectiveness and cost-effectiveness. Several other factors that are outside the scope of the assessment may need to be considered with respect to the potential adoption of AI software assistance into clinical practice and service delivery:

- choice between AI software in the absence of comparative accuracy and clinical evidence
- estimating the effectiveness and cost-effectiveness of AI software capable of detecting and analysing multiple disease conditions
- integration of the technologies into existing PACS and workflow; compatibility with existing CT scanners and workstations
- different costs and costing structures in relation to the volume of CT scans and patient characteristics for individual institutions
- training required for using AI software and learning curve
- ongoing update and user support
- potential impact of increased CT surveillance on patients' mental well-being and quality of life, and issues related to overdiagnosis
- potential impact on radiology service planning and delivery and human resource management, including impact on other services requiring CT scans
- potential interruption to service due to cyber-security issues, and network and data security issues for cloud-based system.

# Chapter 10 Discussion

## Statement of principal findings

- Twenty-seven studies evaluating eight of the 13 technologies specified in the assessment protocol were included. All studies reported findings related to test accuracy. No study providing direct evidence on clinical effectiveness and cost-effectiveness was found. All included studies were judged to be at high risk of bias, and most studies have several applicability concerns for the UK setting.
- The majority of studies (24/27) used retrospective data sets and were conducted in research settings. Only two of these studies were undertaken in the UK. Seventeen studies compared the performance of readers with and without concurrent software use (primary comparison of interest). Additional evidence related to stand-alone AI software and non-comparative evidence from these retrospective studies was also reviewed to provide supplementary information. The remaining three studies reported on prospective screening experiences based on the same screening pilot trial conducted in the Republic of Korea.
- Evidence suggests that AI-assisted CT image analysis may increase the sensitivity of lung nodule detection but may also increase false-positive findings. Consistency between readers in the detection of nodules may improve and variability may reduce when they are assisted by AI. Evidence from research settings suggests that reading time for CT image analysis may be reduced with the assistance of AI software. All these findings require further validation in studies using prospectively collected data in clinical practice settings.
- Segmentation failure by AI-derived software is not uncommon and may impact on its performance in clinical practice settings.
- Different AI software may have different test accuracy and performance to identify lung nodules among patients with different clinical features, and different types of lung nodules. However, there is an absence of direct comparative evidence between (analysis assisted by) different AI software.
- The limited number of studies available and concerns related to risk of bias and applicability mean that estimates of test accuracy for individual technologies are either absent or highly uncertain and require further validation and confirmation.
- In the absence of direct evidence on clinical effectiveness and cost-effectiveness, the EAG created a de novo full model to link up the long causal chain between test accuracy and clinical and economic outcomes. The paucity of data and methodological challenges mean that the findings of the linked evidence approach are highly uncertain and need to be interpreted with great caution.
- Acknowledging the above caveats, the EAG's cost-effectiveness analysis suggests that test accuracy of unaided readers and of AI-assisted reading, radiologists reporting time with and without AI-assistance, prevalence of lung nodules and disutility associated with CT surveillance are likely to be key drivers of cost-effectiveness. AI-assisted reading is likely to be dominated by unaided reading unless AI-assisted reading could improve both sensitivity and specificity compared with unaided reading.

## Strengths and limitations of the assessment

### *Strengths*

The strengths of this technology assessment include:

- Comprehensive and systematic searches of relevant literature were undertaken, supplemented by requests for evidence and data from the companies.
- Rigorous systematic review methods were followed for the selection of studies for inclusion, critical appraisal and synthesis.
- Despite the absence of direct evidence quantifying the impact of AI-derived software on clinical and patient outcomes, we have developed an innovative framework linking up test accuracy evidence with subsequent clinical process and patient outcomes using a decision tree and simulations through a linked evidence approach. This framework may facilitate future evaluation of similar technologies as new evidence emerges.

### *Limitations*

First, the limitations of the review methodology are considered. This is followed by a discussion of the limitations of the evidence identified and included in the review and specific limitations related to economic modelling and simulations adopted by the EAG.

## Limitations of the review

- We excluded literature before 2012; however, studies on AI software published before this date are unlikely to be relevant to the current assessment.
- Only seven companies (Aidence, contextflow, Infervision, JLK, MeVis, Riverain and Siemens Healthineers) submitted information and/or replied to our questions for clarification.
- Fifteen records were excluded at full-text level as the software name was unclear and we received no replies from authors.
- MeVis: excluded studies using the research software CIRRUS as well as studies on the computer-aided detection software Visia.
- Siemens Healthineers: excluded studies on any other technologies, for example, syngo.
- Due to the limited evidence and heterogeneity, neither meta-analysis nor subgroup analysis by ethnicity, nodule type, dose or reader speciality was performed.
- The review did not specifically consider the differences in test accuracy of different AI software because no evidence of direct comparisons between different software was identified, and the included studies were too heterogeneous in design and patient population to allow reliable indirect comparison.
- The adaptation of the QUADAS-2 tool for this review was a first iteration and requires refinement, taking into consideration the QUADAS-2 AI version and AI reporting guides, such as STARD-AI and CONSORT-AI, which are expected to come out in due time.
- The potential impact of AI-assisted image analysis on the overdiagnosis of lung cancer was not considered in this technology assessment.

## Limitations of the evidence

### *Volume and nature of available evidence*

- No studies were identified on five out of 13 technologies. All studies meeting our inclusion criteria reported evidence on test accuracy. No studies reporting direct evidence on clinical effectiveness or cost-effectiveness, or direct evidence comparing different technologies included in this assessment, were found. This made any attempt to evaluate comparative effectiveness and cost-effectiveness of technologies of interest infeasible.
- Of the 27 test accuracy studies included in our review, only two were conducted in the UK: one each on Veolity (MeVis) and Veye Lung Nodules (Aidence). Of the eight technologies for which at least one study was available, only AI-Rad Companion (Siemens Healthineers), ClearRead CT (Riverain Technologies), Veolity (MeVis) and Veye Lung Nodules (Aidence) had at least two studies conducted in Western Europe or North America. This imposes major limitations on the applicability of evidence to the UK setting. The number of studies available for each technology ranges from six (ClearRead CT, Riverain Technologies) to one (Contextflow SEARCH Lung CT, contextflow; VUNO Med-LungCT AI, VUNO). The small number of available studies for most of the technologies also means that the estimation of test accuracy for individual technologies often relies on evidence from a single study (as different papers for the same technology tended to report different outcomes). This, combined with risk of bias and other applicability concerns detailed below, results in a very high level of uncertainty for test accuracy estimates related to individual technologies.
- There is a paucity of evidence on AI-assisted CT image analysis in relation to symptomatic and incidental populations.
- Given all the issues highlighted above, the EAG has summarised and presented the available evidence in a way that provides an overview of the potential impact of using AI-derived software to support nodule detection and analysis compared with current practice (without AI assistance) rather than focusing on the performance of individual technologies, for most of which evidence is still immature. Readers are reminded that such an overview does not imply that key conclusions drawn in this assessment are generalisable to all similar technologies. Rather, our conclusions may serve as a tentative benchmark for individual technologies to demonstrate their performance

by providing equal or better evidence, and as indicators for undertaking further research in many areas of major uncertainty.

- Only some of the potential impacts of introducing software with AI-derived algorithms (as depicted in *Figure 1*) were covered in the studies included in our review. For example, we did not identify evidence evaluating how the change in the number of nodules detected may influence risk prediction using the Brock model, or the actual impact of using volume rather than diameter measurement on patient management. There are also significant gaps between outcomes reported in published studies and downstream measures of clinical processes and patient outcomes. Future studies should be clear about what use cases they are trying to address and pay more attention to demonstrating downstream impact.
- There were inconsistencies in numbers and results between the journal article by Murchison *et al.*[33] and the clinical evaluation report by Aidence.[30] In the DAR results section, we have reported only the results by Murchison *et al.*[33] as this publication was newer and published in a peer-reviewed journal.

### *Applicability concerns, risk of bias and data inconsistency*

- This review focused on identifying evidence that would allow the evaluation of the future integration of AI-based software into UK clinical practice (diagnostic or screening). The most applicable evidence to address this question comes from studies where the index test is the AI software integrated into the diagnostic or screening pathway, as it would be used in clinical or screening practice. These studies need to report the change of the whole pathway when AI is added in concurrent mode. However, the review identified only one non-UK study in which AI software was used prospectively in screening practice.[51]
- Furthermore, the evidence from studies reporting the test accuracy of AI assistance in informing management decision (e.g. discharge, CT surveillance, diagnostic work-up) was scarce and heterogeneous. Most studies focused on only a separate software function, such as nodule detection, nodule measurement or nodule type determination.
- There were no prospective test accuracy studies of consecutive cohorts in clinical practice. The majority of studies were small and used enriched data sets.
- In addition to study location, most studies had additional applicability concerns regarding the target population, for example, nodule-and/or cancer-enriched, undertaken retrospectively in research settings (further discussed below), and slice thickness of CT scans.
- In the only identified study in a symptomatic population,[59] it was unclear whether CT was used according to UK practice as a second-line test after chest X-ray.
- Many studies evaluated AI algorithms as stand-alone systems rather than as an aid to radiologists – raising applicability concerns.
- The reference standard for nodule detection was usually based on majority or consensus of two or more expert chest radiologists.[89]
- Studies evaluating AI algorithms as reader aids mostly used enriched test set MRMC laboratory study designs. These studies used CT images retrospectively collected during routine screening or clinical practice and, under research conditions, requested readers to prospectively read the CT images unaided and AI aided. This results in the well-known laboratory effect, whereby readers under study conditions behave differently from how they would under routine clinical conditions.[90]
- MRMC studies were mainly performed with US or Asian radiologists with different reading experience and specialties. Consequently, the study results have limited applicability to the UK context.
- Further methodological issues of the included studies include the focus on single-centre studies and the reporting of per-nodule sensitivity and number of false-positive detections per image instead of per person-level sensitivity and specificity.
- The applicability of the current evidence to the UK screening context is limited. Studies did not resemble the complete diagnostic pathway in the UK based on the 2015 BTS guidelines;[12] in contrast with clinical practice, readers in the included studies usually had no access to relevant earlier CT images.

- There were inconsistencies in numbers and results between the journal article by Murchison *et al.*[33] and the clinical evaluation report by Aidence.[30] In the DAR results section, we have reported only the results by Murchison *et al.*[33] as this publication was newer and published in a peer-reviewed journal.

## Uncertainties

Uncertainties were associated with high risk of bias and applicability concerns of the available evidence:

- All the issues related to risk of bias and applicability presented in *Methodological quality of the evidence* and highlighted above increase the uncertainty of the estimated test accuracy, clinical effectiveness and cost-effectiveness of the technologies being evaluated in this technology assessment.
- Per-person versus per-nodule analyses: data from per-person analyses would better reflect clinical management related to lung nodules as many people would have more than one lung nodule. Although the BTS guidelines[12] recommend that lung nodules are managed based on the largest one (risk-dominating nodule), in practice other nodules with sizes or features that are not safe to ignore may also be measured and analysed during the same reading session and be followed up during surveillance. Consequently, a per-person analysis of clinical management decisions would reflect the real impact of AI assistance on clinical practice more closely. Nevertheless, the results from per-person analyses or per-nodule analysis based on the risk-dominating nodule are infrequently presented, and in some cases, we have had to use data from per-nodule analyses to inform our model. The impact of this is uncertain and it is difficult to estimate using sensitivity analysis.

Uncertainties were associated with the long causal chain modelled using linked evidence approach:

- One of the main purported benefits of AI-assisted image analysis is the improved precision and accuracy of the measurement of nodule size (diameter or volume) and, by extension, of the estimation of nodule growth. Evidence of the impact of AI assistance on these was reviewed and presented in *Nodule diameter measurement* (diameter measurement), *Nodule volume measurement* (volume measurement) and *Use case 2: nodule growth monitoring in people with previously identified lung nodules* (nodule growth monitoring), respectively. While there is good evidence of improved consistency in nodule measurement between different readers when assisted by AI, evidence of measurement accuracy (e.g. whether measurements assisted by AI systematically over- or underestimate the sizes/volumes of the nodules) is less clear. Furthermore, while separate evidence of measurement precision and accuracy was reported in some studies, evidence of their collective impact on nodule management is scant. In an attempt to capture the potential impact of AI assistance on measurement accuracy and precision, the EAG conducted a series of simulations and developed a nodule growth model to link these pieces of evidence to nodule management decisions to facilitate modelling of health and cost outcomes further downstream. However, the simulation exercise and nodule growth modelling themselves require several parameter inputs and assumptions, which also contribute to the overall uncertainties in cost-effectiveness estimates.

Uncertainty was associated with other methodological challenges:

- Difficulties in defining reference standard for nodule detection.[89]

## Other relevant factors

AI has increasingly been applied to directly predict the risk of malignancy of lung nodules, which could change future clinical management. This is outside the scope of this assessment but is an area of active research.

Our cost-effectiveness analysis only considered the use of AI in the detection and analysis of lung nodules, and its impact on clinical management and patient outcomes related to lung nodules and cancers. A number of AI software capable of detecting and analysing multiple health conditions in chest CT scans have been developed. Evaluating the use of these software, taking into account their impact related to multiple conditions, is beyond the scope of this DAR. Such an evaluation is likely to be highly complex and data- and resource-demanding and may be an area warranting further research.

We were aware that an economic model has been built to support the NSC's assessment of cost-effectiveness of a lung cancer screening programme in the UK. While the model allowed an evaluation of the impact of timing and frequency of LDCT scans on cancer detection, it was not designed to assess the impact of different strategies for nodule detection and analysis within and across individual CT scans, for which we have constructed the full de novo economic model for this technology assessment. Further rationale for developing our model and comparison with the NSC model can be found in *Report Supplementary Material 5*.

## Equality, diversity and inclusion

We set out to explore whether the accuracy of CT image analysis assisted by AI-based software for automated detection and analysis of lung nodules, its practical implications and impact on patient management differ by patient ethnicity. This was important and was prespecified as one of our sub-questions (see *Chapter 1*, *Objectives*), because people from different ethnic groups might have different levels of comorbidities such as tuberculosis that might impact on software's performance in detecting lung nodules. However, no relevant evidence was found.

## Patient and public involvement

We talked to a member of the public who has undergone CT scan and surveillance following participation in the UK's lung cancer screening pilot, and they highlighted the potential impact on a person when a lung nodule requiring follow-up is identified.

# Chapter 11 Conclusions

AI-assisted detection and analysis of lung nodules has the potential to improve the sensitivity of nodule detection and to increase the consistency in nodule measurement compared with unaided reading, but its impact on measurement accuracy is unclear. Current evidence suggests that AI-assisted reading tends to reduce specificity and results in nodules being classified into higher risk categories based on current clinical guidelines, although this may not always be the case. The reported performance of AI-assisted reading varies substantially among published studies (per-person sensitivity for any nodules 0.79–0.99; specificity 0.81–0.97), possibly attributable to heterogeneous study population, reader experience, specialty and reading conditions, other study design features and risk of bias, in addition to potential differences in the performance of individual technologies.

No studies that directly compared the analyses of CT scan images assisted by different technologies were found. Given the paucity of evidence, it is currently not possible to reliably establish the relative effectiveness and cost-effectiveness of strategies adopting different AI software to assist nodule detection and analysis.

No direct evidence on the clinical effectiveness and cost-effectiveness of AI-assisted reading compared with unaided reading for chest CT image analysis related to pulmonary nodules was found. Evaluation of cost-effectiveness using a linked evidence approach undertaken by the EAG was associated with very high levels of uncertainty arising from both paucity of evidence and methodological challenges in modelling the long causal chain between test accuracy and clinical and economic outcomes. Bearing these caveats in mind, the EAG's assessment suggested that, for the 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 taken into account, AI-assisted CT reading is dominated by unaided reading. This is driven by the costs and disutilities associated with false-positive results and CT surveillance. In the screening population, AI-assisted CT image analysis was cost-effective in the base case and all sensitivity and scenario analysis. This was driven by a more favourable profile of model inputs, including an estimate of apparently improved test specificity for AI (the evidence of which was from a single study). Sensitivity and scenario analyses showed that the impact of AI assistance on radiologists' reporting time, prevalence of lung nodules and disutility associated with CT surveillance is likely to be an important factor, in addition to accuracy in nodule detection, in driving cost-effectiveness.

## Implications for service provision

Current evidence concerning the use of AI software to assist radiologists' detection and analysis of lung nodules that is directly applicable to the UK NHS is very limited, although this is an area of active research and further evidence will become available in the coming years. Based on the findings from our assessment, potential implications for service provision include:

- The availability of evidence on test accuracy varies substantially between different technologies, and direct evidence on clinical effectiveness and cost-effectiveness evidence is lacking. Potential adoption of these technologies will need to consider uncertainties associated with quality, quantity and applicability of available evidence on individual technologies in addition to their functionality, relevant costs and costing structure. Further research to generate evidence may be needed to inform decisions about the adoption of these technologies.
- Furthermore, the practical impact of incorporating these technologies into clinical practice, such as their impact on radiologists' reporting time, may need to be evaluated through pilot testing.
- Current evidence indicates a possibility of increased demand for CT surveillance with the adoption of AI-assisted image analysis. The potential impact on costs and service organisation needs to be carefully considered.
- Most technologies undergo regular update, which may involve changes in AI-derived algorithms. An ongoing audit of the potential impact of these updates on test accuracy and service provision may be desirable.

## Suggested research priorities

Published studies have largely been conducted retrospectively in a research environment. The vast majority of studies identified in this DAR were judged to be at high risk of bias and have multiple applicability concerns for the UK setting. No prospective studies evaluating intermediate clinical process and downstream clinical outcomes were identified. Further prospective studies of the use of software derived from AI algorithms to aid chest CT image analysis that adopts per-person analysis for estimating test accuracy, incorporates clinical process and outcome measures, and are undertaken in clinical practice settings are required.

Additional areas of interest that may influence clinical practice include:

- Does the accuracy of AI-assisted chest CT image analysis vary by specialty and experience of readers and reasons for chest CT scans?
- Does the accuracy of AI-assisted chest CT image analysis differ between symptomatic, incidental and screening populations?
- What is the impact of using AI software to assist chest CT image analysis on radiologists' reporting time in clinical practice?
- More precise quantification of potential harm associated with CT surveillance, including potential disutility incurred associated with anxiety during surveillance and effect of exposure to radiation.
- Comparison of accuracy for lung cancer detection based on unaided reading or AI-assisted reading and current clinical guidelines versus nodule management strategy based on cancer risk prediction informed by AI-derived algorithms.

Value-of-information analyses could be conducted to help prioritise required research. Further methodological research that may be needed include:

- Establishing and validating frameworks for linking test accuracy evidence to clinical and economic outcomes to facilitate evaluation of emerging and evolving AI software for chest CT scan analysis and other similar technologies.
- Establishing and validating frameworks for evaluating the cost-effectiveness of AI software capable of analysing chest CT scans for multiple clinical indications (in addition to lung nodule detection and analysis).

# Additional information

## Contributions of authors

**Julia Geppert** (**https://orcid.org/0000-0001-6446-6094**) (Research Fellow) performed the test accuracy and clinical effectiveness reviews and wrote associated sections of this report.

**Peter Auguste** (**https://orcid.org/0000-0001-5143-3218**) (Assistant Professor, Health Economics) led the cost-effectiveness components of this project, undertook the systematic review of the health economic literature, constructed the health economic models and wrote the economics sections of this report.

**Asra Asgharzadeh** (**https://orcid.org/0000-0002-1068-8537**) (Research Fellow) performed the test accuracy and clinical effectiveness reviews and wrote associated sections of this report.

**Hesam Ghiasvand** (**https://orcid.org/0000-0002-3110-6954**) (Research Fellow, Health Economics) performed the systematic review of the health economic literature, undertook health economic modelling and wrote the economics sections of this report.

**Mubarak Patel** (**https://orcid.org/0000-0001-7573-1447**) (Research Fellow, Medical Statistics) performed statistical analyses and wrote associated sections of the report.

**Anna Brown** (**https://orcid.org/0000-0002-4541-6232**) (Information Specialist) developed the search strategies, undertook searches, managed references and wrote the search methods sections of this report.

**Surangi Jayakody** (**https://orcid.org/0000-0001-7995-391X**) (Visiting Research Fellow) performed a review of the literature on overdiagnosis and supported the clinical effectiveness review and report writing of associated sections.

**Emma Helm** (**https://orcid.org/0000-0002-2194-0929**) (Consultant Radiologist) provided expert clinical advice and helped to develop the economic models.

**Dan Todkill** (**https://orcid.org/0000-0002-4325-4786**) (Associate Clinical Professor, Public Health Medicine) commented on earlier versions of the report and assisted in revising the report.

**Jason Madan** (**https://orcid.org/0000-0003-4316-1480**) (Professor, Health Economics) provided methodological advice on economic modelling and revised associated sections of the report.

**Chris Stinton** (**https://orcid.org/0000-0001-9054-1940**) (Senior Research Fellow) provided methodological advice and training for the clinical effectiveness team, acted as third reviewer and assisted in revising the report.

**Daniel Gallacher** (**https://orcid.org/0000-0003-0506-9384**) (Assistant Professor, Medical Statistics and Health Technology Assessment) provided statistical advice on simulation and assisted in revising the report.

**Sian Taylor-Phillips** (**https://orcid.org/0000-0002-1841-4346**) (Professor, Population Health) contributed to study design and protocol, acted as senior advisor, provided methodological support and assisted in revising the report.

**Yen-Fu Chen** (**https://orcid.org/0000-0002-9446-2761**) (Associate Professor, Evidence Synthesis) led the project, its co-ordination and implementation and write-up.

## Acknowledgements

Professor Ben Glocker (Professor in Machine Learning for Imaging, Imperial College London) provided expert technical advice to the EAG during the preparation of the report.

Professor Charles Hutchinson (Professor of Clinical Imaging, CSRL, University Hospitals Coventry and Warwickshire) provided expert clinical advice to the EAG during the preparation of the report.

Professor Aileen Clarke helped with the revision of various summary sections of the report.

We are grateful to the NICE Specialist Committee members for their advice during our preparation of the economic models.

We thank Sarah Abrahamson for her assistance in editing and proofreading the report, and Pearl Pawson for her administrative support.

## Data-sharing statement

We have included most of the data collected in this report, its appendices and supplementary materials. Additional data can be obtained from the corresponding author.

## Ethics statement

Ethical approval is not required for this piece of work as it is an evidence synthesis based on data that are either publicly available or submitted by the companies and that do not include any identifiable patient data.

## Information governance statement

This research does not involve handling of any personal information.

## Disclosure of interests

**Full disclosure of interests:** Completed ICMJE forms for all authors, including all related interests, are available in the toolkit on the NIHR Journals Library report publication page at https://doi.org/10.3310/JYTW8921.

**Primary conflicts of interest:** Dan Todkill is partly supported by the NIHR Applied Research Collaboration West Midlands. Sian Taylor-Phillips was partially funded by an National Institute for Health and Care Research (NIHR) Career Development Fellowship (CDF-2016-09-018) and an NIHR Research Professorship (NIHR302434) during the preparation of the report and is a member of the UK National Screening Committee and Chair of the UK National Screening Committee Research and Methodology Group.

# References

1. Geppert J, Auguste P, Asgharzadeh A, Brown A, Jayakody S, Ghiasvand H, *et al*. *Diagnostic Assessment Report Commissioned by the NIHR HTA Programme on Behalf of the National Institute for Health and Care Excellence – Final Protocol. Software with Artificial Intelligence Derived Algorithms for Automated Detection and Analysis of Lung Nodules from CT Scan Images [DAP60]* National Institute for Health and Care Research; 2021. URL: www.nice.org.uk/guidance/dg55/documents/final-protocol (accessed 3 October 2023).
2. Larici AR, Farchione A, Franchi P, Ciliberto M, Cicchetti G, Calandriello L, *et al*. Lung nodules: size still matters. *Eur Respir Rev* 2017;**26**:170025. https://doi.org/10.1183/16000617.0025-2017
3. Bankier AA, MacMahon H, Goo JM, Rubin GD, Schaefer-Prokop CM, Naidich DP. Recommendations for measuring pulmonary nodules at CT: a statement from the Fleischner Society. *Radiology* 2017;**285**:584–600. https://doi.org/10.1148/radiol.2017162894
4. Horeweg N, van Rosmalen J, Heuvelmans MA, van der Aalst CM, Vliegenthart R, Scholten ET, *et al*. Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening. *Lancet Oncol* 2014;**15**:1332–41. https://doi.org/10.1016/s1470-2045(14)70389-4
5. Cancer Research UK. *Lung Cancer Incidence Statistics*. URL: www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence (accessed 4 October 2022).
6. National Disease Registration Service. *Staging Data in England*. Public Health England; 2020. URL: www.cancerdata.nhs.uk/stage_at_diagnosis (accessed 5 October 2022).
7. NHS. *NHS Long Term Plan*. URL: www.longtermplan.nhs.uk/ (accessed 5 October 2022).
8. National Institute for Health and Care Excellence (NICE). *Suspected Cancer: Recognition and Referral.* NICE Guideline [NG12]. NICE; 2015. URL: www.nice.org.uk/guidance/ng12 (accessed 5 October 2022).
9. National Institute for Health and Care Excellence (NICE). *Lung Cancer: Diagnosis and Management*. NICE Guideline [NG122]. NICE; 2019. URL: www.nice.org.uk/guidance/ng122 (accessed 5 October 2022).
10. UK National Screening Committee. *Adult Screening Programme: Lung Cancer*. URL: https://view-health-screening-recommendations.service.gov.uk/lung-cancer/ (accessed 6 October 2022).
11. NHS England. *Evaluation of the Targeted Lung Health Check Programme*. URL: www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/our-services/evaluation-of-the-targeted-lung-health-check-programme/ (accessed 5 October 2022).
12. Callister ME, Baldwin DR, Akram AR, Barnard S, Cane P, Draffan J, *et al*.; British Thoracic Society Pulmonary Nodule Guideline Development Group. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax* 2015;**70**:ii1–54. https://doi.org/10.1136/thoraxjnl-2015-207168
13. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, *et al*. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med* 2013;**369**:910–9. https://doi.org/10.1056/NEJMoa1214726
14. British Thoracic Society. *PN Risk Calculator*. URL: www.brit-thoracic.org.uk/quality-improvement/guidelines/pulmonary-nodules/pn-risk-calculator/ (accessed 5 October 2022).
15. Herder GJ, van Tinteren H, Golding RP, Kostense PJ, Comans EF, Smit EF, Hoekstra OS. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. *Chest* 2005;**128**:2490–6. https://doi.org/10.1378/chest.128.4.2490
16. British Thoracic Society. *BTS Guidelines for the Investigation and Management of Pulmonary Nodules*. URL: www.brit-thoracic.org.uk/quality-improvement/guidelines/pulmonary-nodules/ (accessed 1 December 2021).

17. American College of Radiology. *Lung CT Screening Reporting & Data System (Lung-RADS). Version 1.1*. 2019. URL: www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads (accessed 24 October 2022).

18. NHS England National Cancer Programme. *Targeted Screening for Lung Cancer with Low Radiation Dose Computed Tomography: Standard Protocol Prepared for The NHS England Targeted Lung Health Checks Programme. Version 2*. NHS; 2022. URL: www.england.nhs.uk/wp-content/uploads/2019/02/B1646-standard-protocol-targeted-lung-health-checks-programme-v2.pdf (accessed 5 December 2022).

19. British Society of Thoracic Imaging, Royal College of Radiologists. *Considerations to Ensure Optimum Roll-out of Targeted Lung Cancer Screening Over the Next Five Years*. Royal College of Radiologists; 2020. URL: www.rcr.ac.uk/posts/considerations-ensure-optimum-roll-out-targeted-lung-cancer-screening-over-next-five-years (accessed 5 October 2022).

20. HM Government. *National AI Strategy*. Office for Artificial Intelligence; 2021. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf (accessed 5 October 2022).

21. National Institute for Health and Care Excellence (NICE). *Software with Artificial Intelligence Derived Algorithms for Automated Detection and Analysis of Lung Nodules from CT Scan Images: Final Scope*. 2021. URL: www.nice.org.uk/guidance/dg55/documents/final-scope (accessed 24 June 2024).

22. National Institute for Health and Care Excellence (NICE). *Diagnostics Assessment Programme Manual*. Manchester: NICE; 2011. URL: www.nice.org.uk/media/default/about/what-we-do/nice-guidance/nice-diagnostics-guidance/diagnostics-assessment-programme-manual.pdf (accessed 5 October 2022).

23. Cochrane Screening and Diagnostic Test Methods Group. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. Cochrane Collaboration; 2013. URL: https://methods.cochrane.org/sdt/handbook-dta-reviews (accessed 5 October 2022).

24. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, *et al.*; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

25. Yang B, Mallett S, Takwoingi Y, Davenport C, Hyde C, Whiting P, *et al*. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Ann Intern Med* 2021;**174**:1592–9. https://doi.org/10.7326/m21-2234

26. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, *et al*. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol* 2020;**20**:293. https://doi.org/10.1186/s12874-020-01179-5

27. Hall H, Ruparel M, Quaife SL, Dickson JL, Horst C, Tisi S, *et al*. The role of computer-assisted radiographer reporting in lung cancer screening programmes. *Eur Radiol* 2022;**32**:6891–9. https://doi.org/10.1007/s00330-022-08824-1

28. Murchison J, Ritchie G, Senszak D, Van Beek EJR. *Evaluation of Deep Learning Software Tool for CT Based Lung Nodule Growth Assessment*. Paper presented at European Congress of Radiology 2019, Vienna, Austria, 27 February–3 March 2019. URL: https://epos.myesr.org/poster/esr/ecr2019/C-3685

29. Murchison J, Ritchie G, Senszak D, Van Beek EJR. *Evaluation of Deep Learning Software Tool for CT Based Lung Nodule Growth Segmentation*. Paper presented at European Congress of Radiology 2019, Vienna, Austria, 27 February–3 March 2019. URL: https://epos.myesr.org/poster/esr/ecr2019/C-3686

30. Wakkie J, Doorn L. *Clinical Evaluation Report: Veye Lung Nodules (and Veye Chest 2.15). LN-CER-003. Document Version 3.1. Company submission*. Amsterdam: Aidence; 2020.

31. Röhrich S, Heidinger BH, Prayer F, Weber M, Krenn M, Zhang R, *et al*. Impact of a content-based image retrieval system on the interpretation of chest CTs of patients with diffuse parenchymal lung disease. *Eur Radiol* 2023;**33**:360–7. https://doi.org/10.1007/s00330-022-08973-3

32. Lancaster HL, Zheng S, Aleshina OO, Yu D, Yu Chernina V, Heuvelmans MA, *et al*. Outstanding negative prediction performance of solid pulmonary nodule volume AI for ultra-LDCT baseline lung cancer screening risk stratification. *Lung Cancer* 2022;**165**:133–40. https://doi.org/10.1016/j.lungcan.2022.01.002

33. Murchison JT, Ritchie G, Senyszak D, Nijwening JH, van Veenendaal G, Wakkie J, van Beek EJR. Validation of a deep learning computer aided system for CT based lung nodule detection, classification, and growth rate estimation in a routine clinical population. *PLOS ONE* 2022;**17**:e0266799. https://doi.org/10.1371/journal.pone.0266799

34. Hempel HL, Engbersen MP, Wakkie J, van Kelckhoven BJ, de Monyé W. Higher agreement between readers with deep learning CAD software for reporting pulmonary nodules on CT. *Eur J Radiol Open* 2022;**9**:100435. https://doi.org/10.1016/j.ejro.2022.100435

35. Hall H, Ruparel M, Quaife S, Dickson JL, Horst C, Tisi S, *et al*. P78. The role of computer-assisted radiographer reporting in lung cancer screening programmes. *Thorax* 2019;**74**:A131. https://doi.org/10.1136/thorax-2019-BTSabstracts2019.221

36. Ahn Y, Lee SM, Noh HN, Kim W, Choe J, Do KH, Seo JB. Use of a commercially available deep learning algorithm to measure the solid portions of lung cancer manifesting as subsolid lesions at CT: comparisons with radiologists and invasive component size at pathologic examination. *Radiology* 2021;**299**:202–10. https://doi.org/10.1148/radiol.2021202803

37. Cohen JG, Goo JM, Yoo RE, Park CM, Lee CH, van Ginneken B, *et al*. Software performance in segmenting ground-glass and solid components of subsolid nodules in pulmonary adenocarcinomas. *Eur Radiol* 2016;**26**:4465–74. https://doi.org/10.1007/s00330-016-4317-3

38. Cohen JG, Goo JM, Yoo RE, Park SB, van Ginneken B, Ferretti GR, *et al*. The effect of late-phase contrast enhancement on semi-automatic software measurements of CT attenuation and volume of part-solid nodules in lung adenocarcinomas. *Eur J Radiol* 2016;**85**:1174–80. https://doi.org/10.1016/j.ejrad.2016.03.027

39. Garzelli L, Goo JM, Ahn SY, Chae KJ, Park CM, Jung J, Hong H. Improving the prediction of lung adenocarcinoma invasive component on CT: value of a vessel removal algorithm during software segmentation of subsolid nodules. *Eur J Radiol* 2018;**100**:58–65. https://doi.org/10.1016/j.ejrad.2018.01.016

40. Martini K, Blüthgen C, Eberhard M, Schönenberger ALN, De Martini I, Huber FA, *et al*. Impact of vessel suppressed-CT on diagnostic accuracy in detection of pulmonary metastasis and reading time. *Acad Radiol* 2021;**28**:988–94. https://doi.org/10.1016/j.acra.2020.01.014

41. Park S, Park G, Lee SM, Kim W, Park H, Jung K, Seo JB. Deep learning-based differentiation of invasive adenocarcinomas from preinvasive or minimally invasive lesions among pulmonary subsolid nodules. *Eur Radiol* 2021;**31**:6239–47. https://doi.org/10.1007/s00330-020-07620-z

42. Hu Q, Chen C, Kang S, Sun Z, Wang Y, Xiang M, *et al*. Application of computer-aided detection (CAD) software to automatically detect nodules under SDCT and LDCT scans with different parameters. *Comput Biol Med* 2022;**146**:105538. https://doi.org/10.1016/j.compbiomed.2022.105538

43. Wagner AK, Hapich A, Psychogios MN, Teichgräber U, Malich A, Papageorgiou I. Computer-aided detection of pulmonary nodules in computed tomography using ClearReadCT. *J Med Syst* 2019;**43**:58. https://doi.org/10.1007/s10916-019-1180-1

44. Yacoub B, Kabakus IM, Schoepf UJ, Giovagnoli VM, Fischer AM, Wichmann JL, *et al*. Performance of an artificial intelligence-based platform against clinical radiology reports for the evaluation of noncontrast chest CT. *Acad Radiol* 2022;**29**:S108–17. https://doi.org/10.1016/j.acra.2021.02.007

45. Li K, Liu K, Zhong Y, Liang M, Qin P, Li H, *et al*. Assessing the predictive accuracy of lung cancer, metastases, and benign lesions using an artificial intelligence-driven computer aided diagnosis system. *Quant Imaging Med Surg* 2021;**11**:3629–42. https://doi.org/10.21037/qims-20-1314

46. Wang Y, Yan F, Lu X, Zheng G, Zhang X, Wang C, *et al*. IILS: intelligent imaging layout system for automatic imaging report standardization and intra-interdisciplinary clinical workflow optimization. *EBioMedicine* 2019;**44**:162–81. https://doi.org/10.1016/j.ebiom.2019.05.040

47. Abadia AF, Yacoub B, Stringer N, Snoddy M, Kocher M, Schoepf UJ, *et al*. Diagnostic accuracy and performance of artificial intelligence in detecting lung nodules in patients with complex lung disease: a noninferiority study. *J Thorac Imaging* 2021;**37**:154–61. https://doi.org/10.1097/RTI.0000000000000613

48. Chamberlin J, Kocher MR, Waltz J, Snoddy M, Stringer NFC, Stephenson J, *et al*. Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose CT scans for lung cancer screening: accuracy and prognostic value. *BMC Med* 2021;**19**:55. https://doi.org/10.1186/s12916-021-01928-3

49. Rückel J, Sperl JI, Kaestle S, Hoppe BF, Fink N, Rudolph J, *et al*. Reduction of missed thoracic findings in emergency whole-body computed tomography using artificial intelligence assistance. *Quant Imaging Med Surg* 2021;**11**:2486–98. https://doi.org/10.21037/qims-20-1037

50. Hwang EJ, Goo JM, Kim HY, Yi J, Kim Y. Optimum diameter threshold for lung nodules at baseline lung cancer screening with low-dose chest CT: exploration of results from the Korean Lung Cancer Screening Project. *Eur Radiol* 2021;**31**:7202–12. https://doi.org/10.1007/s00330-021-07827-8

51. Hwang EJ, Goo JM, Kim HY, Yi J, Yoon SH, Kim Y. Implementation of the cloud-based computerized interpretation system in a nationwide lung cancer screening with low-dose CT: comparison with the conventional reading system. *Eur Radiol* 2021;**31**:475–85. https://doi.org/10.1007/s00330-020-07151-7

52. Hwang EJ, Goo JM, Kim HY, Yoon SH, Jin GY, Yi J, Kim Y. Variability in interpretation of low-dose chest CT using computerized assessment in a nationwide lung cancer screening program: comparison of prospective reading at individual institutions and retrospective central reading. *Eur Radiol* 2021;**31**:2845–55. https://doi.org/10.1007/s00330-020-07424-1

53. Hsu HH, Ko KH, Wu YC, Chiu SH, Chang CK, Chang WC, *et al*. Performance and reading time of lung nodule identification on multidetector CT with or without an artificial intelligence-powered computer-aided detection system. *Clin Radiol* 2021;**76**:626. https://doi.org/10.1016/j.crad.2021.04.006

54. Lo SB, Freedman MT, Gillis LB, White CS, Mun SK. Journal club: computer-aided detection of lung nodules on ct with a computerized pulmonary vessel suppressed function. *AJR Am J Roentgenol* 2018;**210**:480–8. https://doi.org/10.2214/AJR.17.18718

55. Milanese G, Eberhard M, Martini K, Vittoria De Martini I, Frauenfelder T. Vessel suppressed chest computed tomography for semi-automated volumetric measurements of solid pulmonary nodules. *Eur J Radiol* 2018;**101**:97–102. https://doi.org/10.1016/j.ejrad.2018.02.020

56. Singh R, Kalra MK, Homayounieh F, Nitiwarangkul C, McDermott S, Little BP, *et al*. Artificial intelligence-based vessel suppression for detection of sub-solid nodules in lung cancer screening computed tomography. *Quant Imaging Med Surg* 2021;**11**:1134–43. https://doi.org/10.21037/qims-20-630

57. Takaishi T, Ozawa Y, Bando Y, Yamamoto A, Okochi S, Suzuki H, Shibamoto Y. Incorporation of a computer-aided vessel-suppression system to detect lung nodules in CT images: effect on sensitivity and reading time in routine clinical settings. *Jpn J Radiol* 2021;**39**:159–64. https://doi.org/10.1007/s11604-020-01043-y

58. Wan YL, Wu PW, Huang PC, Tsay PK, Pan KT, Trang NN, *et al*. The use of artificial intelligence in the differentiation of malignant and benign lung nodules on computed tomograms proven by surgical pathology. *Cancers (Basel)* 2020;**12**:2211. https://doi.org/10.3390/cancers12082211

59. Kozuka T, Matsukubo Y, Kadoba T, Oda T, Suzuki A, Hyodo T, *et al*. Efficiency of a computer-aided diagnosis (CAD) system with deep learning in detection of pulmonary nodules on 1-mm-thick images of computed tomography. *Jpn J Radiol* 2020;**38**:1052–61. https://doi.org/10.1007/s11604-020-01009-0

60. Liu K, Li Q, Ma J, Zhou Z, Sun M, Deng Y, *et al*. Evaluating a fully automated pulmonary nodule detection approach and its impact on radiologist performance. *Radiol Artif Intell* 2019;**1**:e180084. https://doi.org/10.1148/ryai.2019180084

61. Zhang Y, Jiang B, Zhang L, Greuter MJW, de Bock GH, Zhang H, *et al*. Lung nodule detectability of artificial intelligence-assisted CT image reading in lung cancer screening. *Curr Med Imaging* 2021;**18**:327–34. https://doi.org/10.2174/1573405617666210806125953

62. Cohen JG, Kim H, Park SB, van Ginneken B, Ferretti GR, Lee CH, *et al*. Comparison of the effects of model-based iterative reconstruction and filtered back projection algorithms on software measurements in pulmonary subsolid nodules. *Eur Radiol* 2017;**27**:3266–74. https://doi.org/10.1007/s00330-016-4716-5

63. Kim H, Park CM, Hwang EJ, Ahn SY, Goo JM. Pulmonary subsolid nodules: value of semi-automatic measurement in diagnostic accuracy, diagnostic reproducibility and nodule classification agreement. *Eur Radiol* 2018;**28**:2124–33. https://doi.org/10.1007/s00330-017-5171-7

64. Jacobs C, Schreuder A, van Riel SJ, Scholten ET, Wittenberg R, Wille MMW, *et al*. Assisted versus manual interpretation of low-dose CT scans for lung cancer screening: impact on lung-RADS agreement. *Radiol Imaging Cancer* 2021;**3**:e200160. https://doi.org/10.1148/rycan.2021200160

65. Blazis SP, Dickerscheid DBM, Linsen PVM, Martins Jarnalo CO. Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system. *Eur J Radiol* 2021;**136**:109526. https://doi.org/10.1016/j.ejrad.2021.109526

66. Martins Jarnalo CO, Linsen PVM, Blazís SP, van der Valk PHM, Dickerscheid DBM. Clinical evaluation of a deep-learning-based computer-aided detection system for the detection of pulmonary nodules in a large teaching hospital. *Clin Radiol* 2021;**76**:838–45. https://doi.org/10.1016/j.crad.2021.07.012

67. Park S, Park H, Lee SM, Ahn Y, Kim W, Jung K, Seo JB. Application of computer-aided diagnosis for Lung-RADS categorization in CT screening for lung cancer: effect on inter-reader agreement. *Eur Radiol* 2022;**32**:1054–64. https://doi.org/10.1007/s00330-021-08202-3

68. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, *et al*. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology* 2017;**284**:228–43. https://doi.org/10.1148/radiol.2017161659

69. Singh R, Nitiwarangkul C, Shepard JAO, Homayounieh F, Padole A, McDermott S, *et al*. *Effect of Artificial Intelligence Based Vessel Suppression and Automatic Detection of Part-solid and Ground-glass Nodules on Low-dose Chest CT*. Paper presented at Radiological Society of North America, 104th Scientific Assembly and Annual Meeting, Chicago, IL, USA, 25–30 November 2018. URL: https://archive.rsna.org/2018/ScienceSessions.pdf (accessed 1 February 2022).

70. Naidich DP, Bankier AA, MacMahon H, Schaefer-Prokop CM, Pistolesi M, Goo JM, *et al*. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology* 2013;**266**:304–17. https://doi.org/10.1148/radiol.12120628

71. Treskova M, Aumann I, Golpon H, Vogel-Claussen J, Welte T, Kuhlmann A. Trade-off between benefits, harms and economic efficiency of low-dose CT lung cancer screening: a microsimulation analysis of nodule management strategies in a population-based setting. *BMC Med* 2017;**15**:162. https://doi.org/10.1186/s12916-017-0924-3

72. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, *et al*. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *Eur J Health Econ* 2013;**14**:367–72. https://doi.org/10.1007/s10198-013-0471-6

73. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, *et al*. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* 2004;**8**:iii–iv. https://doi.org/10.3310/hta8360

74. Bajre MK, Pennington M, Woznitza N, Beardmore C, Radhakrishnan M, Harris R, McCrone P. Expanding the role of radiographers in reporting suspected lung cancer: a cost-effectiveness analysis using a decision tree model. *Radiography* 2017;**23**:273–8. https://doi.org/10.1016/j.radi.2017.07.011

75. Adams SJ, Mondal P, Penz E, Tyan CC, Lim H, Babyn P. Development and cost analysis of a lung nodule management strategy combining artificial intelligence and Lung-RADS for baseline lung cancer screening. *J Am Coll Radiol* 2021;**18**:741–51. https://doi.org/10.1016/j.jacr.2020.11.014

76. Hunink M, Glasziou P. *Decision Making in Health and Medicine, Integrating Evidence and Values*. Cambridge: Cambridge University Press; 2001.

77. Ruparel M, Quaife SL, Dickson JL, Horst C, Tisi S, Hall H, *et al*. Lung screen uptake trial: results from a single lung cancer screening round. *Thorax* 2020;**75**:908–12. https://doi.org/10.1136/thoraxjnl-2020-214703

78. Jones K, Burns A. *Unit Costs of Health and Social Care 2021*. Canterbury: Personal Social Services Research Unit, University of Kent; 2021.

79. Steele JD, Buell P. Asymptomatic solitary pulmonary nodules: host survival, tumor size, and growth rate. *J Thorac Cardiovasc Surg* 1973;**65**:140–51.

80. Field JK, Duffy SW, Baldwin DR, Whynes DK, Devaraj A, Brain KE, *et al*. UK Lung Cancer RCT pilot screening trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. *Thorax* 2016;**71**:161–70. https://doi.org/10.1136/thoraxjnl-2015-207140

81. Birtwistle M, Earnshaw A. *Saving Lives, Averting Costs: An Analysis of the Financial Implications of Achieving Earlier Diagnosis of Colorectal, Lung and Ovarian Cancer*. London: Incisive Health, Cancer Research UK; 2014.

82. Hiten N, Howell D, Su J, Qiu X, Brown C, Vennettilli A, *et al*. Stage specific health utility index scores of Canadian cancer patients. *J Clin Oncol* 2015;**33**:6614.

83. Rickets W, Lau KKW, Pollit V, Mealing S, Leonard C, Mallender P, *et al*. Exploratory cost-effectiveness model of electromagnetic navigation bronchoscopy (ENB) compared with CT-guided biopsy (TTNA) for diagnosis of malignant indeterminate peripheral pulmonary nodules. *BMJ Open Respir Res* 2020;**7**:e000595. https://doi.org/10.1136/bmjresp-2020-000595

84. Jacobs DR Jr, Adachi H, Mulder I, Kromhout D, Menotti A, Nissinen A, Blackburn H. Cigarette smoking and mortality risk: twenty-five-year follow-up of the Seven Countries Study. *Arch Intern Med* 1999;**159**:733–40. https://doi.org/10.1001/archinte.159.7.733

85. Sutton AJ, Sagoo GS, Jackson L, Fisher M, Hamilton-Fairley G, Murray A, Hill A. Cost-effectiveness of a new autoantibody test added to computed tomography (CT) compared to CT surveillance alone in the diagnosis of lung cancer amongst patients with indeterminate pulmonary nodules. *PLOS ONE* 2020;**15**:e0237492. https://doi.org/10.1371/journal.pone.0237492

86. Stevenson M, Lloyd-Jones M, Morgan MY, Wong R. Non-invasive diagnostic assessment tools for the detection of liver fibrosis in patients with suspected alcohol-related liver disease: a systematic review and economic evaluation. *Health Technol Assess* 2012;**16**:1–174. https://doi.org/10.3310/hta16040

87. Lancaster HL, Heuvelmans MA, Pelgrim GJ, Rook M, Kok MGJ, Aown A, *et al*. Seasonal prevalence and characteristics of low-dose CT detected lung nodules in a general Dutch population. *Sci Rep* 2021;**11**:9139. https://doi.org/10.1038/s41598-021-88328-y

88. Davies L, Petitti DB, Martin L, Woo M, Lin JS. Defining, estimating, and communicating overdiagnosis in cancer screening. *Ann Intern Med* 2018;**169**:36–43. https://doi.org/10.7326/m18-0694

89. Armato SG 3rd, Roberts RY, Kocherginsky M, Aberle DR, Kazerooni EA, Macmahon H, *et al*. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of ‘truth’. *Acad Radiol* 2009;**16**:28–38. https://doi.org/10.1016/j.acra.2008.05.022

90. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, *et al*. The ‘laboratory’ effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008;**249**:47–53. https://doi.org/10.1148/radiol.2491072025

91. Oudkerk M, Devaraj A, Vliegenthart R, Henzler T, Prosch H, Heussel CP, *et al*. European position statement on lung cancer screening. *Lancet Oncol* 2017;**18**:e754–66. https://doi.org/10.1016/s1470-2045(17)30861-6

92. Oudkerk M, Liu S, Heuvelmans MA, Walter JE, Field JK. Lung cancer LDCT screening and mortality reduction – evidence, pitfalls and future perspectives. *Nat Rev Clin Oncol* 2021;**18**:135–51. https://doi.org/10.1038/s41571-020-00432-6

93. Gould MK, Sanders GD, Barnett PG, Rydzak CE, Maclean CC, McClellan MB, Owens DK. Cost-effectiveness of alternative management strategies for patients with solitary pulmonary nodules. *Ann Intern Med* 2003;**138**:724–35.

94. Edelsberg J, Weycker D, Atwood M, Hamilton-Fairley G, Jett JR. Cost-effectiveness of an autoantibody test (EarlyCDT-Lung) as an aid to early diagnosis of lung cancer in patients with incidentally detected pulmonary nodules. *PLOS ONE* 2018;**13**:e0197826.

95. Chen X, Foy M, Kimmel M, Gorlova OY. Modeling the natural history and detection of lung cancer based on smoking behavior. *PLOS ONE* 2014;**9**:e93430.

96. Lin RS, Plevritis SK. Comparing the benefits of screening for breast cancer and lung cancer using a novel natural history model. *Cancer Cause Control* 2012 Jan;**23**:175–85. https://doi.org/10.1007/s10552-011-9866-9

97. Heuvelmans MA, Vliegenthart R, de Koning HJ, Groen HJ, van Putten MJ, Yousaf-Khan U, *et al*. Quantification of growth patterns of screen-detected lung cancers: the NELSON study. *Lung Cancer* 2017 Jun 1;**108**:48–54. https://doi.org/10.1016/j.lungcan.2017.02.021

98. Veronesi G, Bellomi M, Mulshine JL, Pelosi G, Scanagatta P, Paganelli G, *et al*. Lung cancer screening with low-dose computed tomography: a non-invasive diagnostic protocol for baseline lung nodules. *Lung Cancer* 2008;**61**:340–9. https://doi.org/10.1016/j.lungcan.2008.01.001

99. Wu MY, Li Y, Fu BJ, Wang GS, Chu ZG, Deng D. Evaluate the performance of four artificial intelligence-aided diagnostic systems in identifying and measuring four types of pulmonary nodules. *J Appl Clin Med Phys* 2021;**22**:318–26. https://doi.org/10.1002/acm2.13142

100. Xie X, Zhao Y, Snijder RA, van Ooijen PM, de Jong PA, Oudkerk M, *et al*. Sensitivity and accuracy of volumetry of pulmonary nodules on low-dose 16- and 64-row multi-detector CT: an anthropomorphic phantom study. *Eur Radiol* 2013;**23**:139–47. https://doi.org/10.1007/s00330-012-2570-7.

101. Kakinuma R, Noguchi M, Ashizawa K, Kuriyama K, Maeshima AM, Koizumi N, *et al*. Natural history of pulmonary subsolid nodules: a prospective multicenter study. *J Thorac Oncol* 2016;**11**:1012–28.

# Appendix 1 Supporting figures and tables

## Supporting figures

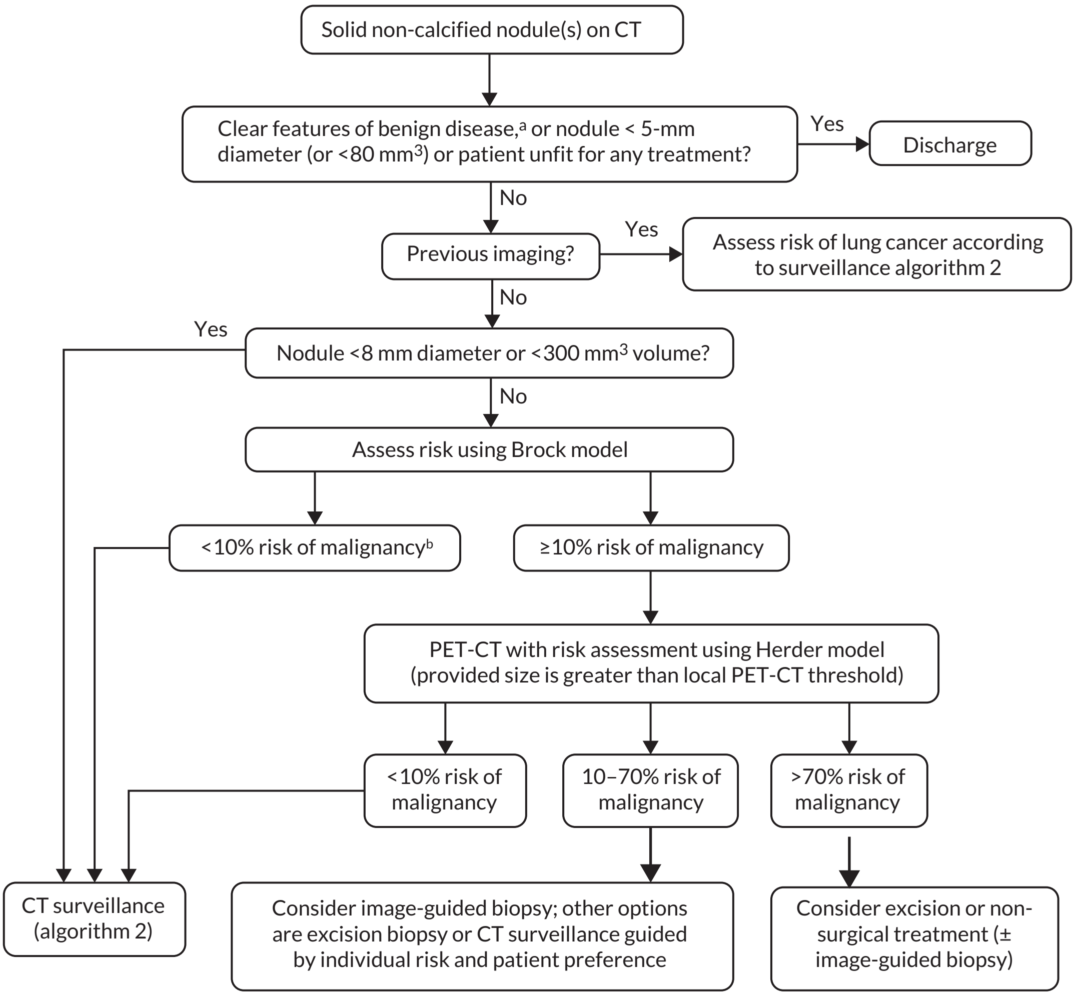

Solid non-calcified nodule(s) on CT
Clear features of benign disease,a or nodule < 5-mm diameter (or <80 mm3) or patient unfit for any treatment?
Yes
Discharge
No
Previous imaging?
Yes
Assess risk of lung cancer according to surveillance algorithm 2
No
Yes
Nodule <8 mm diameter or <300 mm3 volume?
No
Assess risk using Brock model
<10% risk of malignancyb
≥10% risk of malignancy
PET-CT with risk assessment using Herder model (provided size is greater than local PET-CT threshold)
<10% risk of malignancy
10–70% risk of malignancy
>70% risk of malignancy
CT surveillance (algorithm 2)
Consider image-guided biopsy; other options are excision biopsy or CT surveillance guided by individual risk and patient preference
Consider excision or non-surgical treatment (± image-guided biopsy)

**FIGURE 14** Initial assessment of solid lung nodules.[12] Reproduced from British Thoracic Society guidelines for the investigation and management of pulmonary nodules: accredited by NICE, Callister MEJ, Baldwin, DR, Akram AR, Barnard S, Cane P, Draffan J, *et al*., vol. 70, © 2015, with permission from BMJ Publishing Group Ltd.[12] a, Some nodules seen may be attached to or very near the lining of the lungs (perifissural nodules); these are often pulmonary lymph nodes. b, Consider PET-CT for larger nodules in young patients with low risk Brock score as this score was developed in screening cohort (50-75 years) so performance in younger patients is unproven.

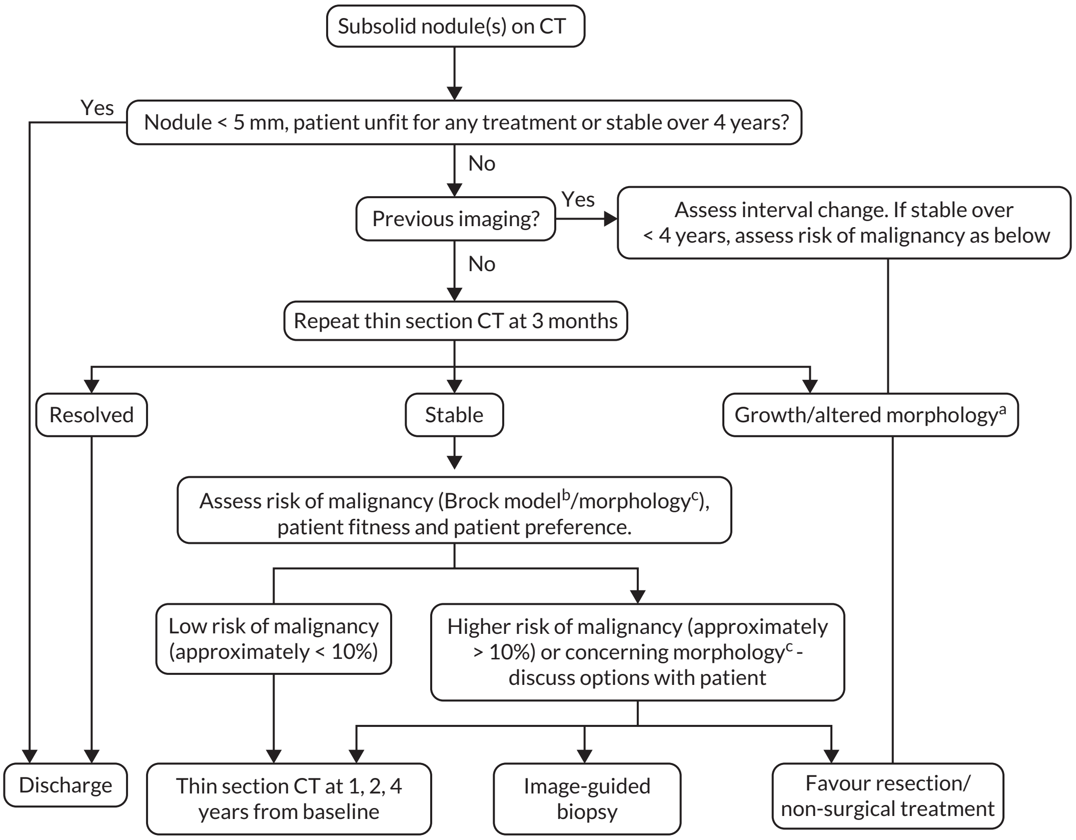

Subsolid nodule(s) on CT
Yes
Nodule < 5 mm, patient unfit for any treatment or stable over 4 years?
No
Previous imaging?
Yes
Assess interval change. If stable over < 4 years, assess risk of malignancy as below
No
Repeat thin section CT at 3 months
Resolved
Stable
Growth/altered morphology^a
Assess risk of malignancy (Brock model^b/morphology^c), patient fitness and patient preference.
Low risk of malignancy (approximately < 10%)
Higher risk of malignancy (approximately > 10%) or concerning morphology^c - discuss options with patient
Discharge
Thin section CT at 1, 2, 4 years from baseline
Image-guided biopsy
Favour resection/ non-surgical treatment

**FIGURE 15** Subsolid pulmonary nodules algorithm. Reproduced from British Thoracic Society guidelines for the investigation and management of pulmonary nodules: accredited by NICE, Callister MEJ, Baldwin, DR, Akram AR, Barnard S, Cane P, Draffan J, *et al*., vol. 70, © 2015, with permission from BMJ Publishing Group Ltd.[12] PSN, part-solid nodule; SSN, subsolid nodule. a, Change in mass/new solid component; b, Brock model may underestimate risk of malignancy in SSNs that persist at 3 months; c, Size of the solid component in PSNs, pleural indentation and bubble-like appearance.

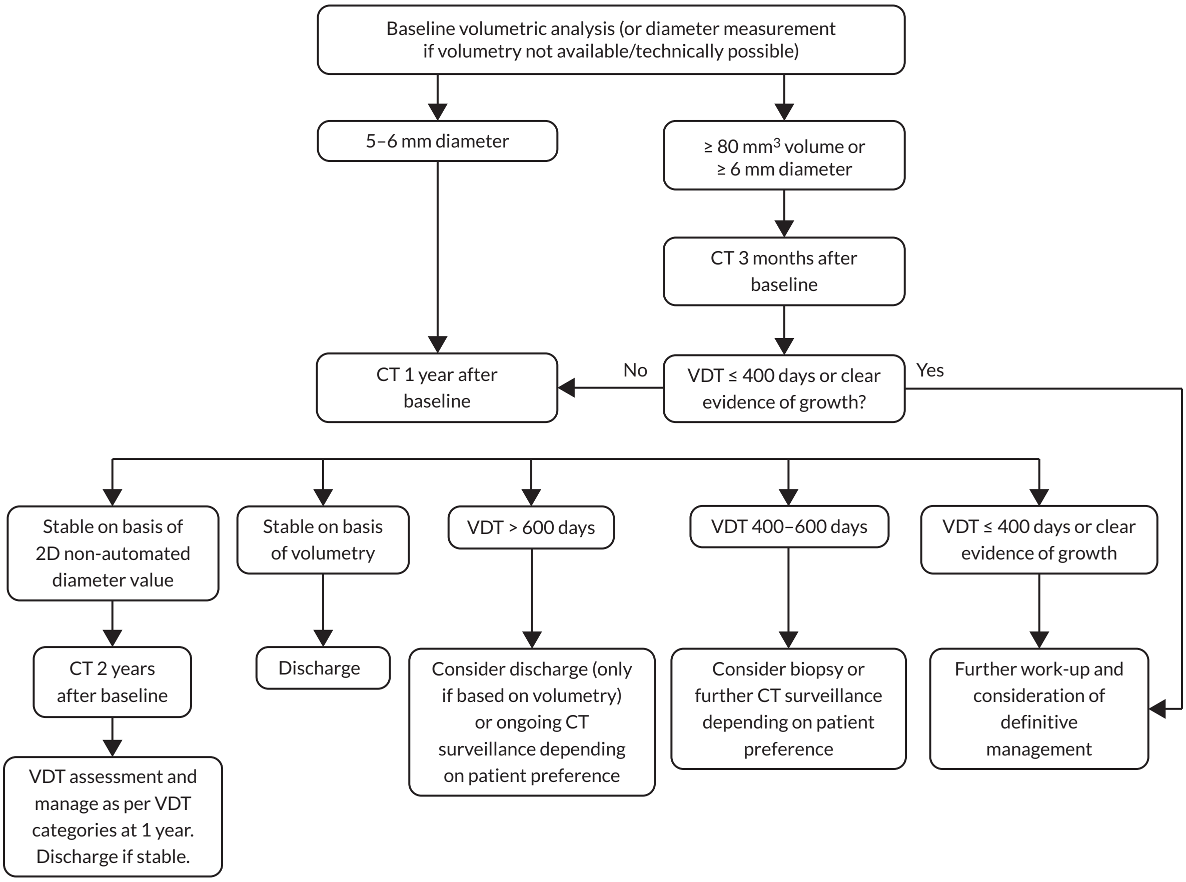

Baseline volumetric analysis (or diameter measurement if volumetry not available/technically possible)
5–6 mm diameter
≥ 80 mm³ volume or ≥ 6 mm diameter
CT 3 months after baseline
CT 1 year after baseline
No
VDT ≤ 400 days or clear evidence of growth?
Yes
Stable on basis of 2D non-automated diameter value
Stable on basis of volumetry
VDT > 600 days
VDT 400–600 days
VDT ≤ 400 days or clear evidence of growth
CT 2 years after baseline
Discharge
Consider discharge (only if based on volumetry) or ongoing CT surveillance depending on patient preference
Consider biopsy or further CT surveillance depending on patient preference
Further work-up and consideration of definitive management
VDT assessment and manage as per VDT categories at 1 year. Discharge if stable.

**FIGURE 16** Computed tomography surveillance of solid lung nodules. Reproduced from British Thoracic Society guidelines for the investigation and management of pulmonary nodules: accredited by NICE, Callister MEJ, Baldwin, DR, Akram AR, Barnard S, Cane P, Draffan J, *et al.*, vol. 70, © 2015, with permission from BMJ Publishing Group Ltd.[12]

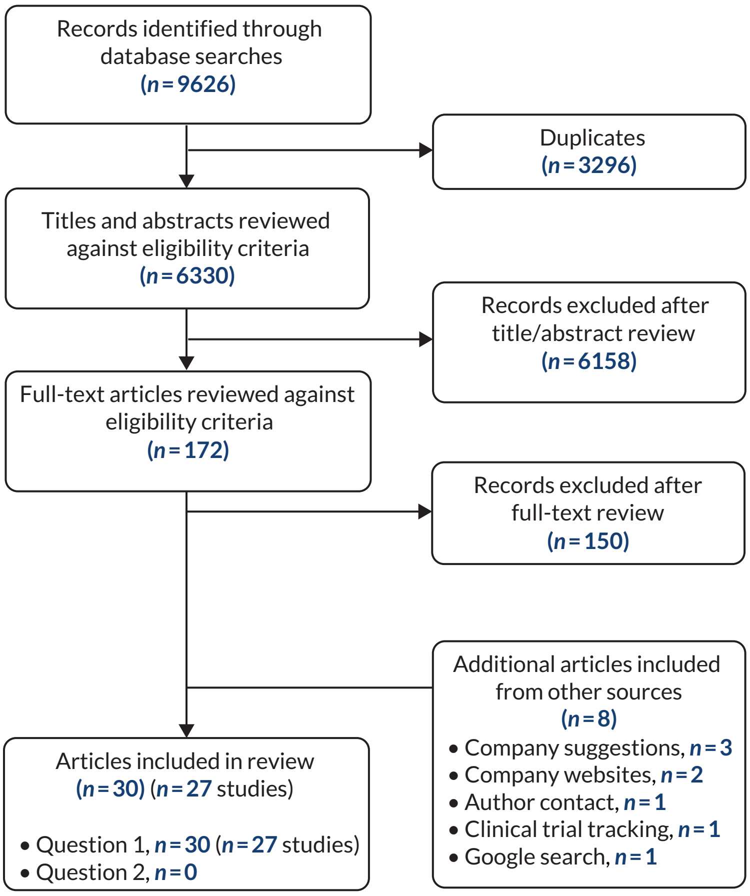
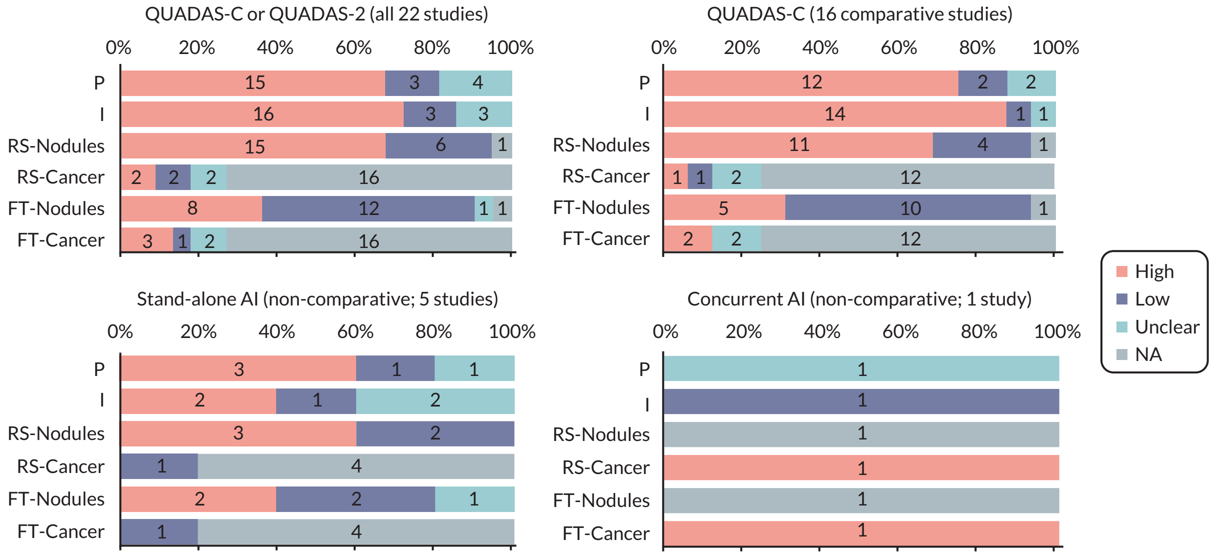

Records identified through database searches (n = 9626)
Duplicates (n = 3296)
Titles and abstracts reviewed against eligibility criteria (n = 6330)
Records excluded after title/abstract review (n = 6158)
Full-text articles reviewed against eligibility criteria (n = 172)
Records excluded after full-text review (n = 150)
Additional articles included from other sources (n = 8)
• Company suggestions, n = 3
• Company websites, n = 2
• Author contact, n = 1
• Clinical trial tracking, n = 1
• Google search, n = 1
Articles included in review (n = 30) (n = 27 studies)
• Question 1, n = 30 (n = 27 studies)
• Question 2, n = 0

**FIGURE 17** Preferred Reporting Items for Systematic Reviews and Meta-Analyses diagram: summary of publications included and excluded at each stage of the review.

| | Test | Risk of bias (QUADAS-2) | | | | | | Applicability concerns (QUADAS-2) | | | | | | | Risk of bias (QUADAS-C) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P | I | R | | FT | | P | | | | I | R | | P | I | R | | FT | |
| | | | | Nodule | Cancer | Nodule | Cancer | INCID | SYMP | SCREEN | SURV | | Nodule | Cancer | | | Nodule | Cancer | Nodule | Cancer |
| Abadia 2021 | A | High | Low | High | | Low | | High | High | High | | High | High | | High | High | High | | Low | |
| | D | High | High | High | | Low | | High | High | High | | High | High | | | | | | | |
| | E | High | Low | High | | Low | | High | High | High | | Unclear | Unclear | | | | | | | |
| Hall 2019 | C | Unclear | High | High | Unclear | High | Unclear | | | High | | High | Low | Unclear | Unclear | High | High | | High | |
| | E | Low | Low | High | | Low | | | | High | | High | Low | | | | | | | |
| Hempel 2022 | C | High | High | High | | Low | | High | | | High | High | High | | High | High | High | | Low | |
| | D | High | High | High | | Low | | High | | | High | High | High | | | | | | | |
| Hsu 2021 | A | High | Unclear | High | | Low | | High | High | High | | High | Low | | High | High | High | | Low | |
| | B | High | High | High | | Low | | High | High | High | | High | High | | | | | | | |
| | C | High | High | High | | Low | | High | High | High | | High | High | | | | | | | |
| | D | High | High | High | | Low | | High | High | High | | High | High | | | | | | | |
| Hwang 2021 | A | Unclear | Low | High | High | High | Unclear | | | High | | High | High | High | High | Low | | High | | Unclear |
| | C | Unclear | Low | | High | | Unclear | | | High | | High | | High | | | | | | |
| | E | Low | Low | | High | | Unclear | | | High | | High | | High | | | | | | |
| Kozuka 2020 | A | High | Unclear | Low | | Low | | | High | | | High | High | | High | High | Low | | Low | |
| | C | High | High | Low | | Low | | | High | | | High | High | | | | | | | |
| | D | High | High | Low | | Low | | | High | | | High | High | | | | | | | |
| Lancaster 2022 | A | HIGH | Low | High | | Low | | | | High | | High | High | | High | High | High | | Low | |
| | C | High | High | High | | Low | | | | High | | High | High | | | | | | | |
| | D | High | High | High | | Low | | | | High | | High | High | | | | | | | |
| Liu 2019 | A | High | High | Low | | Low | | High | High | | | High | High | | High | High | Low | | High | |
| | C | High | High | Low | | High | | High | High | | | High | High | | | | | | | |
| | D | High | High | Low | | High | | High | High | | | High | High | | | | | | | |
| Lo 2018 | A | High | Unclear | Low | Low | Low | High | | | High | | High | High | Low | High | High | Low | Low | Low | High |
| | C | High | High | Low | Low | Low | High | | | High | | High | High | Low | | | | | | |
| | D | High | High | Low | Low | Low | High | | | High | | High | High | Low | | | | | | |
| Milanese 2018 | C | Unclear | High | High | | High | | Unclear | Unclear | High | High | High | Low | | Unclear | High | High | | High | |
| | D | Low | High | High | | High | | Unclear | Unclear | High | High | High | Low | | | | | | | |
| Murchison 2022 | A | High | High | High | | High | | High | High | High | High | High | Low | | High | High | High | | High | |
| | C | High | High | High | | High | | High | High | High | High | High | Low | | | | | | | |
| | D | High | High | High | | High | | High | High | High | High | High | Low | | | | | | | |
| Roehrich 2022 | C | High | High | High | | Low | | High | High | | | High | Unclear | | High | High | High | | Low | |
| | D | High | High | High | | Low | | High | High | | | High | Unclear | | | | | | | |
| Rueckel 2021 | A | Low | Unclear | High | | Low | | Low | | | | High | High | | Low | Unclear | High | | Low | |
| | D | Low | Low | High | | Low | | Low | | | | High | High | | | | | | | |
| Singh 2021 | A | High | Unclear | Low | | High | | | | High | | High | High | | High | High | Low | | High | |
| | C.1 | High | High | Low | | High | | | | High | | High | High | | | | | | | |
| | C.2 | High | High | Low | | High | | | | High | | High | High | | | | | | | |
| | D | High | High | Low | | High | | | | High | | High | High | | | | | | | |
| Takaishi 2021 | C | High | High | High | Unclear | Low | High | High | High | | | High | High | Unclear | High | High | High | Unclear | Low | High |
| | D | High | High | High | Unclear | Low | High | High | High | | | High | High | Unclear | | | | | | |
| Zhang 2021 | C | Low | High | High | | Low | | | | High | | High | High | | Low | High | High | | Low | |
| | D | Low | Low | High | | Low | | | | High | | High | High | | | | | | | |
| Non-comparative accuracy studies | | | | | | | | | | | | | | | | | | | | |
| Blazis 2021 | A | Unclear | High | High | | High | | High | High | High | | High | High | | | | | | | |
| Chamberlin 2021 | A | Low | Low | High | | High | | | | High | | High | High | | | | | | | |
| Hwang 2021b | C | Unclear | Low | | High | | High | | | High | | High | | High | | | | | | |
| Martins Jarnalo 2021 | A | High | Unclear | High | | Low | | High | High | | High | High | High | | | | | | | |
| Wakkie 2020 | A | High | High | Low | | Unclear | | High | High | High | | High | High | | | | | | | |
| Wan 2022 | A | High | Unclear | Low | Low | Low | Low | High | High | High | High | High | High | Low | | | | | | |

**FIGURE 18** Quality assessment results based on QUADAS-2 and QUADAS-C tools (22 studies). A, Stand-alone AI; B, assisted 2nd-read AI; C, concurrent AI; C.1, concurrent AI for vessel-suppression; C.2, concurrent AI for vessel-suppression and nodule detection; D, unaided reader (reader study); E, original radiologist (clinical practice); FT, flow and timing; I, index test; INCID, incidental population; P, population; R, reference standard; SCREEN, screening population; SURV, surveillance population; SYMP, symptomatic population.



FT, Flow and timing; I, Index test; P, Population; RS, Reference standard; a. Nodule detection

**FIGURE 19** Findings of risk-of-bias assessment for all 22 studies as well as separately for comparative (QUADAS-C) and non-comparative (QUADAS-2) studies. (a) Stand-alone AI (*n* = 15); (b) 2nd-read AI (*n* = 1); (c) concurrent AI (*n* = 15); (d) unaided reading (*n* = 16). I, index test; NA, not applicable; P, population; RS, reference standard.

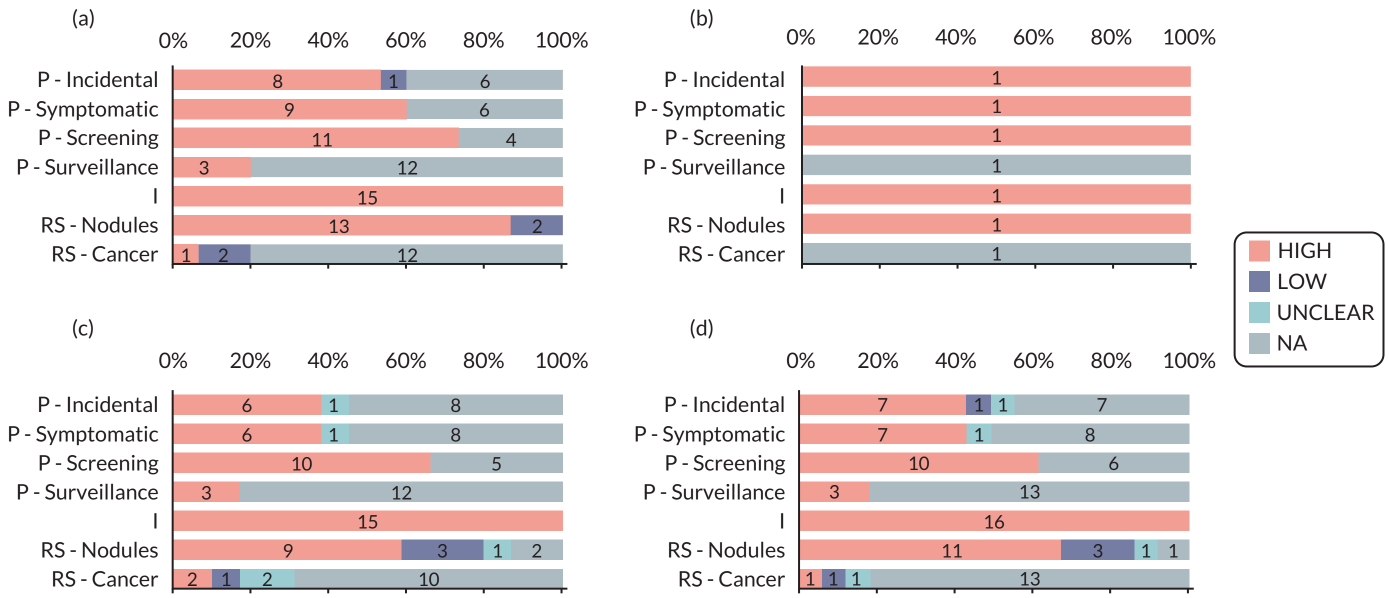


**FIGURE 20** Findings of applicability concern assessment (QUADAS-2) by index test.

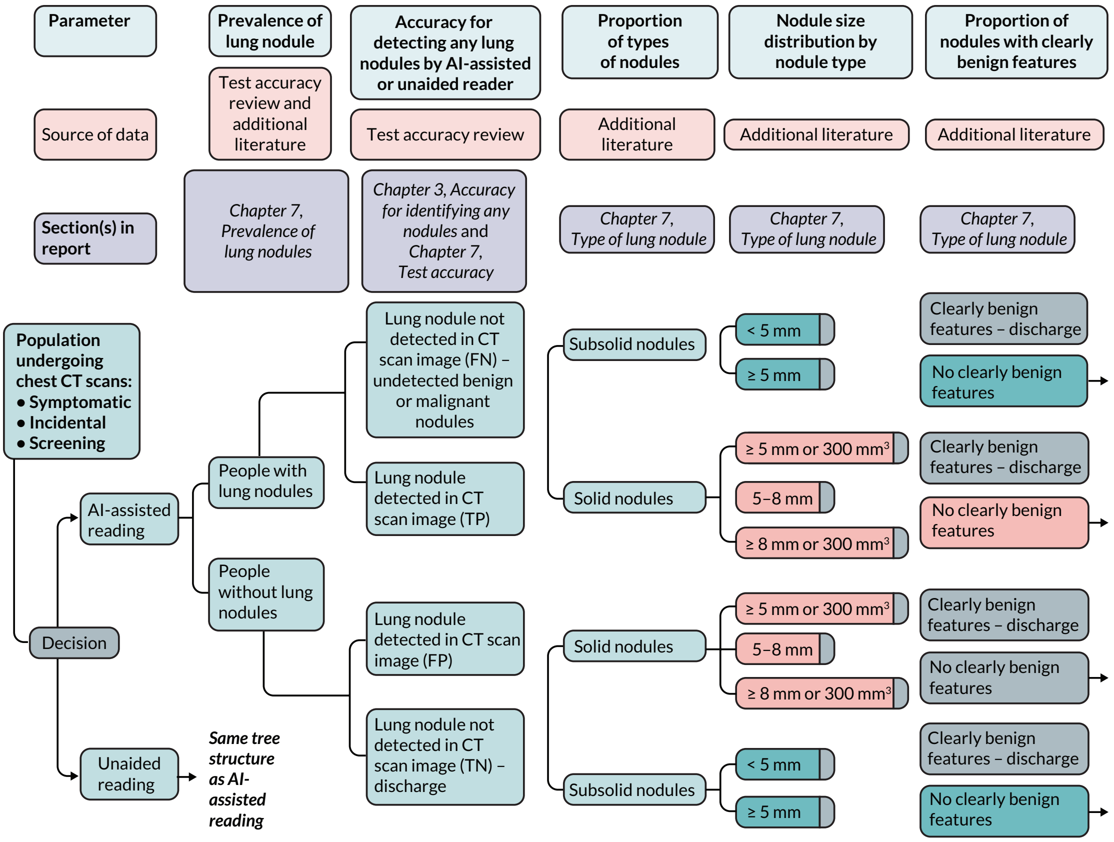

Parameter
Prevalence of lung nodule
Accuracy for detecting any lung nodules by AI-assisted or unaided reader
Proportion of types of nodules
Nodule size distribution by nodule type
Proportion of nodules with clearly benign features
Source of data
Test accuracy review and additional literature
Test accuracy review
Additional literature
Additional literature
Additional literature
Section(s) in report
Chapter 7, Prevalence of lung nodules
Chapter 3, Accuracy for identifying any nodules and Chapter 7, Test accuracy
Chapter 7, Type of lung nodule
Chapter 7, Type of lung nodule
Chapter 7, Type of lung nodule
Population undergoing chest CT scans:
• Symptomatic
• Incidental
• Screening
Decision
AI-assisted reading
Unaided reading
Same tree structure as AI-assisted reading
People with lung nodules
People without lung nodules
Lung nodule not detected in CT scan image (FN) – undetected benign or malignant nodules
Lung nodule detected in CT scan image (TP)
Lung nodule detected in CT scan image (FP)
Lung nodule not detected in CT scan image (TN) – discharge
Subsolid nodules
Solid nodules
Solid nodules
Subsolid nodules
< 5 mm
≥ 5 mm
≥ 5 mm or 300 mm³
5–8 mm
≥ 8 mm or 300 mm³
≥ 5 mm or 300 mm³
5–8 mm
≥ 8 mm or 300 mm³
< 5 mm
≥ 5 mm
Clearly benign features – discharge
No clearly benign features
Clearly benign features – discharge
No clearly benign features
Clearly benign features – discharge
No clearly benign features
Clearly benign features – discharge
No clearly benign features

**FIGURE 21** Abbreviated representation of the decision tree, required model parameters and data source (further parts are shown in *Figures 22* and *23*).

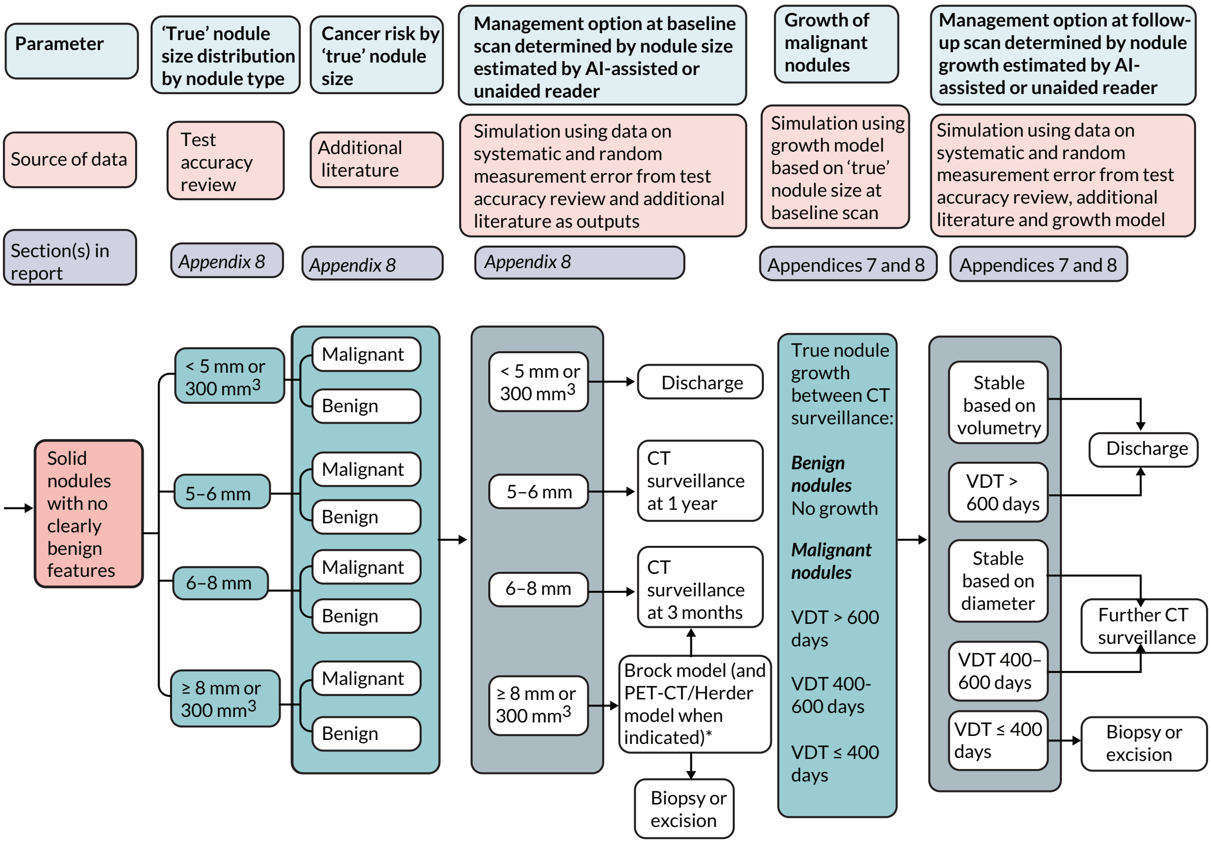

Parameter
'True' nodule size distribution by nodule type
Cancer risk by 'true' nodule size
Management option at baseline scan determined by nodule size estimated by AI-assisted or unaided reader
Growth of malignant nodules
Management option at follow-up scan determined by nodule growth estimated by AI-assisted or unaided reader
Source of data
Test accuracy review
Additional literature
Simulation using data on systematic and random measurement error from test accuracy review and additional literature as outputs
Simulation using growth model based on 'true' nodule size at baseline scan
Simulation using data on systematic and random measurement error from test accuracy review, additional literature and growth model
Section(s) in report
Appendix 8
Appendix 8
Appendix 8
Appendices 7 and 8
Appendices 7 and 8
Solid nodules with no clearly benign features
< 5 mm or 300 mm³
5–6 mm
6–8 mm
≥ 8 mm or 300 mm³
Malignant
Benign
Malignant
Benign
Malignant
Benign
Malignant
Benign
< 5 mm or 300 mm³
5–6 mm
6–8 mm
≥ 8 mm or 300 mm³
Discharge
CT surveillance at 1 year
CT surveillance at 3 months
Brock model (and PET-CT/Herder model when indicated)*
Biopsy or excision
True nodule growth between CT surveillance:
Benign nodules
No growth
Malignant nodules
VDT > 600 days
VDT 400-600 days
VDT ≤ 400 days
Stable based on volumetry
VDT > 600 days
Stable based on diameter
VDT 400–600 days
VDT ≤ 400 days
Discharge
Further CT surveillance
Biopsy or excision

**FIGURE 22** Abbreviated representation of the solid nodule part of the decision tree, required model parameters and data source (continued from *Figure 21*).

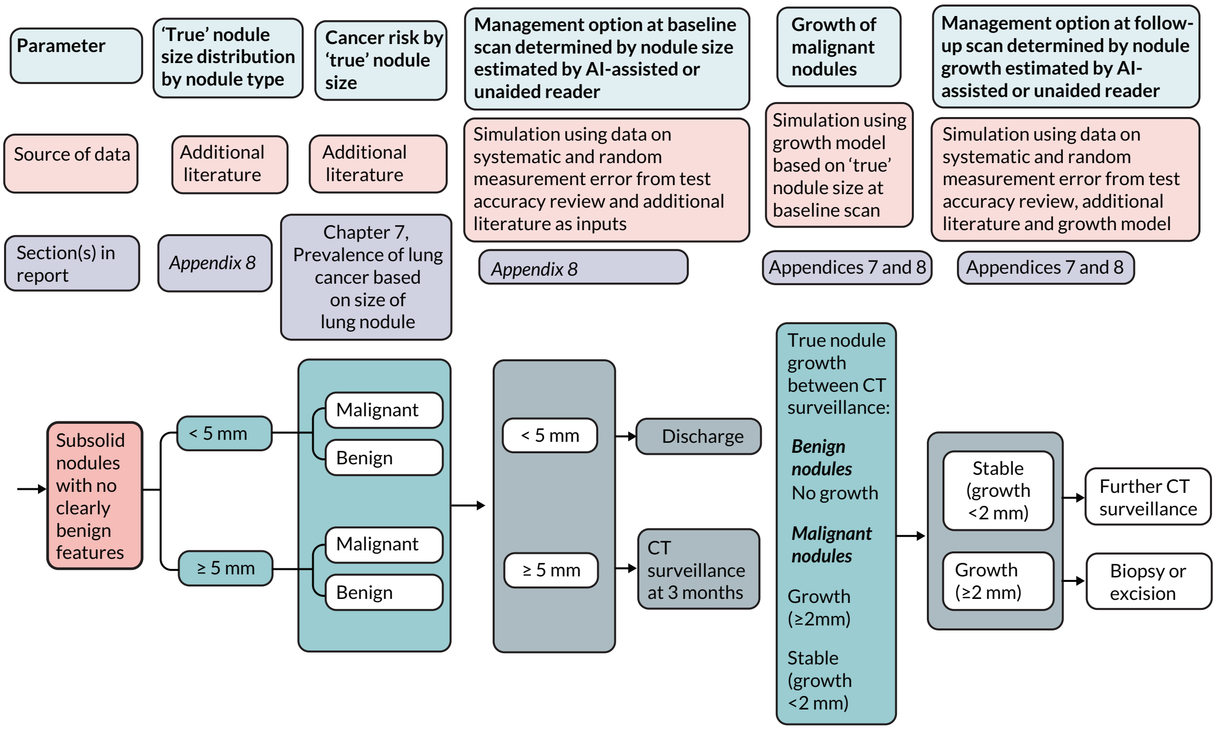

Parameter
'True' nodule size distribution by nodule type
Cancer risk by 'true' nodule size
Management option at baseline scan determined by nodule size estimated by AI-assisted or unaided reader
Growth of malignant nodules
Management option at follow-up scan determined by nodule growth estimated by AI-assisted or unaided reader
Source of data
Additional literature
Additional literature
Simulation using data on systematic and random measurement error from test accuracy review and additional literature as inputs
Simulation using growth model based on 'true' nodule size at baseline scan
Simulation using data on systematic and random measurement error from test accuracy review, additional literature and growth model
Section(s) in report
Appendix 8
Chapter 7, Prevalence of lung cancer based on size of lung nodule
Appendix 8
Appendices 7 and 8
Appendices 7 and 8
Subsolid nodules with no clearly benign features
< 5 mm
≥ 5 mm
Malignant
Benign
Malignant
Benign
< 5 mm
≥ 5 mm
Discharge
CT surveillance at 3 months
True nodule growth between CT surveillance:
Benign nodules
No growth
Malignant nodules
Growth (≥2mm)
Stable (growth <2 mm)
Stable (growth <2 mm)
Growth (≥2 mm)
Further CT surveillance
Biopsy or excision

FIGURE 23 Abbreviated representation of the subsolid nodule part of the decision tree, required model parameters and data source (continued from *Figure 21*).

## Supporting tables

TABLE 38 Outcomes: nodule detection and analysis: accuracy, concordance and variability

| Outcome | Section in report | Comparison | Number of studies | Target population, references |
|---|---|---|---|---|
| ***Use case 1: nodule detection and analysis in people with no known lung nodules*** | | | | |
| Nodule detection: accuracy: any nodule | *Nodule detection*<br>*Appendix 5* | [C] vs. [D] | *n* = 4 | Screening[53,61]<br>Symptomatic[59]<br>Mixed[53,57] |
| | | [B] vs. [D] | *n* = 1 | Screening[53]<br>Mixed[53] |
| | *Appendix 6* | [A] vs. [D] | *n* = 4 | Symptomatic[59]<br>Incidental[49]<br>Mixed[47,60] |
| | | None: [A] | *n* = 6 | Screening[51]<br>Mixed[30,47,58,65,66] |
| Nodule detection: accuracy: actionable nodules | *Nodule detection*<br>*Appendix 5* | [C] vs. [D] | *n* = 5 | Screening[27,54,56]<br>Symptomatic[59]<br>Mixed[33] |
| | *Appendix 6* | [A] vs. [D] | *n* = 2 | Symptomatic[59]<br>Mixed[60] |
| | | None: [A] | *n* = 2 | Screening[48]<br>Mixed[30] |

**TABLE 38** Outcomes: nodule detection and analysis: accuracy, concordance and variability (*continued*)

| Outcome | Section in report | Comparison | Number of studies | Target population, references |
|---|---|---|---|---|
| Nodule detection: accuracy: malignant nodules | *Nodule detection*<br>*Appendix 5* | [C] vs. [D] | *n* = 3 | Screening[54,67]<br>Mixed[57] |
| | *Appendix 6* | None | *n* = 3 | Screening[27,51]<br>Mixed[58] |
| Nodule detection: effect modifiers | *Nodule detection* | Radiation dose | *n* = 2 | Mixed[53,60] |
| | *Nodule detection* | Nodule type | *n* = 7 | Screening[51,54,56,61]<br>Symptomatic[59]<br>Mixed[60,66] |
| | *Nodule detection* | Radiologist experience | *n* = 1 | Screening[53]<br>Mixed[53] |
| Nodule detection: concordance | *Nodule detection* | [A] [C] vs. [D] | *n* = 1 | Mixed[47] |
| | | Inter-observer | *n* = 1 | Screening[56] |
| Nodule type: accuracy | *Nodule type determination* | None: [A] | *n* = 2 | Mixed[33,66] |
| Nodule type: concordance | *Nodule type determination* | Inter-observer | *n* = 2 | Screening[64,67] |
| Diameter measurement: accuracy | *Nodule diameter measurement* | None: [C] | *n* = 1 | Unclear[55] |
| | | None: [A] | *n* = 2 | Screening[56]<br>Mixed[66] |
| Diameter measurement: concordance | *Nodule diameter measurement* | [A] [C] vs. [D] | *n* = 4 | Surveillance with applicability concerns[63]<br>Mixed[33,47,58] |
| | | Inter-observer | *n* = 5 | Screening[64,67]<br>Surveillance with applicability concerns[62,63]<br>Mixed[33] |
| | | Intra-observer | *n* = 2 | Surveillance with applicability concerns[62,63] |
| Volume measurement: accuracy | *Nodule volume measurement* | None: [C] | *n* = 1 | Unclear[55] |
| Volume measurement: concordance | *Nodule volume measurement* | [A] vs. [D] | *n* = 1 | Mixed[33] |
| | | Inter-observer | *n* = 3 | Surveillance with applicability concern[62]<br>Mixed[33]<br>Unclear[55] |
| | | Intra-observer | *n* = 1 | Surveillance with applicability concerns[62] |
| Risk categorisation: accuracy | *Classification into risk categories based on nodule type and size* | [A] [C] vs. [D] | *n* = 3 | Screening[32]<br>Mixed[34]<br>Unclear[55] |

continued

TABLE 38 Outcomes: nodule detection and analysis: accuracy, concordance and variability (*continued*)

| Outcome | Section in report | Comparison | Number of studies | Target population, references |
|---|---|---|---|---|
| Risk categorisation: concordance | *Classification into risk categories based on nodule type and size* | [A] [C] vs. [D] | *n* = 2 | Screening[64,67] |
| | | Inter-observer | *n* = 5 | Screening[64,67]<br>Surveillance with applicability concerns[62,63]<br>Mixed[34] |
| | | Intra-observer | *n* = 2 | Surveillance with applicability concerns[62,63] |
| Whole read: accuracy for lung cancer | *Whole read (Detection plus risk categorisation based on nodule type and size)* | [C] vs. [D] | *n* = 1 | Screening[51] |
| | | None [C] | *n* = 1 | Screening[50] |
| ***Use case 2: nodule growth monitoring in people with previously identified lung nodules*** | | | | |
| Nodule registration: accuracy | *Nodule registration and growth assessment*<br>*Appendix 6* | None [A] | *n* = 1 | Mixed[33] |
| Nodule growth rate: concordance | *Nodule registration and growth assessment*<br>*Appendix 6* | [A] vs. [D] | *n* = 1 | Mixed[33] |
| | | Inter-observer | *n* = 1 | Mixed[33] |

[A] Stand-alone AI; [B] 2nd-read AI; [C] concurrent AI; [D] unaided reading.

TABLE 39 Outcomes: practical implications

| Outcome | Section in report | Number of studies | Target population, references |
|---|---|---|---|
| Technical failure rate | *Technical failure rate (12 studies)* | *n* = 12 | Screening[27,50–52,56,64]<br>Surveillance with applicability concerns[62,63]<br>Mixed[31,33,34,66] |
| Radiologist reading time | *Radiologist reading time (10 studies)* | *n* = 10 | Screening[27,54,64]<br>Symptomatic[59]<br>Mixed[31,34,47,53,57,60] |
| Acceptability and experience of using the software | *Acceptability and experience of using the software (three studies)* | *n* = 3 | Screening[27]<br>Mixed[47,66] |

[A] Stand-alone AI; [B] 2nd-read AI; [C] concurrent AI; [D] unaided reading; NA, not applicable.

**TABLE 40** Outcomes: impact on patient management

| Outcome | Section in report | Comparison | Number of studies | Target population, references |
|---|---|---|---|---|
| ***Characteristics of detected nodules*** | | | | |
| All detected nodules (true positive and false positive) | *Characteristics of detected nodules*<br>*Appendix 5* | [C] vs. [D] | *n* = 2 | Screening[51]<br>Mixed[34] |
| | | [A] vs. [D] | *n* = 1 | Mixed[47] |
| | *Appendix 6* | None | *n* = 3 | Screening[50,52]<br>Mixed[66] |
| True-positive nodules | *Characteristics of detected nodules*<br>*Appendix 5* | [C] vs. [D] | *n* = 2 | Screening[61]<br>Symptomatic[59] |
| | | [A] vs. [D] | *n* = 1 | Mixed[60] |
| | *Appendix 6* | None | *n* = 4 | Screening[32,51,56]<br>Mixed[66] |
| Additional true-positive nodules detected by software | *Characteristics of detected nodules* | [A] vs. [D] | *n* = 1 | Incidental[49] |
| False-positive nodules | *Characteristics of detected nodules*<br>*Appendix 5* | None ([A] only) | *n* = 4 | Screening[48]<br>Incidental[49]<br>Mixed[47,66] |
| FN nodules | *Characteristics of detected nodules*<br>*Appendix 5* | [C] vs. [D] | *n* = 2 | Screening[61]<br>Symptomatic[59] |
| | *Appendix 6* | None | *n* = 5 | Screening[27,51,56]<br>Mixed[58,66] |
| Proportion of detected nodules that are malignant | *Proportion of detected nodules that are malignant (three studies)*<br>*Appendix 5* | [C] vs. [D] | *n* = 2 | Screening[27,51] |
| | | None | *n* = 1 | Screening[50] |
| Impact of test result on clinical decision-making | *Impact of test result on clinical decision-making (six studies)*<br>*Appendix 5* | [C] vs. [D] | *n* = 6 | Screening[27,56,64,67]<br>Surveillance with applicability concerns[63]<br>Unclear[55] |
| Number of people having CT surveillance | *Number of people having computed tomography surveillance (five studies)*<br>*Appendix 5* | [C] vs. [D] | *n* = 2 | Screening[51,64] |
| | *Appendix 6* | None | *n* = 3 | Screening[27,52]<br>Symptomatic[59] |
| Number of people having biopsy or excision | *Number of people having a biopsy or excision (five studies)*<br>*Appendix 5* | [C] vs. [D] | *n* = 2 | Screening[51,64] |
| | *Appendix 6* | None | *n* = 3 | Screening[27,52]<br>Symptomatic[59] |
| Time to diagnosis | *Time to diagnosis (one study)* | [C] vs. [D] | *n* = 1 | Screening[64] |

[A] Stand-alone AI; [B] 2nd-read AI; [C] concurrent AI; [D] unaided reading.

**TABLE 41** Accuracy for the detection of any nodules in standard-dose CT and LDCT scans according to Hsu *et al.*[53]

| | Dose of CT | Total scans | Total nodules | Per-nodule sensitivity, % (95% CI) | Per-patient specificity, % (95% CI) |
|---|---|---|---|---|---|
| 2nd-read AI | Standard dose | 93 | 222 | 83 (81 to 85) | 87 (84 to 87) |
| | Low dose | 57 | 118 | 80 (77 to 83) | 82 (79 to 84) |
| Concurrent AI | Standard dose | 93 | 222 | 81 (79 to 83) | 83 (83 to 87) |
| | Low dose | 57 | 118 | 79 (76 to 81) | 82 (78 to 84) |
| Unaided reading | Standard dose | 93 | 222 | 63 (61 to 66) | 80 (79 to 83) |
| | Low dose | 57 | 118 | 63 (59 to 66) | 72 (74 to 80) |

CI, confidence interval; CT, computed tomography.

**TABLE 42** Effect of nodule type on nodule detection accuracy in screening populations: concurrent AI vs. unaided reader (two studies)

| | | | | Sensitivity, % (95% CI) | | Specificity, % (95% CI) | |
|---|---|---|---|---|---|---|---|
| Authors/year, software | Nodule type | Number of scans | Number of nodules | Concurrent CAD | Unaided reader | Concurrent CAD | Unaided reader |
| Zhang *et al*. 2021,[61] InferRead CT Lung (Infervision) | Solid nodules | 250 | NR | 98.8 (96.5 to 99.8) | 52.4 (46.0 to 58.7) | 99.2 (98.1 to 99.7) | 100.0 (99.4 to 100) |
| | Part-solid nodules | 13 | NR | 100.0 (75.3 to 100) | 23.1 (5.0 to 53.8) | 100.0 (99.6 to 100) | 100.0 (99.6 to 100) |
| | Ground-glass nodules | 111 | NR | 99.1 (95.1 to 99.9) | 25.2 (17.5 to 34.4) | 98.8 (97.7 to 99.5) | 100.0 (99.5 to 100) |
| Singh *et al*. 2021,[56] ClearRead Vessel Suppression (Riverain Technologies) | Subsolid nodules | NR | 310 | 73 | 68 | 74 | 77.5 |
| | Part-solid nodules | NR | 154 | 76 | 70 | 85 | 76 |
| | Ground-glass nodules | NR | 156 | 67 | 67 | 78.5 | 84 |

CAD, computer-aided detection; CI, confidence interval; NR, not reported.

**TABLE 43** Effect of nodule type on nodule detection accuracy in a symptomatic population: concurrent AI/stand-alone AI vs. unaided reader (one study)

| | | | | Per-nodule sensitivity, % (95% CI) | | |
|---|---|---|---|---|---|---|
| Authors/year, software | Nodule type | Number of scans | Number of nodules | Stand-alone AI | Concurrent AI (pooled two readers) | Unaided reader (pooled two readers) |
| Kozuka *et al.* 2020,[59] InferRead CT Lung (Infervision) | Solid nodules | NR | 518 | 68.1 (63.9 to 72.1) | 32.6 (29.8 to 35.6)[a] | 18.6 (16.3 to 21.1) |
| | Part-solid nodules | NR | 65 | 70.8 (58.2 to 81.4) | 58.5 (49.5 to 67.0)[a] | 31.5 (23.7 to 40.3) |
| | Ground-glass nodules | NR | 86 | 72.1 (61.4 to 81.2) | 40.1 (32.7 to 47.9)[a] | 18.0 (12.6 to 24.6) |

AI, artificial intelligence; CI, confidence interval; NR, not reported.
a $p < 0.01$ vs. unaided reader.

**TABLE 44** Effect of nodule type on nodule detection accuracy in screening population: stand-alone AI (two studies)

| Authors/year, software | Nodule type | Number of scans | Number of nodules | Sensitivity, % (95% CI) |
|---|---|---|---|---|
| Hwang *et al.* 2021,[51] AVIEW Lungscreen (Coreline Soft) | Solid nodules | NR | 4032 | 51 (50 to 53) |
| | Part-solid nodules | NR | 70 | 49 (36 to 61) |
| | Ground-glass nodules | NR | 178 | 21 (16 to 29) |
| Lo *et al.* 2018,[54] ClearRead CT (Riverain Technologies) | Solid nodules | NR | 119 | 84 |
| | Part-solid nodules | NR | 35 | 85 |
| | Ground-glass nodules | NR | 24 | 67 |

CI, confidence interval; NR, not reported.

**TABLE 45** Effect of nodule type on nodule detection accuracy in mixed populations: stand-alone AI alone (one study) or vs. unaided readers (one study)

| | | | | Per-nodule sensitivity, % | | |
|---|---|---|---|---|---|---|
| Authors/year, software | Nodule size and dose | Nodule type | Number of nodules | Stand-alone AI | Unaided reader 1 | Unaided reader 2 |
| Liu *et al.* 2018,[60] InferRead CT Lung (Infervision) | > 6 mm, conventional dose | Solid nodules | 215 | 87.9 | 77.2 | 69.3 |
| | ≤ 6 mm, conventional dose | Solid nodules | 2680 | 64.4 | 36.1 | 50.3 |
| | > 6 mm, low dose | Solid nodules | 44 | 88.6 | 93.2 | 81.8 |
| | ≤ 6 mm, low dose | Solid nodules | 719 | 71.9 | 41.7 | 49.8 |

**TABLE 45** Effect of nodule type on nodule detection accuracy in mixed populations: stand-alone AI alone (one study) or vs. unaided readers (one study) (*continued*)

| | | | | Per-nodule sensitivity, % | | |
|---|---|---|---|---|---|---|
| **Authors/year, software** | **Nodule size and dose** | **Nodule type** | **Number of nodules** | **Stand-alone AI** | **Unaided reader 1** | **Unaided reader 2** |
| | > 5 mm, conventional dose | Subsolid nodules | 371 | 81.1 | 58.2 | 85.2 |
| | ≤ 5 mm, conventional dose | Subsolid nodules | 993 | 68.1 | 26.2 | 56.9 |
| | > 5 mm, low dose | Subsolid nodules | 61 | 85.2 | 67.2 | 82.0 |
| | ≤ 5 mm, low dose | Subsolid nodules | 333 | 61.3 | 22.5 | 56.2 |
| Martins Jarnalo *et al*. 2021,[66] Veye Chest (Aidence) | 4–30 mm | Solid nodules | 73 | 89.0 | NA | NA |
| | 4–30 mm | Subsolid | 16 | 81.3 | NA | NA |
| | 4–30 mm | Mixed (solid/ subsolid) | 2 | 100.0 | NA | NA |

**TABLE 46** Accuracy of readers with and without concurrent use of Veye Chest to identify patients with BTS grade A (no clinical follow-up recommended)[34]

| | Sensitivity (95% CI) | | Specificity (95% CI) | |
|---|---|---|---|---|
| | **Unaided** | **Aided** | **Unaided** | **Aided** |
| Reader 1 | 0.83 (0.61 to 0.95) | 0.85 (0.66 to 0.96) | 0.85 (0.66 to 0.96) | 1.00 (0.85 to 1.00) |
| Reader 2 | 0.76 (0.55 to 0.91) | 0.92 (0.73 to 0.99) | 0.84 (0.64 to 0.95) | 0.96 (0.80 to 1.00) |

CI, confidence interval.

**TABLE 47** Risk categorisation using standard CT images and vessel-suppressed CT images for semiautomatic volume measurement

| | | **Reader 1** | **Reader 2** | **Reference standard** |
|---|---|---|---|---|
| Semiautomatic measurement on standard CT images | < 100 $mm^3$ | 48 (73.8%) | 48 (73.8%) | 49 (75.4%) |
| | 100–250 $mm^3$ | 11 (16.9%) | 11 (16.9%) | 10 (15.4%) |
| | > 250 $mm^3$ | 6 (9.2%) | 6 (9.2%) | 6 (9.2%) |
| Semiautomatic measurement on vessel-suppressed CT images | < 100 $mm^3$ | 50 (76.9%) | 49 (75.4%) | NA |
| | 100–250 $mm^3$ | 9 (13.8%) | 9 (13.8%) | NA |
| | > 250 $mm^3$ | 6 (9.2%) | 7 (10.8%) | NA |

Modified from Milanese *et al.*[54]

# Appendix 2 Descriptions of technologies included in this assessment

## AI-Rad Companion Chest CT (Siemens Healthineers)

AI-Rad Companion Chest CT is a CE-marked (class IIa medical device) software. It includes Lung-CAD, a tool that can detect and measure solid lung nodules in CT scans that cover the entire lung, with and without contrast. The algorithms are optimised for nodules between 3 mm and 30 mm. Lung-CAD is compatible with slice thickness of up to 2.5 mm. It is indicated for use in both screening and diagnostic protocols in people without diffuse interstitial or airway diseases, severe pneumonia, extensive granulomatous diseases, prior thoracotomy or history of radiation therapy involving the lung parenchyma who are aged ≥ 22 years. The software integrates with the PACS.

## AVIEW LCS+ (Coreline Soft)

AVIEW LCS+ is a CE-marked (class IIa medical device) software. It can detect, measure and assess the growth of solid and subsolid nodules in low-dose chest CT scans. AVIEW LCS+ is indicated for use in adults. Other indications for use include the detection of emphysema (damage to the air sacs in the lung) and coronary artery calcification. The software integrates with PACS. The software is commercially available to the NHS.

## ClearRead CT (Riverain Technologies)

ClearRead CT is a CE-marked (class IIa medical device) software. It consists of ClearRead CT Vessel Suppress, ClearRead CT Detect and ClearRead CT Compare features. Using these features, the software can detect, measure and assess the growth of solid and subsolid lung nodules in low-dose and regular-dose CT scans where both lungs are visible, with and without contrast. The software is compatible with slice thickness of up to 5 mm. ClearRead CT is indicated for use in people aged ≥ 18 years who are asymptomatic. The software is updated frequently, but none of the functionality is expected to be removed in future updates. The software integrates with, and the findings of the software are visible within, PACS. The company expects that the training of radiologists on how to use ClearRead CT will be usually done within a day. The software is commercially available to the NHS directly from the manufacturer and through partner organisations.

## Contextflow SEARCH Lung CT (contextflow)

Contextflow SEARCH Lung CT is a CE-marked (class IIa medical device) software. It can detect and measure solid and subsolid lung nodules in chest CT scans with and without contrast. It is intended for use in clinically stable, symptomatic patients. Other indications for use include identification of lung-specific image patterns related to diseases such as airway wall thickening, bronchiectasis, emphysema and pneumothorax. contextflow SEARCH Lung CT integrates with PACS. The company expects users to attend a training presentation before using the software. The software is commercially available to the NHS.

## InferRead CT Lung (Infervision)

InferRead CT Lung is a CE-marked (class IIa medical device) software. It can detect, measure and assess the growth of solid and subsolid lung nodules in low-dose or regular-dose CT scans with and without contrast. The company advises that InferRead CT Lung is intended for use in asymptomatic populations. The company also states that use is recommended in people aged ≥ 18 years. Users can dismiss nodules found by the automated analysis, but editing the findings is not possible. Users can add nodules, but the software will not measure the volume of user-added nodules. A new version of InferRead CT Lung is expected to be released within 18 months. The current version will continue to be supported and is available to the NHS. InferRead CT Lung includes rules for reporting that follow the BTS guidelines for the investigation and management of pulmonary nodules.[12] The software integrates with, and the findings of the software are visible within, PACS. The company expects radiologists to complete a 1-hour training session before using the technology. The software is commercially available to the NHS.

## JLD-01K (JLK, Inc.)

JLD-01K is a CE-marked (class I medical device) software. It can detect and measure solid and subsolid lung nodules in chest CT scans without contrast. The software was trained in CT scans where nodules were at least 3 mm in diameter. JLK-01K integrates with PACS.

## Lung AI (Arterys)

Lung AI is a CE-marked (class IIa medical device) software. It can detect, measure and assess the growth of solid and subsolid lung nodules in chest CT scans. The nodule detection and segmentation algorithms are optimised for low-dose chest CT scans, but the software will analyse any chest CT scan including regular-dose CT scans with contrast without generating an error. Users can add, edit or dismiss detected nodules with automatic updates to quantitative nodule information. Lung AI integrates with PACS.

## Lung Nodule AI (Fujifilm)

Lung Nodule AI is a software that can detect, measure and assess the growth of lung nodules in chest CT scans. The software is currently approved for use in Japan. The company plans to introduce the technology in Europe once required regulatory clearances are obtained.

## qCT-Lung (Qure.ai)

qCT-Lung is a CE-marked (class I medical device) software. It can detect lung nodules at least 3 mm in diameter in chest CT scans without contrast. The software can also measure the volume and assess the growth of lung nodules, but these features are currently available for research purposes only. Other indications for use include detection of emphysema. qCT-Lung is intended for use in people aged ≥ 18 years. The software is compatible with slice thickness of up to 6 mm. qCT-Lung integrates with PACS.

## SenseCare-Lung Pro (SenseTime)

SenseCare-Lung Pro is a CE-marked (class IIb medical device) software. It can detect, measure and assess the growth of solid and subsolid lung nodules in chest CT scans without contrast. Other indications for use include detection of pneumonia (including COVID-19) lesions. The software is compatible with slice thickness of up to 5 mm, but the preferred slice thickness is up to 1.5 mm. SenseCare-Lung Pro integrates with PACS.

## Veolity (MeVis)

Veolity is a CE-marked (class IIa medical device) software. It can detect, measure and assess the growth of lung nodules in low-dose and regular-dose CT scans that include the complete chest, with and without contrast. The software is compatible with slice thickness of up to 3 mm. Veolity is indicated for the screening, diagnosis and monitoring of lung cancer. Users can interact with the software by adding and dismissing nodules in the analysis and editing the findings of the software. With input from the user, the software also calculates the malignancy risk of the nodules using the Brock model. Veolity's current detection algorithm only detects solid nodules. A new version of the software (Veolity 2.0) is planned for the beginning of 2022. This version will detect solid and subsolid nodules. Usually, two updates or functional upgrades per year are planned. Existing versions will continue to be supported. Veolity includes rules for reporting following the BTS guidelines for the investigation and management of pulmonary nodules[12] and integrates with the PACS. The company states that usually 4–6 hours of training are needed for radiologists to learn how to use Veolity. The software is commercially available to the NHS, distributed in the UK by SynApps Solutions.

## Veye Lung Nodules (Aidence)

Veye Lung Nodules is a CE-marked (class IIb medical device) software. It can detect, measure and assess the growth of solid and subsolid lung nodules in low-dose or standard-dose CT scans where both lungs are visible, with and without contrast. The software is compatible with slice thickness of up to 3 mm. Veye Lung Nodules is intended for use in people aged ≥ 18 years. Users can dismiss nodules found by the automated analysis, but editing the findings is not possible. Users can add nodules, but the software will not measure the volume of user-added nodules. The software is updated frequently. Veye Lung Nodules includes rules for reporting following the BTS guidelines for the investigation and management of pulmonary nodules.[12] The software integrates with, and findings of the software are visible within, PACS. The company expects radiologists to attend a 1-hour training session before using the technology. The software is commercially available to the NHS.

## VUNO Med-LungCT AI (VUNO)

VUNO Med-LungCT AI is a CE-marked (class IIa medical device) software. It can detect, measure and assess the growth of solid and subsolid lung nodules in low-dose chest CT scans. It is intended for use in lung cancer screening populations. The software integrates with PACS.

# Appendix 3 Literature search strategies

## Search strategies for systematic review of test accuracy and clinical effectiveness

Search dates and number of records retrieved per source are reported below.

| Bibliographic databases and trials registers | | |
|---|---|---|
| Database/register | Date searched | Number of records |
| MEDLINE All | 17 January 2022 | 2740 |
| EMBASE | 17 January 2022 | 3495 |
| Cochrane Library (CENTRAL and Cochrane Database of Systematic reviews) | 17 January 2022 | 131 (all from CENTRAL; 0 results from CDSR) |
| Science Citation Index and Conference Proceedings – Science (Web of Science) | 19 January 2022 | 3210 |
| HTA database (CRD) | 19 January 2022 | 1 |
| INAHTA database | 19 January 2022 | 3 |
| medRxiv | 19 January 2022 | 7 |
| ClinicalTrials.gov | 19 January 2022 | 17 |
| WHO ICTRP | 19 January 2022 | 22 |
| Total number of records retrieved: 9626<br>Duplicates removed (EndNote): 3296<br>Final number for screening: 6330 | | |
| Other sources | | |
| Source | Date searched | Documents retrieved |
| NICE website | 24 January 2022 | 3 |
| Canadian Agency for Drugs and Technologies in Health (CADTH) website | 24 January 2022 | 7 |
| ISPOR conference presentations | 25 January 2022 | 0 |
| HTAi annual meetings | 25 January 2022 | 1 |
| SPIE proceedings | 27 January 2022 | 14 |
| IEEE Engineering in Medicine & Biology Society annual conference | 27 January 2022 | 1 |
| European Congress of Radiology | 31 January 2022 | 47 |
| Radiological Society of North America annual meetings | 1 February 2022 | 55 |
| FDA devices databases | 14 February 2022 | 5 |
| Device/ manufacturer websites | 15–16 February 2022 | 15 documents, plus 1 link to video presentation |
| Forwards citation tracking: Science Citation Index (Web of Science) and Google Scholar | 26 May 2022 and 30 May 2022 | 44 |
| Total: 192 | | |

Search strategies used:

**MEDLINE All**

**Date searched: 17 January 2022**

Ovid MEDLINE(R) ALL 1946–14 January 2022

1 exp artificial intelligence/ or exp machine learning/ or exp deep learning/ or exp supervised machine learning/ or exp support vector machine/ or exp unsupervised machine learning/134,273
2 ai.kf,tw. 34,062
3 ((artificial or machine or deep) adj5 (intelligence or learning or reasoning)).kf,tw. 89,902
4 exp Neural Networks, Computer/42,235
5 (neural network* or convolutional or CNN or CNNs).kf,tw. 73835
6 exp Diagnosis, Computer-Assisted/85,513
7 ((computer aided or computer assisted) adj1 (diagnosis or detection)).kf,tw. 6018
8 (support vector machine* or random forest* or black box learning).kf,tw. 31,141
9 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8322,906
10 exp Lung Neoplasms/di, dg or Solitary Pulmonary Nodule/di, dg56,493
11 ((lung or lungs or pulmon* or bronchial) adj3 (nodul* or cancer* or neoplas* or tumor* or tumour* or carcino* or malignan* or adenocarcinom* or blastoma*)).kf,tw. 274,199
12 ((pulmonary or lung) adj2 lesion*).kf,tw. 14,782
13 10 or 11 or 12302,352
14 Tomography, X-Ray Computed/ or exp Tomography, Spiral Computed/418,962
15 (comput* adj2 tomograph*).kf,tw. 348,023
16 (CT or LDCT).kf,tw.388,825
17 (CAT adj2 (scan* or x-ray* or xray*)).kf,tw. 1342
18 Mass Screening/111,594
19 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw. 4813
20 "Early Detection of Cancer"/ 31,774
21 14 or 15 or 16 or 17 or 18 or 19 or 20893125
22 9 and 13 and 212767
23 (aview* lcs* or clearread* ct* or inferread* ct lung* or lung nodule ai* or veolity* or veye).kf,tw.7
24 ((ai rad companion* and chest) or contextflow* or search lung ct* or "jld 01k*" or qct lung* or sensecare* lung* or visia* ct* or vuno).kf,tw.8
25 (coreline* or riverain* or infervision* or fujifilm* or mevis* or aidence*).in,kf,tw. 1381
26 (siemens* healthineers* or contextflow* or jlk inc* or arterys* or qureai* or qure ai* or sensetime* or canon medical* or vuno*).in,kf,tw. 1407
27 (25 or 26) and (10 or 11)159
28 22 or 23 or 24 or 272867
29 exp animals/ not humans/4,943,529
30 28 not 292851
31 limit 30 to english language2740

The artificial intelligence search terms (lines 1–4 and 6) are based on those used in Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, *et al.* Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* 2021;**374**:n1872.

Selected lung cancer/nodule search terms (lines 11–12) were informed by those used in Duarte A, Corbett M, Melton H, Harden M, Palmer S, Soares M, Simmonds M. *EarlyCDT Lung for Lung Cancer Risk Classification of Solid Pulmonary Nodules: A Diagnostics Assessment Report*. York EAG, 2021. URL: www.nice.org.uk/guidance/indevelopment/gid-dg10041/documents (accessed 9 November 2021)

**EMBASE**

**Date searched: 17 January 2022**

EMBASE 1974–14 January 2022

1 exp artificial intelligence/ or exp machine learning/304,838
2 ai.kf,tw.45,921
3 ((artificial or machine or deep) adj5 (intelligence or learning or reasoning)).kf,tw. 105,922
4 (neural network* or convolutional or CNN or CNNs).kf,tw. 89,201
5 computer assisted diagnosis/40,877
6 ((computer aided or computer assisted) adj1 (diagnosis or detection)).kf,tw. 8264
7 (support vector machine* or random forest* or black box learning).kf,tw. 38,837
8 1 or 2 or 3 or 4 or 5 or 6 or 7420,312
9 exp lung cancer/di or lung nodule/di46,922
10 ((lung or lungs or pulmon* or bronchial) adj3 (nodul* or cancer* or neoplas* or tumor* or tumour* or carcino* or malignan* or adenocarcinom* or blastoma*)).kf,tw. 392,765
11 ((pulmonary or lung) adj2 lesion*).kf,tw. 21,058
12 9 or 10 or 11420,629
13 computer assisted tomography/ or low-dose computed tomography/ or exp x-ray computed tomography/ or multidetector computed tomography/ or spiral computer assisted tomography/ or computed tomography scanner/931,594
14 (comput* adj2 tomograph*).kf,tw. 445,065
15 (CT or LDCT).kf,tw.664,348
16 (CAT adj2 (scan* or x-ray* or xray*)).kf,tw. 2036
17 mass screening/ or cancer screening/142,872
18 screening/184,110
19 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw. 7644
20 early cancer diagnosis/9899
21 13 or 14 or 15 or 16 or 17 or 18 or 19 or 201,643,282
22 8 and 12 and 213370
23 (aview* lcs* or clearread* ct* or inferread* ct lung* or lung nodule ai or veolity* or veye).dv,kf,tw. 11
24 (qct lung* or vuno*).dv.0
25 ((ai rad companion* and chest) or contextflow* or search lung ct* or "jld 01k*" or sensecare* lung* or visia* ct*). dv,kf,tw.4
26 (coreline* or riverain* or infervision* or fujifilm* or mevis* or aidence*).dm,in,kf,tw. 5146
27 (siemens* healthineers* or contextflow* or jlk inc* or arterys* or qureai* or qure ai* or sensetime* or canon medical* or vuno*).dm,in,kf,tw. 4797
28 (26 or 27) and (9 or 10)436
29 22 or 23 or 24 or 25 or 283692
30 (exp animal/ or exp animal experiment/) not (exp human/ or exp human experiment/ or conference abstract. pt.)4,770,834
31 29 not 303673
32 limit 31 to english language3495

**Cochrane Library (via www.cochranelibrary.com)**

**Date searched: 17 January 2022**

Cochrane Central Register of Controlled Trials, Issue 12 of 12, December 2021

Cochrane Database of Systematic Reviews, Issue 1 of 12, January 2022

IDSearchHits

#1 [mh “artificial intelligence”] OR [mh “machine learning”] OR [mh “deep learning”] OR [mh “supervised machine learning”] OR [mh “support vector machine”] OR [mh “unsupervised machine learning”]1261
#2 ai:ti,ab,kw4506
#3 ((artificial OR machine OR deep) NEAR/5 (intelligence OR learning OR reasoning)):ti,ab,kw2857
#4 [mh “Neural Networks, Computer”]148
#5 ((neural NEXT network*) OR convolutional OR CNN OR CNNs):ti,ab,kw1479
#6 [mh “Diagnosis, Computer-Assisted”]1931
#7 ((“computer aided” OR “computer assisted”) NEAR/1 (diagnosis OR detection)):ti,ab,kw1001
#8 ((“support vector” NEXT machine*) OR (random NEXT forest*) OR “black box learning”):ti,ab,kw776
#9 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #810,964
#10 [mh “Lung Neoplasms”/DI,DG] OR [mh ^”Solitary Pulmonary Nodule”/DI,DG]653
#11 ((lung OR lungs OR pulmon* OR bronchial) NEAR/3 (nodul* OR cancer* OR neoplas* OR tumor* OR tumour* OR carcino* OR malignan* OR adenocarcinom* OR blastoma*)):ti,ab,kw25,143
#12 ((pulmonary OR lung) NEAR/2 lesion*):ti,ab,kw533
#13 #10 OR #11 OR #1225,426
#14 [mh ^“Tomography, X-Ray Computed”] OR [mh “Tomography, Spiral Computed”]4555
#15 (comput* NEAR/2 tomograph*):ti,ab,kw20,680
#16 (CT OR LDCT):ti,ab,kw81,013
#17 (CAT NEAR/2 (scan* OR x-ray* OR xray*)):ti,ab,kw34
#18 [mh ^“Mass Screening”]3339
#19 ((lung OR lungs OR pulmon*) NEAR/3 (nodule* OR cancer* OR tumor* OR tumour*) NEAR/3 screen*):ti,ab,kw758
#20 [mh ^“Early Detection of Cancer”]1384
#21 #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #2096,454
#22 #9 AND #13 AND #21125
#23 ((aview* NEXT lcs*) OR (clearread* NEXT ct*) OR (inferread* NEXT “ct” NEXT lung*) OR (“lung nodule” NEXT ai*) OR veolity* OR veye)2
#24 ((“ai rad” NEXT companion*) AND chest) OR contextflow* OR (“search lung” NEXT ct*) OR (jld NEXT 01k*) OR (qct NEXT lung*) OR (sensecare* NEXT lung*) OR (visia* NEXT ct*) OR vuno*2
#25 coreline* OR riverain* OR infervision* OR fujifilm* OR mevis* OR aidence*152
#26 (siemens* NEXT healthineers*) OR contextflow* OR (jlk NEXT inc*) OR arterys* OR qureai* OR (qure NEXT ai*) OR sensetime* OR (canon NEXT medical*) OR vuno*57
#27 (#25 OR #26) AND (#10 OR #11)6
#28 #22 OR #23 OR #24 OR #27 in Cochrane Reviews, Trials131

The Ovid MEDLINE search strategy was translated for use in the Cochrane Library and Web of Science with the aid of the Polyglot Search Translator: Clark JM, Sanders S, Carter M, Honeyman D, Cleo G, Auld Y, *et al.* Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. *J Med Libr Assoc* 2020;**108**:195–207. https://doi.org/10.5195/jmla.2020.834

## Science Citation Index and Conference Proceedings – Science (via Web of Science)

**Date searched: 19 January 2022**

SCI-EXPANDED: 1970–present

CPCI-S: 1990–present

23 (((#17) OR #18) OR #19) OR #22 and English (Languages)3210
22 (#20 OR #21) AND #7 AND #16216

21 ((((TS=("siemens* healthineers*" OR contextflow* OR "jlk inc*" OR arterys* OR qureai* OR "qure ai*" OR sensetime* OR "canon medical*" OR vuno*)) OR OG=("siemens* healthineers*" OR contextflow* OR "jlk inc*" OR arterys* OR qureai* OR "qure ai*" OR sensetime* OR "canon medical*" OR vuno*)) OR AD=("siemens* healthineers*" OR contextflow* OR "jlk inc*" OR arterys* OR qureai* OR "qure ai*" OR sensetime* OR "canon medical*" OR vuno*)) OR FO=("siemens* healthineers*" OR contextflow* OR "jlk inc*" OR arterys* OR qureai* OR "qure ai*" OR sensetime* OR "canon medical*" OR vuno*))2633

20 ((((TS=(coreline* OR riverain* OR infervision* OR fujifilm* OR mevis* OR aidence*)) OR OG=(coreline* OR riverain* OR infervision* OR fujifilm* OR mevis* OR aidence*)) OR AD=(coreline* OR riverain* OR infervision* OR fujifilm* OR mevis* OR aidence*)) OR FO=(coreline* OR riverain* OR infervision* OR fujifilm* OR mevis* OR aidence*))3964

19 TS=(("ai rad companion*" AND chest) OR contextflow* OR "search lung ct*" OR "jld 01k*" OR "qct lung*" OR "sensecare* lung*" OR "visia* ct*" OR vuno)8

18 TS=("aview* lcs*" OR "clearread* ct*" OR "inferread* ct lung*" OR "lung nodule ai*" OR veolity* OR veye)5

17 ((#6) AND #9) AND #163085

16 #10 or #11 or #12 or #13 or #14 or #15655,436

15 TS=("Early Detection of Cancer")2106

14 TS=((lung OR lungs OR pulmon*) NEAR/3 (nodule* OR cancer* OR tumor* OR tumour*) NEAR/3 screen*)6299

13 TS=("Mass Screening")5559

12 TS=(CAT NEAR/2 (scan* OR x-ray* OR xray*))1067

11 TS=(CT OR LDCT)455,518

10 TS=(comput* NEAR/2 tomograph*)361,422

9 #7 OR #8380,001

8 TS=((pulmonary OR lung) NEAR/2 lesion*)14,221

7 TS=((lung OR lungs OR pulmon* OR bronchial) NEAR/3 (nodul* OR cancer* OR neoplas* OR tumor* OR tumour* OR carcino* OR malignan* OR adenocarcinom* OR blastoma*))370,649

6 #1 OR #2 OR #3 OR #4 OR #5901,467

5 TS=("support vector machine*" OR "random forest*" OR "black box learning")133,456

4 TS=(("computer aided" OR "computer assisted") NEAR/2 (diagnosis OR detection))16,891

3 TS=("neural network*" OR convolutional OR CNN OR CNNs)501,511

2 TS=((artificial OR machine OR deep) NEAR/5 (intelligence OR learning OR reasoning))395,814

1 TS=(ai)75,151

The Ovid MEDLINE search strategy was translated for use in the Cochrane Library and Web of Science with the aid of the Polyglot Search Translator: Clark JM, Sanders S, Carter M, Honeyman D, Cleo G, Auld Y, *et al.* Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. *J Med Libr Assoc* 2020;**108**:195–207. https://doi.org/10.5195/jmla.2020.834

**HTA database (via CRD; www.crd.york.ac.uk/CRDWeb/)**

**Date searched: 19 January 2022**

1 MeSH DESCRIPTOR Artificial Intelligence EXPLODE ALL TREES290

2 (ai)202

3 ((artificial OR machine OR deep) ADJ5 (intelligence OR learning OR reasoning))8

4 (neural network* OR convolutional OR CNN OR CNNs)12

5 MeSH DESCRIPTOR Diagnosis, Computer-Assisted EXPLODE ALL TREES108

6 ((computer aided OR computer assisted) ADJ1 (diagnosis OR detection))34

7 (support vector machine* OR random forest* OR black box learning)0

8 (#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7) IN HTA148

9 ((lung* or pulmon*) ADJ3 (nodul* or cancer* or neoplas* or tumor* or tumour* or carcino* or malignan* or adenocarcinom*))1444

10 MeSH DESCRIPTOR Lung Neoplasms EXPLODE ALL TREES1151
11 MeSH DESCRIPTOR Solitary Pulmonary Nodule EXPLODE ALL TREES27
12 (#9 OR #10 OR #11) IN HTA341
13 MeSH DESCRIPTOR Tomography, X-Ray Computed896
14 MeSH DESCRIPTOR Tomography, Spiral Computed EXPLODE ALL TREES75
15 (comput* ADJ2 tomograph*)1395
16 (CT OR LDCT)1231
17 (CAT ADJ2 (scan* OR x-ray* OR xray*))6
18 MeSH DESCRIPTOR Mass Screening2103
19 ((lung OR lungs OR pulmon*) ADJ3 (nodule* OR cancer* OR tumor* OR tumour*) ADJ3 screen*)42
20 MeSH DESCRIPTOR Early Detection of Cancer EXPLODE ALL TREES277
21 (#13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20) IN HTA953
22 #8 AND #12 AND #211

**International HTA database (via INAHTA; https://database.inahta.org/)**

**Date searched: 19 January 2022**

21 #20 AND #14 AND #83
20 #19 OR #16 OR #15417
19 #18 AND #17383
18 nodul* OR cancer* OR neoplas* OR tumor* OR tumour* OR carcino* OR malignan* OR adenocarcinom*3216
17 lung* OR pulmon*866
16 “Lung Neoplasms”[mhe]318
15 “Solitary Pulmonary Nodule”[mh]6
14 #13 OR #12 OR #11 OR #10 OR #92443
13 tomograph* OR radiograph* OR CT OR x-ray* OR xray* OR MRI OR PET813
12 screening1234
11 “Diagnostic Imaging”[mhe]1127
10 “Mass Screening”[mhe]758
9 “Early Detection of Cancer”[mh]71
8 #7 OR #6 OR #5 OR #4 OR #3 OR #2 OR #1189
7 “Artificial Intelligence”[mhe]85
6 “Diagnosis, Computer-Assisted”[mhe]64
5 “Neural Networks, Computer”[mhe]0
4 “artificial intelligence” OR “machine learning” OR “deep learning” OR “deep reasoning” OR “machine reasoning”9
3 “neural network” OR “neural networks” OR convolutional OR CNN OR CNNs5
2 “computer aided” OR “computer assisted”65
1 “support vector machine*” OR “random forest*” OR “black box learning”0

**medRxiv (via medrxivr; https://mcguinlu.shinyapps.io/medrxivr/)**

**Date searched: 19 January 2022**

Advanced search screen:

Topic 1:

[Aa]rtificial [Ii]ntelligence

[Mm]achine [Ll]earning

[Dd]eep [Ll]earning

[Ss]upport [Vv]ector [Mm]achine

\\b[Aa][Ii]\\b

[Nn]eural [Nn]etwork

[Cc]onvolutional

[Rr]andom [Ff]orest

[Bb]lack [Bb]ox [Ll]earning

[Cc]omputer [Aa]ided [Dd]iagnosis

[Cc]omputer [Aa]ssisted [Dd]iagnosis

[Cc]omputer [Aa]ided [Dd]etection

[Cc]omputer [Aa]ssisted [Dd]etection

\\bCNN\\b

\\bCNNs\\b

[Dd]eep [Rr]easoning

[Mm]achine [Rr]easoning

Topic 2:

[Ll]ung

[Pp]ulmon

Topic 3:

[Nn]eoplas

[Cc]ancer

[Nn]odul

[Tt]umor

[Tt]umour

[Cc]arcinoma

[Aa]denocarcinoma

Topic 4:

[Cc]omputed [Tt]omograph

\\bCT\\b

\\bLDCT\\b

screening

Earliest record date:

2016-07-01

Latest record date:

2022-01-19

Remove older versions of the same record

**ClinicalTrials.gov**

**Date searched: 19 January 2022**

Home screen search: https://clinicaltrials.gov/ct2/home

3 studies found for: “aview lcs” OR “aview lcs+” OR “clearread ct” OR “inferread ct lung” OR “inferread lung” OR “lung nodule ai” OR veolity OR veye [Other terms]

10 studies found for: coreline* OR riverain OR infervision OR fujifilm OR aidoc OR mevis OR aidence [‘Other terms’]| lung OR pulmonary [Condition or disease] *(of which 3 studies already found above)*

2 studies found for: “ai rad companion” OR contextflow OR “search lung ct” OR “jld 01k” OR “lung ai” OR “qct lung” OR sensecare OR vuno [Other terms]

5 studies found for: “siemens healthineers” OR jlk OR qureai OR “qure ai” OR sensetime [Other terms]| lung OR pulmonary [Condition or disease]

Total: 17 unique results

**WHO International Clinical Trials Registry Platform (ICTRP) search portal**

**Date searched: 19 January 2022**

Home screen search: https://trialsearch.who.int/Default.aspx

7 records for 7 trials found for: aview lcs* OR clearread ct OR inferread ct lung OR inferread lung OR lung nodule ai OR veolity OR veye

9 records for 9 trials found for: (coreline* OR riverain OR infervision OR fujifilm OR aidoc OR mevis OR aidence) AND (lung OR pulmonary)

9 records for 8 trials found for: ai rad companion OR contextflow OR search lung ct OR jld 01k OR qct lung OR sensecare OR vuno

No results were found for: (siemens healthineers OR jlk OR qureai OR qure ai OR sensetime OR arterys) AND (lung OR pulmonary)

Advanced search screen: https://trialsearch.who.int/AdvSearch.aspx

1 record for 1 trial found for: lung ai [in the intervention]

without synonyms selected; recruitment status is ALL

Total number of trials after 3 duplicates removed (using EndNote): **22**

**NICE website: www.nice.org.uk/**

Date searched: 24 January 2022

Browsed: NICE Guidance > Conditions and diseases > Cancer > Lung cancer: www.nice.org.uk/guidance/conditions-and-diseases/cancer/lung-cancer

found 76 published products, of which **3** downloaded/of potential interest

Searched published guidance: www.nice.org.uk/guidance/published?sp=on

Filters (Guidance programme): Technology appraisal guidance, NICE guidelines, Clinical guidelines, Medical technologies guidance, Diagnostics guidance, Highly specialised technologies guidance, Cancer service guidelines.

lung cancer51 results, of which 1 potentially relevant, already identified above

nodule3 results, of which 1 potentially relevant, already identified above

Searched published guidance: www.nice.org.uk/guidance/published?sp=on

No filters.

artificial intelligence3 results, of which 1 potentially relevant, already identified above

machine learning0 results

deep learning0 results

ai1 result, of which 0 relevant

neural network0 results

Browsed guidance in consultation: www.nice.org.uk/guidance/inconsultation

12 results, 0 relevant to lung cancer/pulmonary nodules or artificial intelligence

**Total unique results downloaded: 3**

**Canadian Agency for Drugs and Technologies in Health (CADTH) website: www.cadth.ca/**

Date searched: 24 January 2022

Search screen: www.cadth.ca/search, results limited to Reports tab.

Search terms:

lung cancer [contains all words]74 results; 8 potentially relevant, of which 1 already identified via bibliographic database searches

nodules nodule [contains any words]9 results; 5 potentially relevant, all 5 already identified above

artificial intelligence [contains all words]31 results; 3 potentially relevant, all 3 already identified above

machine learning [contains all words]17 results; 2 potentially relevant, both already identified above

deep learning [contains all words]11 results; 2 potentially relevant, both already identified above

ai20 results; 2 potentially relevant, both already identified above

neural networks [contains all words]5 results; 1 potentially relevant, already identified above

**Total unique results downloaded: 7**

**ISPOR presentations database: www.ispor.org/heor-resources/presentations-database/search**

Date searched: 25 January 2022

As there was no option to export results in bulk, titles and, where necessary, abstracts were scanned for potential relevance and only those potentially relevant to AI technologies *and* CT imaging *and* lung cancer/pulmonary nodules were retrieved (where not already identified by previous searches).

| Search | Hits | Documents retrieved |
|---|---|---|
| lung cancer AND (tomograph* OR CT OR LDCT OR screening) | 70 | 0 (1 potentially relevant already identified via database searches) |
| pulmonary nodule* AND (tomograph* OR CT OR LDCT OR screening) | 3 | 0 |
| lung nodule* AND (tomograph* OR CT OR LDCT OR screening) | 4 | 0 |
| lung AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR ai OR "neural networks" OR "neural network") | 15 | 0 |
| pulmonary AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR ai OR "neural networks" OR "neural network") | 7 | 0 |
| **Total documents retrieved:** | | **0** |

**Health Technology Assessment International (HTAi) Annual Meetings: https://htai.org/annual-meetings/**

Date searched: 25 January 2022

HTAi 2021 Virtual (Manchester). Full programme available at:

https://htai.org/wp-content/uploads/2021/06/HTAi_AM21_Full-Program.pdf

Searched (Ctrl + F) for:

lung

pulmon

chest

thora

artificial int

learning

neural*nothing relevant found*

HTAi 2020 Beijing (virtual). Poster abstracts and Oral abstracts available from https://htai.eventsair.com/htaibeijing2020

Scanned titles in poster and abstract e-books (no search function available); 1 potentially relevant (oral abstract)

HTAi 2019 Cologne. Abstract book available at

https://htai.org/wp-content/uploads/2019/08/htai_AM19_abstracts_20190812.pdf

Searched (Ctrl + F) for:

lung

pulmon

chest

thora

artificial int

learning

neural*nothing relevant found*

**Total documents retrieved: 1**

**SPIE Proceedings (via SPIE Digital Library; www.spiedigitallibrary.org/)**

Date searched: 26 January 2022

Advanced search screen; search in: Proceedings

("lung cancer" OR "pulmonary nodule") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network") AND (screening OR tomography OR CT OR LDCT)

Refine by: Year 2012-2022

285 results; of which 14 potentially relevant *and* not already identified via the bibliographic database searches

**Annual International Conference of the IEEE Engineering in Medicine & Biology Society (via IEEE Xplore)**

Date searched: 27 January 2022

Command search screen: https://ieeexplore.ieee.org/search/advanced/command

“Parent Publication Number”:1000269 AND ((lung OR pulmonary) NEAR/3 (nodule OR cancer OR neoplas* OR tumor OR tumour OR carcinoma OR malignan* OR adenocarcinoma)) AND (ai OR ((artificial OR machine OR deep) NEAR/5 (intelligence OR learning OR reasoning)) OR “neural network” OR “neural networks” OR convolutional OR CNN OR CNNs OR ((“computer aided” OR “computer assisted”) NEAR/1 (diagnosis OR detection)) OR “support vector machine*” OR “random forest*” OR “black box learning”) AND (tomograph* OR CT OR LDCT OR screening)

14 results; of which 13 already identified via the bibliographic database searches

**1 paper downloaded**

**European Congress of Radiology (via European Society of Radiology website: www.myesr.org/congress/about-ecr/past-congresses)**

Date searched: 31 January 2022

ECR 2021. Abstract book available at https://insightsimaging.springeropen.com/track/pdf/10.1186/s13244-021-01014-5.pdf

ECR 2020. Abstract book available at https://insightsimaging.springeropen.com/track/pdf/10.1186/s13244-020-00851-0.pdf

ECR 2019. Abstract book available at https://insightsimaging.springeropen.com/track/pdf/10.1186/s13244-019-0713-y.pdf

ECR 2018. Abstract book available at https://link.springer.com/article/10.1007/s13244-018-0603-8

ECR 2017. Abstract book available at https://insightsimaging.springeropen.com/track/pdf/10.1007/s13244-017-0546-5.pdf

ECR 2016. Abstract book B – Scientific Sessions and Clinical Trials in Radiology, available at https://link.springer.com/content/pdf/10.1007/s13244-016-0475-8.pdf

ECR 2015. Abstract book B – Scientific Sessions and Late-Breaking Clinical Trials, available at https://link.springer.com/content/pdf/10.1007/s13244-015-0387-z.pdf

ECR 2014. Abstract book B – Scientific Sessions, available at https://link.springer.com/content/pdf/10.1007/s13244-014-0317-5.pdf

Searched (Ctrl + F) for:

lung ca

lung nod

pulmonary nod

artificial int

machine learning

deep learning

neural net

Number of abstracts downloaded (potentially relevant to AI + CT/screening + lung cancer/nodules; obvious phantom studies, prediction models and PET-CT excluded):

2021: 5

2020: 17

2019: 19

2018: 4

2017: 2

2016: 1

2015: 3

2014: 1

**Total: 47** (0 already identified via other searches)

**Radiological Society of North America annual meetings (via RSNA website: www.rsna.org/annual-meeting/future-and-past-meetings)**

Date searched: 1 February 2022

RSNA 2020 meeting programme available at www.rsna.org/-/media/Files/RSNA/Annual-meeting/Program/RSNA-2020-program.ashx

posters: *unable to access posters without an RSNA members' login*

RSNA 2019

scientific sessions available at https://archive.rsna.org/2019/ScienceSessions.pdf

posters: a *list of titles is available, but no abstracts/further details accessible without an RSNA members' login*

RSNA 2018:

scientific sessions available at https://archive.rsna.org/2018/ScienceSessions.pdf

posters and exhibits available at https://archive.rsna.org/2018/PostersandExhibits.pdf

RSNA 2016 meeting programmes

scientific sessions available at https://archive.rsna.org/2016/ScienceSessions.pdf posters and exhibits available at https://archive.rsna.org/2016/PostersandExhibits.pdf

Searched (Ctrl + F) within documents for:

lung ca

lung nod

pulmonary nod

artificial int

machine learning

neural net

deep learning *[except in 2019 & 2018 Scientific Sessions, where there were too many (200+) results to scan]*

RSNA 2017:

No PDF documents available.

Meeting programme available at: http://rsna2017.rsna.org/program/index.cfm

Searched for:

lung cancer

pulmonary nodule

pulmonary nodules

lung nodule

lung nodules

artificial intelligence

machine learning

Number of abstracts downloaded (potentially relevant to AI + CT/screening + lung cancer/nodules; obvious phantom studies, prediction models and PET-CT excluded):

2020: 2

2019: 17

2018: 17

2017: 14

2016: 5

**Total: 55**

### U.S. Food & Drug Administration (FDA) Premarket Notification, Premarket Approval & De novo databases (via FDA website)

Date searched: 14 February 2022

Search interfaces:

- Premarket Approval (PMA) database, ‘Device’ field www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma.cfm
- 510(k) Premarket Notification database, ‘Device Name’ field www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm
- Device Classification Under Section 513(f)(2)(De Novo) database, ‘device name’ field www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/denovo.cfm

| Search terms | PMA database results | 510(k) database results | De novo database results | Documents downloaded (judged to contain potentially useful/relevant information not already identified in previous sets) |
|---|---|---|---|---|
| ai rad companion | 0 | 7 | 0 | 1 |
| aview lcs | 0 | 1 | 0 | 1 |
| clearread | 1 | 2 | 0 | 1 |
| contextflow | 0 | 0 | 0 | |
| search lung | 0 | 0 | 0 | |
| inferread | 0 | 2 | 0 | 1 |
| jld-01k | 0 | 0 | 0 | |
| lung AI | 0 | 3 | 0 | |
| lung nodule | 0 | 4 | 0 | |
| qct lung | 0 | 1 | 0 | |
| search lung | 0 | 0 | 0 | |
| sensecare | 0 | 0 | 0 | |
| veolity | 0 | 1 | 0 | 1 |
| veye | 0 | 0 | 0 | |
| vuno | 0 | 0 | 0 | |
| **Total:** | | | | **5** |

### Websites relating to the technologies of interest/their manufacturers

Dates searched: 15–16 February 2022

AI-Rad Companion Chest CT/Siemens Healthineers

www.siemens-healthineers.com/ searched for ‘AI-Rad Companion’.

**Downloaded 1 ‘White paper’** and checked its references (all potentially relevant references already identified via database searches).

AVIEW LCS+/Coreline Soft. We browsed the following pages at https://www.corelinesoft.com/: ‘AVIEW LCS plus’, ‘AVIEW LCS 2’ and ‘Newsroom’.

0 documents to download

ClearRead CT/Riverain Technologies

www.riveraintech.com/clearread-ai-solutions/clearread-ct/ 1 reference on page, already identified via database searches

www.riveraintech.com/resources/clinical-evidence/#clearread-ct-studies links to 5 papers, of which 1 not already found via database searches; **1 downloaded (Van Leeuwen 2021)**

SEARCH Lung CT/ contextflow

https://contextflow.com/solution/search-for-3d-medical-imaging/ 0 to download

https://contextflow.com/startup-news/ **1 press release mentions not-yet-published study and 1 video presentation about the same study.**

InferRead CT Lung/Infervision. On the website https://global.infervision.com/

we browsed the pages 'InferRead CT Lung' and 'Newsroom'.

0 documents to download

JLD-01K/JLK, Inc.

On the website https://www.jlkgroup.com/en/

we browsed the page 'MEDIHUB Products'; 0 documents to download

Lung AI/Arterys

www.arterys.com/clinicalapp/lungapp – references 'Arterys Lung AI Nodule Detection study – University of California, San Diego' – unable to find this via Google search

www.arterys.com/clinical-evidence – nothing on Lung AI; 0 documents to download

Lung Nodule AI/Fujifilm. Browsed:

www.fujifilm.com/uk/en/healthcare/healthcare-it;

0 documents to download

qCT-Lung/Qure.ai. https://www.qure.ai/:

'QCT Lung' and 'Evidence'.

0 documents to download

SenseCare-Lung Pro/Sensetime. Browsed:

www.sensetime.com/en/product-detail?categoryId=32629

www.sensetime.com/en/news-index

0 documents to download

MeVis/Veolity. Browsed

www.veolity.com/

www.veolity.com/news-events

0 documents to download

Aidence/Veye Lung Nodules

www.aidence.com/veye-lung-nodules/

www.aidence.com/development-clinical-validation/ **2 conference posters and 1 unpublished manuscript downloaded**

www.aidence.com/clinical-research/ 5 articles/reports, of which 1 CQC report not identified via previous searches; **1 document downloaded**

www.aidence.com/resources/

www.aidence.com/articles/ **6 articles downloaded** (including 3 from an external site, 2 of which are in Dutch)

VUNO Med-LungCT AI/VUNO

www.vuno.co/en/lung

www.vuno.co/en/publication/lists/medical_image 10 articles/abstracts of potential interest, of which 2 RSNA abstracts not already identified via other searches; **2 downloaded**

| Forwards citation tracking:Paper EN ID | Web of Science,* searched 26 May 2022 | Google Scholar, searched 30 May 2022 |
|---|---|---|
| Abadia 2021 54 | 0 citations | |
| Cohen 2016 | 28 citations | |
| Cohen 2017 | 12 citations | |
| Hsu 2021 3060 | 3 citations | |
| Hwang 2021 491 | 0 citations | |
| Hwang 2021 662 | 4 citations | |
| Hwang 2021 671 | 5 citations | |
| Jacobs 2021 393 | Not found | 1 citation |
| Kim 2018 1197 | 14 citations | |
| Kozuka 2020 683 | 6 citations | |
| Martins Jarnalo 2021 345 | 2 citations | |
| Milanese 2018 1158 | 12 citations | |
| Park 2022 503 | 2 citations | |
| Park 2022 57 | 0 citations | |

continued

| Forwards citation tracking:Paper EN ID | Web of Science,* searched 26 May 2022 | Google Scholar, searched 30 May 2022 |
|---|---|---|
| Singh 2021 255 | 4 citations | |
| Takaishi 2021 607 | 0 citations | |
| Wan 2020 3913 | 4 citations | |
| Zhang 2021 56 | 0 citations | |
| **Total:** | **96** | **1** |

53 duplicates removed (both within set of 96, and against previous clinical systematic review search results) using EndNote 20.

**Total for screening: 44**

*Science Citation Index Expanded 1970–present, Social Sciences Citation Index 1900–present, Arts & Humanities Citation Index 1975–present, Conference Proceedings Citation Index – Science, 1990–present, Conference Proceedings Citation Index – Social Science & Humanities 1990–present, Emerging Sources Citation Index 2015–present.

## Search strategies for searches to inform the economic model

### *Searches for information on model structures, costs and utility values to inform the economic model*

Search dates and number of records retrieved per source are reported below.

| *Bibliographic databases* | | |
|---|---|---|
| Database | Date searched | Number of records |
| MEDLINE All | 1 December 2021 | 549 |
| EMBASE | 1 December 2021 | 970 |
| NHS EED (CRD) | 1 December 2021 | 122 |
| HTA database (CRD) | 1 December 2021 | 90 |
| INAHTA HTA database | 1 December 2021 | 107 |
| Cost-Effectiveness Analysis registry (Tufts Medical Center) | 1 December 2021 | 33 |
| EconPapers [Research Papers in Economics (RePEc)] | 2 December 2021 | 69 |
| ScHARRHUD | 2 December 2021 | 13 |
| Total number of records retrieved: 1953<br>Duplicates removed (EndNote): 689<br>Final number for screening: 1264 | | |
| *Other sources* | | |
| Source | Date searched | Documents retrieved |
| NICE website | 7 December 2021 | 0 |
| Canadian Agency for Drugs and Technologies in Health (CADTH) website | 7 December 2021 | 4 |

| | | |
|---|---|---|
| Google | 7 December 2021,<br>8 December 2021 | 15, plus 1 ongoing study |
| ISPOR conference presentations | 9 December 2021 | 7; plus 5 posters related to abstracts previously identified |
| HTAi annual meetings | 9 December 2021 | 2 |
| iHEA congresses | 9 December 2021 | 0 (2 potentially relevant abstracts unavailable) |
| Total number sought for retrieval: 35 (+ 1 ongoing study)<br>Reports not retrieved/available: 2 (iHEA abstracts)<br>Final number for screening: 33 (+ 1 ongoing study) | | |

Search strategies used:

**MEDLINE All**

**Date searched: 1 December 2021**

Ovid MEDLINE(R) ALL 1946–30 November 2021

1 exp Lung Neoplasms/dg or Solitary Pulmonary Nodule/dg26,945
2 exp Lung Neoplasms/ or Solitary Pulmonary Nodule/253,279
3 ((lung or lungs or pulmon* or bronchial) adj3 (nodul* or cancer* or neoplas* or tumor* or tumour* or carcino* or malignan* or adenocarcinom* or blastoma*)).kf,tw.271,939
4 ((pulmonary or lung) adj2 lesion*).kf,tw.14,650
5 2 or 3 or 4357,079
6 Mass Screening/111,107
7 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw.4748
8 “Early Detection of Cancer”/31,301
9 exp Radiography, Thoracic/ or Diagnostic Imaging/ or exp Image Interpretation, Computer-Assisted/ or exp Positron Emission Tomography Computed Tomography/ or exp Tomography, Emission-Computed/ or exp Tomography, X-Ray/665,323
10 (radiograph* or tomograph* or imaging or x-ray* or xray* or CT or PET or PET-CT or MRI or (CAT adj2 scan*)).kf,tw.2,037,778
11 6 or 7 or 8 or 9 or 102,374,261
12 5 and 1168,258
13 1 or 12 [lung neoplasms; diagnostic imaging or screening]74,051
14 *economics/10,766
15 exp *“costs and cost analysis”/76,423
16 (economic adj2 model*).mp.14,167
17 (cost minimi* or cost-utilit* or health utilit* or economic evaluation* or economic review* or cost outcome* or cost analys?s or economic analys?s or budget* impact analys?s).ti,ab,kf,kw.37,484
18 (cost-effective* or pharmacoeconomic* or pharmaco-economic* or cost-benefit or costs).ti,kf,kw.79,637
19 (life year or life years or qaly* or cost-benefit analys?s or cost-effectiveness analys?s).ab,kf,kw.34,382
20 (cost or economic*).ti,kf,kw. and (costs or cost-effectiveness or markov or monte carlo or model or modeling or modelling).ab.74,307
21 or/14-20 [CADTH Narrow Economic Filter – OVID Medline, EMBASE https://www.cadth.ca/strings-attached-cadths-database-search-filters]201,764
22 13 and 21481
23 Quality-Adjusted Life Years/14,121
24 (quality adjusted or adjusted life year*).ti,ab,kf.19,799
25 (qaly* or qald* or qale* or qtime*).ti,ab,kf.12,541
26 (illness state*1 or health state*1).ti,ab,kf.7368

27 (hui or hui1 or hui2 or hui3).ti,ab,kf.1749
28 (multiattribute* or multi attribute*).ti,ab,kf.1057
29 (utility adj3 (score*1 or valu* or health* or cost* or measur* or disease* or mean or gain or gains or index*)).ti,ab,kf.17,499
30 utilities.ti,ab,kf.8178
31 (eq-5d or eq5d or eq-5 or eq5 or euro qual or euroqual or euro qual5d or euroqual5d or euro qol or euroqol or euro qol5d or euroqol5d or euro quol or euroquol or euro quol5d or euroquol5d or eur qol or eurqol or eur qol5d or eur qol5d or eur?qul or eur?qul5d or euro* quality of life or European qol).ti,ab,kf.14,119
32 (euro* adj3 (5 d or 5d or 5 dimension* or 5dimension* or 5 domain* or 5domain*)).ti,ab,kf.4937
33 (sf36* or sf 36* or sf thirtysix or sf thirty six).ti,ab,kf.24,278
34 (time trade off*1 or time tradeoff*1 or tto or timetradeoff*1).ti,ab,kf.2105
35 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 [Filter FSF3 - precision maximizing filter to identify HSU studies, from Arber *et al.*, 2017 https://doi.org/10.1017/S0266462317000897}81,449
36 13 and 35193
37 22 or 36549

**EMBASE**

**Date searched: 1 December 2021**

EMBASE 1974–30 November 2021

1 exp lung cancer/di or lung nodule/di46,425
2 ((lung or lungs or pulmon* or bronchial) adj3 (nodul* or cancer* or neoplas* or tumor* or tumour* or carcino* or malignan* or adenocarcinom* or blastoma*)).kf,tw.388,958
3 ((pulmonary or lung) adj2 lesion*).kf,tw.20,844
4 1 or 2 or 3416,579
5 mass screening/ or cancer screening/141,880
6 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw.7543
7 early cancer diagnosis/9533
8 diagnostic imaging/ or exp thorax radiography/ or computer assisted tomography/ or low-dose computed tomography/ or exp x-ray computed tomography/ or multidetector computed tomography/ or spiral computer assisted tomography/ or exp computer assisted emission tomography/1,351,059
9 (radiograph* or tomograph* or imaging or x-ray* or xray* or CT or PET or PET-CT or MRI or (CAT adj2 scan*)).kf,tw.2,769,230
10 5 or 6 or 7 or 8 or 93,410,416
11 4 and 10113,394
12 *economics/27,332
13 exp *"costs and cost analysis"/84,204
14 (economic adj2 model*).mp.8559
15 (cost minimi* or cost-utilit* or health utilit* or economic evaluation* or economic review* or cost outcome* or cost analys?s or economic analys?s or budget* impact analys?s).ti,ab,kf,kw.57,878
16 (cost-effective* or pharmacoeconomic* or pharmaco-economic* or cost-benefit or costs).ti,kf,kw.117,531
17 (life year or life years or qaly* or cost-benefit analys?s or cost-effectiveness analys?s).ab,kf,kw.53,133
18 (cost or economic*).ti,kf,kw. and (costs or cost-effectiveness or markov or monte carlo or model or modeling or modelling).ab.112,254
19 or/12-18 [CADTH Narrow Economic Filter – OVID Medline, EMBASE https://www.cadth.ca/strings-attached-cadths-database-search-filters]286,393
20 11 and 19767
21 Quality-Adjusted Life Years/30,198
22 (quality adjusted or adjusted life year*).ti,ab,kf.28,814
23 (qaly* or qald* or qale* or qtime*).ti,ab,kf.23,274
24 (illness state*1 or health state*1).ti,ab,kf.12,756

25 (hui or hui1 or hui2 or hui3).ti,ab,kf.2685
26 (multiattribute* or multi attribute*).ti,ab,kf.1305
27 (utility adj3 (score*1 or valu* or health* or cost* or measur* or disease* or mean or gain or gains or index*)).ti,ab,kf.27,682
28 utilities.ti,ab,kf.13,218
29 (eq-5d or eq5d or eq-5 or eq5 or euro qual or euroqual or euro qual5d or euroqual5d or euro qol or euroqol or euro qol5d or euroqol5d or euro quol or euroquol or euro quol5d or euroquol5d or eur qol or eurqol or eur qol5d or eur qol5d or eur?qul or eur?qul5d or euro* quality of life or European qol).ti,ab,kf.25,481
30 (euro* adj3 (5 d or 5d or 5 dimension* or 5dimension* or 5 domain* or 5domain*)).ti,ab,kf.7449
31 (sf36* or sf 36* or sf thirtysix or sf thirty six).ti,ab,kf.41,638
32 (time trade off*1 or time tradeoff*1 or tto or timetradeoff*1).ti,ab,kf.3088
33 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 [Filter FSF3 - precision maximizing filter to identify HSU studies, from Arber *et al.*, 2017 https://doi.org/10.1017/S0266462317000897}133,355
34 11 and 33400
35 20 or 34970

**NHS EED and HTA database (CRD) www.crd.york.ac.uk/CRDWeb/HomePage.asp**

**Date searched: 1 December 2021**

| Line | Search | Hits |
|---|---|---|
| 1 | ((lung* or pulmon*) ADJ3 (nodul* or cancer* or neoplas* or tumor* or tumour* or carcino* or malignan* or adenocarcinom*)) | 1444 |
| 2 | MeSH DESCRIPTOR Lung Neoplasms EXPLODE ALL TREES | 1151 |
| 3 | MeSH DESCRIPTOR Solitary Pulmonary Nodule EXPLODE ALL TREES | 27 |
| 4 | (#1) OR (#2) OR (#3) IN NHSEED, HTA | 677 |
| 5 | MeSH DESCRIPTOR Diagnostic Imaging EXPLODE ALL TREES | 4336 |
| 6 | (screening) | 5030 |
| 7 | MeSH DESCRIPTOR Mass Screening EXPLODE ALL TREES | 2347 |
| 8 | MeSH DESCRIPTOR Early Detection of Cancer EXPLODE ALL TREES | 277 |
| 9 | (tomograph* OR radiograph* OR CT OR x-ray* OR xray* OR MRI OR PET) | 4288 |
| 10 | (#5 OR #6 OR #7 OR #8 OR #9) IN NHSEED, HTA | 5965 |
| 11 | (#4 AND #10) IN NHSEED | 122 |
| 12 | (#4 AND #10) IN HTA | 90 |

**INAHTA HTA database**

**Date searched: 1 December 2021**

| Line | Query | Hits |
|---|---|---|
| 75 | #74 AND #66 | 107 |
| 74 | #73 OR #72 OR #71 OR #70 OR #69 OR #68 OR #67 | 2412 |
| 73 | “Early Detection of Cancer”[mh] | 71 |
| 72 | “Mass Screening”[mhe] | 751 |

continued

| Line | Query | Hits |
|---|---|---|
| 71 | (screening)[Title] OR (screening)[abs] OR (screening)[Keywords] | 1222 |
| 70 | "Diagnostic Imaging"[mhe] | 1124 |
| 69 | (tomograph* OR radiograph* OR CT OR x-ray* OR xray* OR MRI OR PET)[Keywords] | 14 |
| 68 | (tomograph* OR radiograph* OR CT OR x-ray* OR xray* OR MRI OR PET)[abs] | 591 |
| 67 | (tomograph* OR radiograph* OR CT OR "x-ray*" OR xray* OR MRI OR PET)[Title] | 461 |
| 66 | #65 OR #64 OR #63 OR #62 OR #61 | 415 |
| 65 | "Solitary Pulmonary Nodule"[mh] | 6 |
| 64 | "Lung Neoplasms"[mhe] | 317 |
| 63 | ((lung* OR pulmon*) AND (nodul* OR cancer* OR neoplas* OR tumor* OR tumour* OR carcino* OR malignan* OR adenocarcinom*))[Keywords] | 15 |
| 62 | ((lung* OR pulmon*) AND (nodul* OR cancer* OR neoplas* OR tumor* OR tumour* OR carcino* OR malignan* OR adenocarcinom*))[abs] | 243 |
| 61 | (lung* OR pulmon*)[Title] AND (nodul* OR cancer* OR neoplas* OR tumor* OR tumour* OR carcino* OR malignan* OR adenocarcinom*)[Title] | 278 |

**Cost-effectiveness Analysis Registry (via Tufts Medical Center website) https://cevr.tuftsmedicalcenter.org/databases/cea-registry**

Date searched: 1 December 2021

Basic search screen: Methods selected

Results of each search were copied and pasted into Microsoft Excel®, to easily identify unique results, which were then found in PubMed for easy export/import into EndNote.

| Search term/s | Results |
|---|---|
| lung nodule | 0 |
| pulmonary nodule | 9 |
| lung cancer screening | 19 |
| lung CT | 1 |
| lung computed tomography | 1 |
| chest CT | 4 |
| chest computed tomography | 5 |
| thoracic CT | 0 |
| thoracic computed tomography | 0 |
| thorax CT | 0 |
| thorax computed tomography | 0 |
| lung imaging | 0 |
| lung radiography | 0 |
| lung x-ray | 0 |

| Search term/s | Results |
|---|---|
| lung xray | 0 |
| **Total:** | **39** |
| **Total unique results (after deduplication in Excel)** | **33** |

**EconPapers (via Research Papers in Economics (RePEc)) https://econpapers.repec.org/**

Date searched: 2 December 2021

Advanced search screen: https://econpapers.repec.org/scripts/search.pf

**50** documents matched the search for ("pulmonary nodule*" OR "lung nodule*" OR "lung cancer") AND (tomograph* OR radiograph* OR CT OR x-ray* OR xray* OR MRI OR PET OR screening) in titles and keywords in working papers, articles, books and chapters.

**19** documents matched the search for ("artificial intelligence" OR "machine learning" OR "deep learning" OR "support vector machine*" OR "neural network*" OR "random forest" OR "black box learning") AND ("pulmonary nodule*" OR "lung nocule*" OR "lung cancer*") AND (CT OR "computed tomography" OR screening) in working papers, articles, books and chapters. *[Free text search]*

**Total: 69 records**

**ScHARRHUD www.scharrhud.org/index.php?recordsN1&m=search**

Date searched: 2 December 2021

(lung OR lungs OR pulmonary) AND (nodule OR nodules OR cancer OR cancers OR neoplasm OR neoplasms OR tumor OR tumors OR tumour OR tumours OR carcinoma OR carcinomas OR malignancy OR malignancies OR malignant OR adenocarcinoma OR adenocarcinomas) **13 results**

**NICE website www.nice.org.uk/**

Date searched: 7 December 2021

Browsed: NICE Guidance > Conditions and diseases > Cancer > Lung cancer: www.nice.org.uk/guidance/conditions-and-diseases/cancer/lung-cancer

found 75 published products, of which none included economic evaluation of diagnostic imaging

Searched published guidance: www.nice.org.uk/guidance/published?sp=on

Filters: Diagnostics guidance, Technology appraisal guidance

lung cancer 48results, of which 0 relevant

nodule0 results

nodules0 results

Browsed guidance in consultation: www.nice.org.uk/guidance/inconsultation

20 results, 0 relevant to lung cancer/pulmonary nodules

Searched guidance in development: www.nice.org.uk/guidance/indevelopment

Filters: Diagnostics guidance, Technology appraisal guidance

lung cancer51 results, of which 0 relate to diagnostic imaging

nodule1 result; 0 unique results

nodules1 result; 0 unique results

**Canadian Agency for Drugs and Technologies in Health (CADTH) website www.cadth.ca/**

Date searched: 7 December 2021

Search box on homepage, results limited to Reports tab.

Search terms:

lung cancer76 results; 6 on imaging;of which 1 not a cost-effectiveness/economic evaluation; 1 already retrieved by database searches; **4 reports retrieved**

nodules 7 results; 3 on imaging; all 3 already identified above

**Google www.google.co.uk**

Dates searched: 7–8 December 2021

Results (10 per page) were browsed until yielding very few results containing all search terms.

Documents were retrieved if judged to be potentially useful, and if they had not already been identified via the database searches or earlier Google searches. Documents without English-language abstracts were also excluded.

| Search string | Number of results browsed | Documents retrieved |
|---|---|---|
| lung nodules HTA imaging OR diagnosis OR detection OR screening | 30 | 3 (Department of Health, ECRI, Ministry of Health) |
| pulmonary nodules HTA imaging OR diagnosis OR detection OR screening | 22 | 0 |
| lung cancer HTA imaging OR diagnosis OR detection OR screening | 30 | 3 [2 x HTA Austria reports; 1 review (van Meerbeeck 2021)] |
| lung nodules HTA CT OR tomography OR radiography OR xray OR PET | 47 | 0 |
| lung cancer HTA CT OR tomography OR radiography OR xray OR PET | 50 | 1 (Bucher 2020) |
| lung nodules economic imaging OR diagnosis OR detection OR screening OR CT OR tomography OR radiography OR xray OR PET | 50 | 1 ongoing study<br>4 (LeMense 2020, Edelman Saul 2020, Pyenson 2019, Gilbert 2018) |

| Search string | Number of results browsed | Documents retrieved |
|---|---|---|
| lung cancer economic imaging OR diagnosis OR detection OR screening OR CT OR tomography OR radiography OR xray OR PET | 50 | 2 (Health Policy Partnership, EEPRU) |
| lung nodules cost effectiveness imaging OR diagnosis OR detection OR screening OR CT OR tomography OR radiography OR xray OR PET | 50 | 2 (Lu 2014, Gilbert 2021) |
| **Total documents retrieved:** | | **15; plus 1 ongoing study** |

**ISPOR presentations database www.ispor.org/heor-resources/presentations-database/search**

Date searched: 9 December 2021

As there was no option to export results in bulk, titles and, where necessary, abstracts were scanned for potential relevance and only those including economic evaluation, costs or utilities information for diagnostic imaging of lung cancer/pulmonary nodules were retrieved (where not already identified by previous searches).

| Search | Hits | Documents retrieved |
|---|---|---|
| lung cancer AND (imaging OR tomograph* OR radiograph* OR CT OR “x-ray*” OR xray* OR MRI OR PET OR screening) | 73 | 7 unique results, plus:<br>5 posters related to abstracts already identified via database searches |
| pulmonary nodule* AND (imaging OR tomograph* OR radiograph* OR CT OR “x-ray*” OR xray* OR MRI OR PET OR screening) | 3 | 0 |
| lung nodule* AND (imaging OR tomograph* OR radiograph* OR CT OR “x-ray*” OR xray* OR MRI OR PET OR screening) | 5 | 0 |
| **Total documents retrieved:** | | **7; plus 5 posters related to abstracts previously identified** |

**Health Technology Assessment International (HTAi) Annual Meetings https://htai.org/annual-meetings/**

Date searched: 9 December 2021

HTAi 2021 Virtual (Manchester). Full programme available at:

https://htai.org/wp-content/uploads/2021/06/HTAi_AM21_Full-Program.pdf

Searched (Ctrl + F) for:

lung

pulmon

chest

thoran*othing relevant found*

HTAi 2020 Beijing (virtual). Poster abstracts and Oral abstracts available from: https://htai.eventsair.com/htaibeijing2020

Scanned titles in poster and abstract e-books (no search function available); *nothing relevant found*

HTAi 2019 Cologne. Abstract book available at:

https://htai.org/wp-content/uploads/2019/08/htai_AM19_abstracts_20190812.pdf

Searched (Ctrl + F) for:

lung*2 abstracts retrieved*

pulmon*nothing relevant found*

chest*nothing relevant found*

thora`*nothing relevant found*

**International Health Economics Association (iHEA) Congresses www.healtheconomics.org/page/PastCongresses**

**Abstracts not available**

Date searched: 9 December 2021

Searched (Ctrl + F) for:

- lung
- pulmon
- chest
- thora

in all of the following:

Beijing 2009. Programme available at www.healtheconomics.org/wp-content/uploads/2022/07/ihea-2009-beijing-programme.pdf

Toronto 2011. Programme available at www.healtheconomics.org/wp-content/uploads/2022/07/ihea-2011-toronto-programme.pdf

Sydney 2013. Programme available at www.healtheconomics.org/wp-content/uploads/2022/07/ihea-2013-sydney-programme.pdf

Dublin 2014. Programme available at www.healtheconomics.org/wp-content/uploads/2022/07/ihea-2014-dublin-programme.pdf

Milan 2015. Programme available at www.healtheconomics.org/wp-content/uploads/2022/07/ihea-2015-milan-programme.pdf

Boston 2017. Programme available at www.healtheconomics.org/wp-content/uploads/2022/07/iHEA_Program_2017.pdf

Basel 2019. Programme available at www.healtheconomics.org/wp-content/uploads/2022/07/Basel-2019-Program.pdf

2 potentially relevant presentations identified (both from Boston 2017):

Title: Cost Utility Analysis of Lung Cancer Screening for High Risk Patients in Germany

Presenter: Florian Hofer, Hamburg Center for Health Economics, Germany

Author(s): Tom Stargard

*no abstract available, but a full journal article with very similar authors and title was identified via the database searches (Endnote ID #148)*

Title: Risk Stratified Lung Cancer Screening – A Cost-Effectiveness Analysis

Presenter: Vaibhav Kumar, Tufts Medical Center, USA

Author(s): Joshua T Cohen, David van Klaveren, Djøra I Soeteman, John Wong, Peter J Neumann, David M Kent

*no abstract available, but a full journal article with very similar authors and title was identified via the database searches (Endnote ID #169)*

0 documents retrieved.

## Searches for pulmonary nodule growth rates/volume doubling times

Search dates and number of records retrieved per source are reported below.

| Database/source | Date searched | Number of results |
|---|---|---|
| MEDLINE | 2 March 2022 | 375 |
| EMBASE | 2 March 2022 | 517 |
| CISNET website: publications list | 3 March 2022 | 144 |
| *Total:* | | 1036 |
| *Total after deduplication within set:* | | 810 |
| *Total after deduplication against previous search (economics SLR):* | | **786** |
| Internet (Google) and website (NCCN, NHS Digital, plus others identified via Google) searches, 3–9 March 2022 | | 10 potentially relevant documents retrieved (9 articles, 1 conference abstract)<br>0 potentially useful registries/websites identified<br>2 ongoing studies of potential interest identified (IDEAL, Watch the Spot) |
| Google Dataset Search, 29–30 March 2022 | | 1 potentially relevant data set retrieved |

Search strategies used:

**MEDLINE via Ovid**

Date searched: 2 March 2022

Ovid MEDLINE(R) ALL 1946–1 March 2022

1 (growth rate* or growth curve* or doubling time*).kf,tw.95,469
2 Lung Neoplasms/di, dg50,814
3 Solitary Pulmonary Nodule/4475
4 (lung nodule* or pulmonary nodule*).kf,tw.11,482
5 2 or 3 or 458,727
6 1 and 5375

**EMBASE via Ovid**

Date searched: 2 March 2022

EMBASE Classic+EMBASE 1947–1 March 2022

1 (growth rate* or growth curve* or doubling time*).mp.139,373
2 exp lung cancer/di [Diagnosis]43,090
3 lung nodule/24,693
4 (lung nodule* or pulmonary nodule*).kf,tw.19,482
5 2 or 3 or 468,270
6 1 and 5517

**CISNET: Cancer Intervention and Surveillance Modeling Network https://cisnet.cancer.gov/**

Date searched: 3 March 2022

Publications list – Lung: https://cisnet.cancer.gov/publications/cancer-site.html#lung_header

144 publications listed. Citations retrieved using Citation Finder https://citation-finder.vercel.app/

**Google (Chrome browser) 3 March 2022**

search terms: list patient registries*browsed first 30 results. Checked:*

www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries

> https://epi.grants.cancer.gov/cancer-registries/

> https://cancer.ca/en/

www.ema.europa.eu/en/documents/regulatory-procedural-guideline/encepp-resource-database-inventory-patient-registries_en.pdf

www.encepp.eu/encepp/search.htm searched:

Data source > 'lung cancer'

lung

nodule

cancer*nothing relevant found*

www.ncri.ie/*has good survival statistics, but nothing on growth*

www.infodesk.com/blog/types-of-patient-registries-and-where-to-find-them/life-sciences

> CDC resources: browsed www.cdc.gov/cancer/lung/ *lung cancer stats are available (USCS) but not growth rates.*

CDC search box: pulmonary nodules *nothing relevant* https://web.archive.org/web/20220302143708/www.cdc.gov/cancer/npcr/meaningful_use.htm

> www.pcori.org/

browsed www.pcori.org/topics/cancer

search box: nodules:

> Watch the Spot:

www.pcori.org/research-results/pcori-literature/methods-watch-spot-trial-pragmatic-trial-more-vs-less-intensive-strategies-active-surveillance-small-pulmonary-nodules

www.pcori.org/research-results/2015/comparing-more-versus-less-frequent-monitoring-diagnose-lung-cancer-early-watch-spot-trial

*this ongoing trial may be of interest*

www.eunethta.eu/parent/ *appears to be closed; links are dead*

www.ncra-usa.org/Advocacy/IMSWR/List-of-Medical-Registries

www.safetyandquality.gov.au/publications-and-resources/australian-register-clinical-registries

search box:

lung

> https://vlcr.org.au/

pulmonary

Sorted by 'prioritised clinical domain' and scanned list*nothing relevant*

**Google (Chrome browser) 7** March 2022

search terms: pulmonary nodule growth dataset OR registry OR audit*browsed first 30 results. Checked:*

BTS guidelines www.brit-thoracic.org.uk/document-library/guidelines/pulmonary-nodules/bts-guidelines-for-the-investigation-and-management-of-pulmonary-nodules/ pages ii18–20; *all relevant references identified by MEDLINE/EMBASE searches*

IDEAL study:

https://thorax.bmj.com/content/thoraxjnl/75/4/306.full.pdf

https://clinicaltrials.gov/ct2/show/NCT03753724

https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0044-3

*this ongoing trial may be of interest*

search terms: diagnostic radiology professional bodies*browsed first 10 results. Checked:*

www.bir.org.uk/useful-information/professional-links.aspx

search box:

pulmonary nodules

registry

audit lung

nodule surveillance

> National Lung Cancer Audit: https://nlca.rcp.ac.uk/Home/Index *has good survival statistics, but nothing on growth*

www.rcr.ac.uk/

search box:

pulmonary nodule

registry

audit lung*nothing relevant found*

https://ektron.rsna.org/Radiology-Organizations/

*browsed and/or searched for 'pulmonary nodules' and 'lung cancer' on each of these listed sites:*

www.theabr.org/

www.acr.org/ *2 'incidental findings' papers on adherence/real life follow up may be of interest*

www.ahra.org/Default.aspx

https://car.ca/

www.myesr.org/

www.myesti.org/

https://fleischner.memberclicks.net/

www.hkcr.org/

www.icimagingsociety.org.uk/

www.iria.in/

www.isradiology.org/

www.ranzcr.com/

www.radiology.ie/

www.rsna.org/

www.rssa.co.za/ *need membership to access most documents*

www.scardweb.org/ *need membership to access 'Resources' section*

https://siim.org/

https://srs.org.sg/

https://thoracicrad.org/

*nothing relevant found*

**Google (Chrome browser) 9 March 2022**

search terms: pulmonary nodule natural history database OR registry OR audit*browsed first 50 results. Checked:*

https://clinicaltrials.gov/ct2/show/NCT01540552

> www.ncbi.nlm.nih.gov/pmc/articles/PMC4405280/

*potentially relevant study/article*

www.frontiersin.org/articles/10.3389/fonc.2020.00318/full – *potentially relevant study/article*

www.appliedradiology.com/articles/rsna-2019-tracking-improves-follow-up-imaging-compliance-in-incidental-lung-nodules

> additional Google search: national jewish health lung nodule registry

> www.nationaljewish.org/directory/lung-nodule-registry-program

>https://doi.org/10.1016/j.jacr.2021.01.018

> www.jtocrr.org/article/S2666-3643(22)00021-2/pdf – *includes nothing on nodule growth but they should be able to assess this from their registry data…?*

https://ascopubs.org/doi/abs/10.1200/JCO.2021.39.15_suppl.1564 *conference abstract, mainly about increasing follow up*

search terms: pulmonary nodule surveillance dataset OR registry OR audit *browsed first 30 results. Checked:*

www.ncbi.nlm.nih.gov/pmc/articles/PMC6784443/ – *potentially useful paragraph: 'Nodule growth rate' – checked references:*

>ACCP guidelines – *see section 4.5 'CT Scan Surveillance') – checked references:*

>https://pubmed.ncbi.nlm.nih.gov/10942328/ *potentially relevant article*

https://pubs.rsna.org/doi/full/10.1148/radiol.2017151022#_i27 *potentially useful section on 'Clinical Applicability of Volumetry in Nodule Management' – checked references:*

>https://erj.ersjournals.com/content/42/6/1706 – *potentially useful; see table 1*

>https://doi.org/10.1016/0007-0971(79)90002-0 – *potentially useful*

https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/1857093 – *not about growth but may be of interest because looks at resource use*

https://doi.org/10.1016/S0169-5002(19)30071-6 – *potentially useful conference abstract*

**Additional websites and databases: 9 March 2022**

https://data.gov.uk/

searched (topic: health):

lung cancer

lung nodules

pulmonary nodules

nodule

*nothing relevant found*

National Comprehensive Cancer Network www.nccn.org/

search box:

pulmonary nodules

nodule

lung ct

lung computed tomography

Browsed ‘Education & Research’

browsed ‘Shared Resources’ database

*nothing relevant found*

NHS Digital https://digital.nhs.uk/

search box:

pulmonary nodules

nodule

lung cancer

*nothing relevant to growth rates*

ICPSR (Inter-university Consortium for Political and Social Research) www.icpsr.umich.edu/web/pages/

search box:

lung nodules

“pulmonary nodule”

“computed tomography”

“lung cancer”

*nothing relevant found*

UK Data Service https://ukdataservice.ac.uk/

search box ‘search our data catalogue’:

lung cancer

pulmonary nodules

nodule

*nothing relevant found*

**Google Dataset Search https://datasetsearch.research.google.com/ (Chrome browser)**

**29–30 March 2022**

pulmonary nodule growth rate20 data sets found1 *potentially relevant dataset downloaded*

pulmonary nodules doubling time2 results; both already found above

lung nodules doubling time same2 results retrieved

## Searches for pulmonary nodule prevalence by size and type

Search dates and number of records retrieved per source are reported below.

| Database/source | Date searched | Results (titles/ abstracts) screened | Results selected as potentially relevant |
|---|---|---|---|
| MEDLINE | 30 June 2022 | 228 | 20 |
| Google | 23 June 2022 | 20 | 1, plus section of BTS guideline on prevalence (see below) |
| Reference checking from BTS guideline | 23 June 2022 | 32 | 8 |
| *Total* | | *280* | ***29*** |

Search strategies used:

**MEDLINE via Ovid**

Date searched: 30 June 2022

Database: Ovid MEDLINE(R) ALL 1946–29 June 2022

1 exp Lung Neoplasms/dg (27,068)

2 Solitary Pulmonary Nodule/di, dg (3694)

3 ((lung or lungs or pulmon* or bronchial) adj3 (nodul* or cancer* or neoplas* or tumor* or tumour* or carcino* or malignan* or adenocarcinom* or blastoma*)).kf,tw. (283,697)

4 1 or 2 or 3 [lung cancer or SPNs] (293,093)

5 Mass Screening/ (113,855)

6 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw. (5092)

7 5 or 6 [screening] (117,282)

8 Tomography, X-Ray Computed/ or exp Tomography, Spiral Computed/ (424,801)

9 (comput* adj2 tomograph*).kf,tw. (359,889)

10 (CT or LDCT).kf,tw. (402,762)

11 8 or 9 or 10 [CT] (782,143)

12 Prevalence/ (332,019)

13 “prevalen*”.kf,tw. (895,491)

14 12 or 13 [prevalence] (975,826)

15 Incidental Findings/ (11,566)

16 (incidental* adj2 (finding* or found or discover* or diagnos* or detect*)).kf,tw. (29,485)

17 “incidentaloma*”.kf,tw. (2592)

18 15 or 16 or 17 [incidental findings] (36,802)

19 4 and 7 and 11 and 14 [lung ca/PN screening CT prevalence] (337)

20 (pulmonary nodule* or lung nodule*).kf,tw. (11,812)

21 2 or 20 [PNs - not Ca] (12,891)

22 11 and 14 and 21 [PNs prevalence CT] (316)

23 4 and 11 and 18 [lung ca/PNs CT Incidental findings] (1007)

24 19 or 22 or 23 (1499)

25 exp United Kingdom/ (385,304)

26 (national health service* or nhs*).ab,in,ti. (247,302)

27 (english not ((published or publication* or translat* or written or language* or speak* or literature or citation*) adj5 english)).ab,ti. (45,087)

28 (gb or “g.b.” or britain* or (british* not “british columbia”) or uk or “u.k.” or united kingdom* or (england* not “new england”) or northern ireland* or northern irish* or scotland* or scottish* or ((wales or “south wales”) not “new south wales”) or welsh*).ab,in,jw,ti. (2,322,787)

29 (bath or “bath’s” or ((birmingham not alabama*) or (“birmingham’s” not alabama*) or bradford or “bradford’s” or brighton or “brighton’s” or bristol or “bristol’s” or carlisle* or “carlisle’s” or (cambridge not (massachusetts* or boston* or harvard*)) or (“cambridge’s” not (massachusetts* or boston* or harvard*)) or (canterbury not zealand*) or (“canterbury’s” not zealand*) or chelmsford or “chelmsford’s” or chester or “chester’s” or chichester or “chichester’s” or coventry or “coventry’s” or derby or “derby’s” or (durham not (carolina* or nc)) or (“durham’s” not (carolina* or nc)) or ely or “ely’s” or exeter or “exeter’s” or gloucester or “gloucester’s” or hereford or “hereford’s” or hull or “hull’s” or lancaster or “lancaster’s” or leeds* or leicester or “leicester’s” or (lincoln not nebraska*) or (“lincoln’s” not nebraska*) or (liverpool not (new south wales* or nsw)) or (“liverpool’s” not (new south wales* or nsw)) or ((london not (ontario* or ont or toronto*)) or (“london’s” not (ontario* or ont or toronto*)) or manchester or “manchester’s” or (newcastle not (new south wales* or nsw)) or (“newcastle’s” not (new south wales* or nsw)) or norwich or “norwich’s” or nottingham or “nottingham’s” or oxford or “oxford’s” or peterborough or “peterborough’s” or plymouth or “plymouth’s” or portsmouth or “portsmouth’s” or preston or “preston’s” or ripon or “ripon’s” or salford or “salford’s” or salisbury or “salisbury’s” or sheffield or “sheffield’s” or southampton or “southampton’s” or st albans or stoke or

"stoke's" or sunderland or "sunderland's" or truro or "truro's" or wakefield or "wakefield's" or wells or westminster or "westminster's" or winchester or "winchester's" or wolverhampton or "wolverhampton's" or (worcester not (massachusetts* or boston* or harvard*)) or ("worcester's" not (massachusetts* or boston* or harvard*)) or (york not ("new york*" or ny or ontario* or ont or toronto*)) or ("york's" not ("new york*" or ny or ontario* or ont or toronto*))))).ab,in,ti. (1,633,647)

30 (bangor or "bangor's" or cardiff or "cardiff's" or newport or "newport's" or st asaph or "st asaph's" or st davids or swansea or "swansea's").ab,in,ti. (65,320)

31 (aberdeen or "aberdeen's" or dundee or "dundee's" or edinburgh or "edinburgh's" or glasgow or "glasgow's" or inverness or (perth not australia*) or ("perth's" not australia*) or stirling or "stirling's").ab,in,ti. (240,883)

32 (armagh or "armagh's" or belfast or "belfast's" or lisburn or "lisburn's" or londonderry or "londonderry's" or derry or "derry's" or newry or "newry's").ab,in,ti. (31,250)

33 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 (2,915,825)

34 (exp africa/ or exp americas/ or exp antarctic regions/ or exp arctic regions/ or exp asia/ or exp australia/ or exp oceania/) not (exp United Kingdom/ or europe/) (3,215,213)

35 33 not 34 [UK search filter, Ayiku *et al.* 2017 https://onlinelibrary.wiley.com/doi/10.1111/hir.12187} (2,762,901)

36 **24 and 35 (114)**

37 **from 36 keep 6,10,15,23,36,40,44,46-47,62,65,69,99,102,114 (15)**

38 ((larger or smaller or bigger or greater or more than or less than) adj4 mm).tw. (48,708)

39 ((larger or smaller or bigger or greater or more than or less than) adj4 millimet*).tw. (718)

40 21 and (38 or 39) [PNs - size] (346)

41 (nodule* adj4 (size or type or characteristic*)).kf,tw. (5085)

42 38 or 39 or 41 [nodule type or size] (54,250)

43 21 and 42 (1401)

44 35 and 43 (85)

45 **44 not 36 (77)**

46 **from 45 keep 23,26-27,36 (4)**

47 **37 or 46 (19)**

48 (distribution adj5 (size? or type? or characteristic? or solidity)).kf,tw. (66,593)

49 ((prevalence or proportion or percentage or distribution) adj5 (solid or nonsolid or partsolid or subsolid or ground glass or SSN or PSN or GGN or GGO or SSNs or PSNs or GGNs or GGOs)).kf,tw. (2138)

50 48 or 49 (68,596)

51 4 and 50 (792)

52 35 and 51 (41)

53 **52 not 45 (37)**

54 **from 53 keep 8 (1)**

Lines 25–35 of the MEDLINE search are the UK search filter described and validated in: Ayiku L, Levay P, Hudson T, Craven J, Barrett E, Finnegan A, *et al.* The MEDLINE UK filter: development and validation of a geographic search filter to retrieve research about the UK from OVID MEDLINE. *Health Inform Libr J* 2017;**34**:200–16. https://doi.org/10.1111/hir.12187

**Google (Chrome browser) 23 June 2022**

search terms: lung nodule prevalence UK*browsed first 20 results. 2 potentially relevant, one of which is the BTS guidelines:*

**Checked references related to prevalence in the BTS guideline (32):**

Callister ME, Baldwin DR, Akram AR, Barnard S, Cane P, Draffan J, *et al.* British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax* 2015;**70**(Suppl. 2):ii1–54. https://doi.org/10.1136/thoraxjnl-2015-207168

*8 potentially relevant papers*

# Appendix 4 Further details on key features of included studies

**TABLE 48** Study-level description of the 27 included studies for key question 1

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| ***AI-Rad Companion (Siemens Healthineers) (three studies)*** | | | | | | | |
| Abadia *et al*. 2021,[47] USA; retrospective test accuracy and MRMC study; VA10A prototype | Mixed population<br>Lung cancer screening, abnormal X-rays, suspicious nodule follow-up, abnormal lung-function tests, respiratory symptoms or history of lung diseases[b]<br>Selected 143 patients with least one lung condition[b] present and by nodule presence/absence in radiology report: (1) 103 with nodules, (2) 40 without nodules | Low dose, no contrast, 1 mm | Any type | [A] Stand-alone AI; one expert chest radiologist; [C] with concurrent AI (MRMC study); [D] without AI (MRMC study); [E] original radiology reports (one of five experienced chest radiologists without AI) | Per-nodule assessment/per-subject assessment<br>[1] [D] + AI-Rad (2nd-read AI)<br>[2] [E] + AI-Rad (2nd-read AI)[c]<br>AI-Rad vs. radiology reports:<br>[1] [E] + AI-Rad (2nd-read AI) | Nodule detection accuracy; nodule size measurement ([A] vs. [D]); characteristics of nodules (FN, FP); reading times; confidence in lung nodule detection | N/A |
| Chamberlin *et al*. 2021,[48] USA; retrospective test accuracy study; VA10A prototype | Screening population: randomly selected 117 patients from a single US institution | Low dose, no contrast, 1 mm | Any type, > 6 mm | [A] Stand-alone AI | Nodule detection: Consensus expert reading (two readers) | Nodule detection accuracy; characteristics of detected nodules | Quantification of coronary artery calcium volume; prediction of major cardiopulmonary outcomes; false-positive analysis |
| Rückel *et al*. 2021,[49] Germany; retrospective test accuracy study; prototype | Incidental population: 105 shock-room whole-body CT scans (consecutively included) from a single hospital | Standard dose, with contrast, 0.75 mm | Any type | [A] Stand-alone AI; [E] original radiologist report [single radiologist (18 images), or by a radiology resident and radiologist (87 images)]<br>25 different radiology residents and 18 different radiologists | Initial radiologist report plus additionally AI-identified and expert-confirmed nodules (2nd-read AI) | Accuracy to detect lung nodules; characteristics of detected nodules | N/A |

continued

**TABLE 48** Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/ comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| ***AVIEW LCS+ (Coreline Soft) (four studies)*** | | | | | | | |
| Hwang *et al*. 2021,[51] Republic of Korea; before-and-after study; A-view Lungscreen | Screening population: 6487 consecutive participants (1821 pre-AI implementation; 4666 post-AI implementation) from 14 institutions (K-LUCAS project) | Low dose, no contrast, < 1.5 mm | Solid, part-solid, ground glass | [A] Stand-alone AI for nodule detection; [B] assisted 2nd-read AI for nodule detection; [C] concurrent AI for nodule measurement and whole read including Lung-RADS categorisation | Lung nodules: Radiologist with second read AI [B]<br>Lung cancer: Medical record review | Characteristics of detected nodules; % detected nodules being malignant; nodule detection accuracy of [A]; accuracy to detect lung cancer (whole read [C] with Lung-RADS); number of people with positive screening result; technical failure rate | Nodule size measured on transverse planes vs. any maximum plane or maximum orthogonal plane |
| Hwang *et al*. 2021,[50] Republic of Korea; retrospective analyses of prospective cohort study; A-View Lungscreen | Screening population: 10,424 consecutive participants from the K-LUCAS project (14 institutions) | Low dose, no contrast, < 1.5 mm (1 mm, *n* = 9,514; 1.25 mm, *n* = 910) | Solid, part-solid, ground glass | [B] Second read AI for nodule detection; [C] concurrent AI for nodule measurement and whole read including Lung-RADS categorisation | Lung cancer: Medical record review | Accuracy to detect lung cancer; characteristics of detected nodules; % of nodules being malignant; number of people with positive screening result; technical failure rate | Agreement between average transverse and effective diameters and their diagnostic performance at various thresholds; proportional reduction of unnecessary follow-up CTs and frequency of delayed lung cancer diagnosis for each elevated threshold |
| Hwang *et al*. 2021,[52] Republic of Korea; prospective screening cohort and retrospective central reading; A-View Lungscreen | Screening population: 3,353 consecutive participants from the K-LUCAS project (14 institutions) | Low dose, no contrast, < 1.5 mm | Solid, part-solid, ground glass | [B] 2nd-read AI for nodule detection; [C] concurrent AI for nodule measurement and whole read including Lung-RADS categorisation | N/A | Characteristics of detected nodules; number of people having CT surveillance; number of people having excision/biopsy; technical failure rate | Positivity rates by Lung-RADS and NELSON criteria, segmentation failure/number of nodules per participant: Inter-radiologist variability; inter-institution variability; disagreement between the institutional reading and central review |
| Lancaster *et al*. 2022,[32] Russia; MRMC study; AVIEW LCS v1.0.34 | Screening population: enriched sample of 283 scans with at least one solid nodule | Ultra-low dose, no contrast, 1 mm | Solid | [A] Stand-alone AI for nodule detection and classification based on volume; [C] concurrent AI for nodule volume measurement (three experienced chest radiologists); [D] unaided reader: two experienced chest radiologist using other semiautomated volumetric software | Nodule categorisation: Consensus expert reading (three radiologists with > 10 years of experience and one experienced IT technologist) | Accuracy of nodule categorisation (< 100 mm$^3$, ≥ 100 mm$^3$); characteristics of detected nodules; simulated radiologist workload reduction when stand-alone AI software would be used as pre-screen to rule out negative CT images | N/A |

TABLE 48 Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| ***ClearRead CT (Riverain Technologies) (six studies)*** | | | | | | | |
| Singh *et al.* 2021,[56]<br>USA; MRMC study; ClearRead CT with vessel suppression and nodule detection | Screening population: enriched sample of 123 patients (100 with subsolid nodules and 23 with no nodules) from the NLST | Low dose, contrast use unclear, 1.2–2 mm | Part-solid, ground glass | [A] Stand-alone AI-AD (with vessel suppression and autodetection of pulmonary nodules);<br>[C.1] concurrent AI – two experienced radiologists reading AI vs. images (with vessel suppression without automatic nodule detection feature); [C.2] concurrent AI – two experienced radiologists reading AI-AD images (with vessel suppression and autodetection of pulmonary nodules);<br>[D] two experienced radiologists reading standard CT images | Nodule detection: Consensus expert reading (two readers) | Nodule detection accuracy; characteristics of detected nodules; size measurement accuracy; inter-observer agreement to detect the dominant nodule; technical failure rate; impact on clinical decision making (change in Lung-RADS category) | N/A |
| Lo *et al.* 2018,[54]<br>USA; MRMC study; ClearRead CT with vessel suppression and nodule detection; pre-market version (first-generation system) | Screening population: 324 enriched cases (including 95 cancers, 83 benign nodules; 216 nodule free vs. 108 cases with actionable nodules) from the NLST and two hospitals | Low dose, contrast and slice thickness unclear | Solid, part-solid, ground glass; 5–44 mm | [A] Stand-alone AI; 12 experienced general radiologists:<br>[C] With concurrent AI;<br>[D] without AI | Nodule detection: Consensus expert reading (three readers) assisted by corresponding NLST or source documentations containing radiologic, pathologic and follow-up reports | Accuracy of nodule detection; radiologist reading time | N/A |

continued

**TABLE 48** Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/ comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| Milanese *et al*. 2018, Switzerland,[55] MRMC study; ClearRead CT for vessel suppression; pre-market version (first generation system) | Unclear indication for CT: 93 consecutive patients referred to University Hospital Zurich for clinical non-enhanced chest CT | Low dose, no contrast, 2 mm | Solid; 13–366 mm$^3$ | [C] Nodule measurement on vessel suppressed CT images (one general radiologist with 3 years of experience, one resident radiologist) using semiautomatic segmentation software (MM Oncology, Siemens Healthcare)<br>[D] Nodule measurement on standard CT images (one general radiologist with 3 years of experience, one resident radiologist) using semiautomatic segmentation software (MM Oncology, Siemens Healthcare) | Nodule measurement: Volumes and longest diameters measured on standard CT images [D] by reader 1 and reader 2 for each nodule averaged | Measurement accuracy; inter-reader variability in nodule measurement; impact on clinical decision-making (categorisation according to Fleischner guidelines) | N/A |
| Hsu *et al*. 2021,[53] Taiwan; MRMC study; ClearRead CT with vessel suppression and nodule detection | Mixed population: 93 clinical routine; 57 screening population Outcomes for screening population reported separately<br>150 consecutive cases with lung nodules ≤ 1 cm or no nodules | Low dose (*n* = 57), standard dose (*n* = 93), no contrast, 2.5 mm | Any type; ≤ 10 mm | [A] Stand-alone AI; six chest radiologists – three less experienced and three experienced:<br>[B] With 2nd-read AI; [C] with concurrent AI; [D] without software | Nodule detection: consensus expert reading (two readers) | Nodule detection accuracy; radiologist reading time | N/A |
| Takaishi *et al*. 2021,[57] Japan; MRMC study; ClearRead CT for vessel suppression | Mixed population:[d] unclear how selected, 61 thoracic or thoracic-abdominal CT images conducted at one Japanese hospital in September 2019 | Standard dose, no contrast, 5 mm | Solid, ground glass; 4–54 mm diameter | Three general radiologists with 2–8 years of experience:<br>[C] With concurrent AI;<br>[D] without software | Nodule detection: consensus expert reading (two readers) | Nodule detection accuracy; radiologist reading time | N/A |

**TABLE 48** Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| Wan *et al*. 2020,[58] Taiwan; MRMC study; ClearRead CT with vessel suppression and nodule detection | Mixed population: selected only patients with previously identified nodules that had subsequent excision, 75 nodules in 50 cases[e] | Low dose, unclear contrast | Solid, part-solid, ground glass; ≤ 2 cm | [A] Stand-alone AI;<br>[D] consensus of two radiologists with 2–38 years of experience measuring diameter manually | Lung nodules and lung cancer: excision and pathological results | Nodule detection accuracy; lung cancer detection accuracy; characteristics of missed nodules; measurement concordance between stand-alone AI and unaided reader | |
| ***Contextflow SEARCH Lung CT (contextflow) (one study)*** | | | | | | | |
| Röhrich *et al*. 2023,[31] Austria, MRMC study, prototype version | Mixed population[f] (follow-up of a known lung disease, suspected lung disease, incidental): 100 with confirmed diffuse parenchymal lung disease, eight with inconspicuous chest CT scans from one hospital in Austria | Unclear dose, with or without contrast | Any type | Four radiology residents (2.1 ± 0.7 years of experience) and four general radiologists (12 ± 1.8 years of experience)<br>[C] With concurrent AI;<br>[D] without AI | Lung nodule detection: One experienced thoracic radiologist (20 years of experience) where available using prior and follow-up examinations, clinical symptoms, pathology and histology reports, and interdisciplinary board decisions | Radiologist reading time; technical failure rate | Overall diagnostic accuracy for diffuse parenchymal lung disease |
| ***InferRead CT Lung (Infervision) (three studies)*** | | | | | | | |
| Kozuka *et al*. 2020,[59] Japan; MRMC study; version NR | Symptomatic population (suspected cancer): Random 120 chest CT images from one hospital in Japan | Standard dose; no contrast; 1 mm | Solid, part-solid, calcified, ground glass | [A] Stand-alone AI;<br>two less experienced radiologists:<br>[C] With concurrent AI;<br>[D] without AI | Nodule detection: consensus expert reading (three readers) | Nodule detection accuracy; radiologist reading time; characteristics of detected nodules | N/A |
| Liu *et al*. 2019,[60] China; MRMC study; software name and version NR | Mixed population: screening and inpatient, convenience sample, 1129 CT scans from > 10 hospitals in China<br>Evaluation 1: *N* = 1,129;<br>Evaluation 4: *N* = 123 (batch 1); *N* = 148 (batch 2) | Standard dose or low dose; unclear regarding contrast; 0.8–2.0 mm | Solid, subsolid, calcified, pleural | Evaluation 1:<br>[A] Stand-alone AI; [D.1] two experienced general radiologists without AI<br>Evaluation 4:<br>Two experienced general radiologists: [C] with concurrent AI; [D.2] without AI | Nodule detection: consensus expert reading (three readers) | Nodule detection accuracy; comparison of AI performance by radiation dose; radiologist reading time | AI performance by patient age (evaluation 2) and CT manufacturer (evaluation 3) |

continued

**TABLE 48** Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| Zhang *et al.* 2021,[61] China; retrospective test accuracy and MRMC study; software version NR | Screening population: 860 consecutive patients from one hospital in China (part of NELCIN-B3 project) | Low dose; no contrast; 0.625–1.0 mm | Solid, part-solid, ground glass | One radiology resident with supervision of one experienced radiologist:<br>[C] With concurrent AI (MRMC study: one radiology resident and one experienced radiologist)<br>[E] Without AI (clinical practice: 14 different radiology residents and 15 different experienced radiologists) | Nodule detection: consensus expert reading (two readers) | Nodule detection accuracy; characteristics of detected nodules | N/A |
| ***JLD-01K (JLK, Inc.)*** | | | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | | | |
| ***Lung AI (Arterys)*** | | | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | | | |
| ***Lung Nodule AI (Fujifilm)*** | | | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | | | |
| ***qCT-Lung (Qure.ai)*** | | | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | | | |
| ***SenseCare-Lung Pro (SenseTime)*** | | | | | | | |
| No relevant evidence was identified by the EAG or supplied by the company | | | | | | | |
| ***Veolity (MeVis) (four studies)*** | | | | | | | |
| Cohen *et al.* 2017,[62] Republic of Korea, MRMC study, version 1.1 | Surveillance population with applicability concerns: 73 patients with preoperative CT scan for subsolid nodules and subsequent surgical resection at one Korean hospital | Standard dose; no contrast; 0.625 mm | Subsolid nodules | Two radiologists with 4–5 years of experience:<br>[C] Concurrent AI, assessing CT images reconstructed using FBP and MBIR algorithms, respectively | No reference standard | Diameter and volume measurement: technical failure rate; inter-observer variability; repeatability/reproducibility; concordance between readers with software: FBP vs. MBIR | N/A |

**TABLE 48** Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/ comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| Kim *et al.* 2018,[63] Republic of Korea, MRMC study, version 1.2 | Surveillance population with applicability concerns:<br>89 consecutive patients with preoperative CT scan for subsolid nodules and subsequent surgical resection at one Korean hospital | Standard dose; no contrast; 0.625 mm | Subsolid nodules | Two experienced radiologists:<br>[C] With concurrent AI;<br>[D] without AI | No reference standard for nodule size measurement | Diameter measurement: concordance between readers with and without software; inter-observer variability; repeatability/ reproducibility; technical failure rate | Diagnostic performance using binary logistic regression analysis for invasive adeno-carcinoma |
| | | | | | | Nodule classification by size of solid portion: inter-observer variability; repeatability/ reproducibility | |
| Hall *et al.* 2022,[27] UK, retrospective test accuracy study and MRMC study, version 1.2 | Screening population: All 770 available CT scans from LSUT | Low dose; no contrast; 0.5–1.0 mm | Solid, part-solid, ground glass; ≥ 5 mm or ≥ 80 $mm^3$ | [C] Concurrent AI: Two radiographers without prior experience in chest CT (MRMC study)<br>[E] Without AI: one of five original study chest radiologists with 5–28 years of experience (clinical practice); 95% single reading, 5% double reading | Nodule detection: Nodules identified by study radiologists without AI [D], plus review of any additional nodules identified by the radiographers with concurrent AI [C] by one radiologist (if needed two) for consensus | Nodule detection accuracy; lung cancer detection accuracy; impact on decision-making; radiologist reading time; software acceptability and experience; proportion of scans referred for CT surveillance; proportion of scans referred to MDT; characteristics of missed nodules; % of detected nodules being malignant | N/A |
| Jacobs *et al.* 2021,[64] USA, Denmark, the Netherlands; MRMC study, version 1.5 | Screening population: Selected 160 patients (80 round 1 and 80 round 2) from NLST: 40 Lung-RADS 1 or 2; 40 Lung-RADS 3; 40 Lung-RADS 4A; 40 Lung-RADS 4B | Low dose; no contrast; 1.0–3.2 mm | Any nodules | Three experienced radiologists and four radiology residents from Denmark and the Netherlands:<br>[C] With concurrent AI;<br>[D] without AI | No reference standard | Lung-RADS categorisation: inter-observer variability; repeatability/ reproducibility<br>Radiologist reading time; technical failure rate; impact on decision-making | N/A |

continued

**TABLE 48** Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/ comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| ***Veye Lung Nodules (Aidence) (five studies )*** | | | | | | | |
| Blazis *et al*. 2021,[65] the Netherlands, retrospective test accuracy study; Veye Chest, version NR | Mixed indication (ranging from pulmonary nodule follow-up to primary staging of abdominal malignancy): sampling method unclear, 31 patients (384 CT reconstructions from 24 patients included in analyses) from one Dutch hospital | Unclear dose, unclear contrast use, 1 mm and 3 mm | Any nodules; > 4 mm or > 30 mm$^3$ | [A] Stand-alone AI | Nodule detection: consensus expert reading (three readers) | Nodule detection accuracy | N/A |
| Hempel *et al*. 2022,[34] the Netherlands; MRMC study; Veye Chest v2.15.3 | Mixed indication: 50 patients with incidentally detected nodules or no nodules from one Dutch hospital: 5 no nodules, 45 with ≤ 5 nodules (10 no prior CT, 35 with prior CT)<br>Incidental population (*n* = 15); surveillance population (*n* = 35) | Unclear dose; with or without contrast; 2.00 mm (*n* = 73), 3.0 mm (*n* = 12) | Actionable nodules: 65–14,000 mm$^3$ or 5–30 mm | One experienced chest radiologist and one experienced general radiologist:<br>[C] With concurrent AI;<br>[D] without AI | Risk categorisation based on 2015 BTS grades:<br>All cases with discrepant BTS grades between readers re-evaluated during a consensus meeting and a consensus BTS grade determined | BTS grade category: Accuracy; characteristics of detected nodules; radiologist reading time; technical failure rate; inter-observer variability | N/A |
| Martins Jarnalo *et al*. 2021,[66] the Netherlands, retrospective test accuracy study, Veye Chest; versions (25 May 2018), and (18 March 2019) | Mixed indications (ruling out metastasis, follow-up of nodules or other pulmonary abnormalities, other miscellaneous indications):<br>145 randomly selected CT images performed at one Dutch teaching hospital | Unclear dose; with or without contrast; 1 mm or 3 mm | Solid, subsolid; 4–30 mm | [A] Stand-alone AI | Nodule detection, composition and measurement: consensus expert reading (three readers) | Nodule detection accuracy; nodule type accuracy (solid, subsolid); size measurement accuracy; characteristics of detected (TP, FP) and missed (FN) nodules; technical failure rate; software acceptability and experience | N/A |

TABLE 48 Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/ comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| Murchison *et al.* 2022,[33] UK, MRMC study, Veye Chest version 2.0 | Mixed indications (clinical routine mimicking a screening population in age and smoking history[h]):<br>337 CT scans of 314 subjects from one hospital in Edinburgh<br>[1] No nodules in original report (*n* = 178); [2] with 1–10 nodule in original report (*n* = 95); [3] 23 baseline scans that were followed up for presence of a lung nodules; [4] 23 follow-up CT scans of [3]; [5] with subsolid nodules in original report (*n* = 18) | Standard dose, with (*n* = 22) or without contrast (*n* = 315);<br>1.0–2.5 mm | Any type; 3–30 mm, 5–30 mm | [A] Stand-alone AI<br>Two experienced chest radiologists:<br>[C] With concurrent AI;<br>[D] without AI | Nodule detection and composition:<br>Majority expert reading (two index test readers with discrepancies adjudicated by a third experienced chest radiologist)<br>Nodule measurement and growth rate:<br>No consensus requirement for the reference standard of segmentation. All segmentations were retained | Nodule detection accuracy; nodule type accuracy; measurement (volume, diameter): inter-observer variability; concordance between stand-alone software and readers without software<br>Technical failure rate<br>Growth rate:<br>Nodule registration accuracy; inter-observer variability; concordance between stand-alone software and readers without software | N/A |

continued

TABLE 48 Study-level description of the 27 included studies for key question 1 (*continued*)

| Study, country, design and software version[a] | Study population | CT acquisition details | Type and size of nodules | Index test(s) ([A], [B], [C])/comparator ([D], [E]) | Reference standard | Relevant outcomes reported | Other outcomes (not reported in this report) |
|---|---|---|---|---|---|---|---|
| ***VUNO Med-Lung CT AI (VUNO) (one study)*** | | | | | | | |
| Park *et al*. 2022,[67] USA, Republic of Korea, MRMC study, v.1.0.1 | Screening population: 200 cases randomly selected from a nodule- and cancer-enriched subset of the NLST database | Low dose, no contrast | Solid, part-solid, non-solid | [A] Stand-alone AI;<br>1 resident radiologist and 4 radiologists with 1–20 years of experience:<br>[C] With concurrent AI;<br>[D] without AI | Lung cancer detection: NR (same-year positive cancer diagnosis) | Nodule detection and Lung-RADS categorisation:<br>Lung cancer detection accuracy; concordance between stand-alone software and readers; inter-observer variability; impact on decision-making | Assignment of risk-dominant nodules |

[A] Stand-alone AI; [B] reader with 2nd-read AI; [C] reader with concurrent AI; [D] unaided reader; [E] original radiologist report. AI, artificial intelligence; BTS, British Thoracic Society; CT, computed tomography; FN, false negative; FP, false positive; K-LUCAS, Korean Lung Cancer Screening project; LIDC-IDRI, Lung Image Database Consortium image collection; LSUT, Lung Screen Uptake Trial; Lung-RADS, Lung CT Screening Reporting And Data System; MDT, multi-disciplinary team; MRMC, multi-reader multi-case study; N/A, not applicable; NELCIN-B3, Netherlands-China Big-3 disease screening: lung cancer, coronary atherosclerosis, and chronic obstructive pulmonary disease; NELSON, Dutch-Belgian Randomized Lung Cancer Screening Trial; NLST, National Lung Screening Trial; NR, not reported; TP, true positive.

a Where the software evaluated in the study had a different name from that listed in the NICE final scope, but the company confirmed its relevance.

b Interstitial lung disease, chronic obstructive lung disease, respiratory bronchiolitis, pulmonary oedema or pulmonary embolism.

c If AI-Rad found additional nodules, the expert radiologist verified if the nodules were TP or FP.

d Postoperative follow-up (*n* = 14), to identify the cause of fever (*n* = 11), to identify the cause of abdominal pain (*n* = 9), scrutiny of abnormality in chest X-ray (*n* = 7), annual medical check-up (*n* = 4), cancer staging (prostate, colon, etc.) (*n* = 3), trauma survey (*n* = 2), other (*n* = 11).

e For 561 patients screened for eligibility: LDCT health examination at one's own expense (*n* = 207), malignant neoplasms of other organs (*n* = 127), chief complaints other than respiratory symptoms (*n* = 103), symptoms or signs of respiratory diseases (*n* = 68), follow-up CT of lung cancer after treatment (*n* = 56). Inclusion criteria state that the CT scan must have been low dose, and patients with a previous history of thoracic surgery and/or a final pathological diagnosis with metastases were excluded.

f Most of the indications for the 108 CT scans were either follow-up examination in case of an already known disease or the primary CT-scan in case of a clinically suspected disease. In some cases, the CT findings were incidental, and the scan was conducted for another reason not covered by the exclusion criteria.

g One study considered confidential was removed from the table.

h Current smokers, people with a smoking history and/or people with radiological evidence of pulmonary emphysema.

# **Appendix 5** Descriptions of evidence from individual studies providing data on nodule detection and impact on patient management

## Descriptions of evidence from individual studies providing data on nodule detection (summarised in *Nodule detection*)

### *Accuracy for identifying any nodules*

#### Concurrent AI versus unassisted reader (four studies)

- **Screening population (two studies)**

*Hsu et al. 2021,*[53] *Taiwan: ClearRead CT (Riverain Technologies)*

Hsu *et al.*'s study[53] comprised 150 consecutive cases with lung nodules ≤ 1 cm or no nodules [93 standard-dose CT images from clinical routine and 57 LDCT images from lung cancer screening]. Six readers participated in the MRMC study: three residents in radiology (junior group) and three experienced chest radiologists (senior group). Accuracy results were reported separately for the 57 LDCT images obtained for screening purposes. The mean per-nodule sensitivity of all six readers increased significantly from 63% (95% CI 59% to 66%) without software use to 79% (95% CI 76% to 81%) with software use ($p < 0.001$). The mean per-person specificity did not change significantly: 81% (95% CI 78% to 84%) with software use and 77% (95% CI 74% to 80%) for unaided readers ($p = 0.449$).

*Zhang et al. 2021,*[61] *China: InferRead CT Lung (Infervision)*

Zhang *et al.*[61] included 860 consecutive patients who underwent chest CT from November to December 2019 at one Chinese hospital as part of the Netherlands-China Big-3 disease screening (NELCIN-B3) project. One resident drafted the diagnostic report, and a board-certified radiologist supervised the final version without software use in clinical practice or with concurrent software use under laboratory conditions. The per-subject sensitivity for detecting any nodules was 98.9% (370/374) with versus 43.3% (162/374) without software use. No level of significance was reported for all nodule types combined, but the sensitivities for the detection of solid, part-solid and ground-glass nodules, respectively, were all significantly higher with AI software use ($p < 0.001$ for all). The per-subject specificity was 97.1% (472/486) with versus 100.0% (486/486) without software use (no level of significance reported).

- **Symptomatic population (one study)**

*Kozuka et al. 2020,*[59] *Japan: InferRead CT Lung (Infervision)*

Kozuka *et al.*[59] reported per-nodule and per-patient accuracy for concurrent AI and unaided readers by nodule type and size. This study was a retrospective analysis of 120 randomly selected chest CT images (117 cases included in analysis) from cases with suspected lung cancer. Two less experienced radiologists (1 and 5 years of diagnostic experience) assessed the CT images with and without software use. The per-nodule sensitivity for the pooled readers increased significantly from 20.9% (95% 18.8% to 23.0%) for the unaided reader to 38.0% (95% CI 35.5% to 40.5%) with concurrent AI ($p < 0.01$). The pooled positive predictive value was 61.8% (95% CI 58.6% to 65.0%) with and 70.5% (95% CI 66.0% to 74.7%) without software. The pooled per-patient sensitivity increased significantly with software use from 68.0% (95% CI 61.4% to 74.1%) to 85.1% (95% CI 79.8% to 89.5%) ($p < 0.001$). The pooled specificity decreased from 91.7% (11/12; 95% CI 61.5% to 99.8%) to 83.3% (10/12; 95% CI 51.6% to 97.9%) with concurrent software use.

- **Mixed population (two studies)**

*Hsu et al. 2021,[53] Taiwan: ClearRead CT (Riverain Technologies)*

Hsu *et al.*'s study[53] comprised 150 consecutive cases with lung nodules ≤ 1 cm or no nodules (93 standard-dose CT images from clinical routine and 57 LDCTs from lung cancer screening). Six readers participated in the MRMC study: three residents in radiology (junior group) and three experienced chest radiologists (senior group). For all readers, the mean per-nodule sensitivity was significantly improved with software use: 80% (95% CI 79% to 82%) versus 64% (95% CI 62% to 66%) without software use ($p$ < 0.001). The mean specificity was 83% (95% CI 82% to 85%) with software use and 80% (95% CI 78% to 81%) without software use ($p$ = 0.25).

In the junior group, the mean per-nodule sensitivity increased significantly from 52% (95% CI 49% to 55%) without software use to 77% (95% CI 74% to 79%) with software use ($p$ < 0.001). The mean specificity was 78% (95% CI 76% to 81%) with and 71% (95% CI 69% to 74%) without software use ($p$ = 0.152). In the senior group, the mean per-nodule sensitivity was significantly higher with software use: 84% (95% CI 82% to 86%) compared with 74% (95% CI 72% to 77%) without software use ($p$ < 0.001). The mean specificity was 88% (95% CI 87% to 90%) with and 87% (95% CI 85% to 89%) without software use ($p$ = 0.729).

*Takaishi et al. 2021,[57] Japan: ClearRead CT (Riverain Technologies)*

Takaishi *et al.*[57] performed a retrospective analysis of 61 thoracic or thoracic-abdominal unenhanced CT images produced at Konan Kosei hospital during September 2019. The MRMC study assessed the nodule detection accuracy of three radiologists (8, 6 and 2 years' experience, respectively) with and without software support. The study found significantly higher average per-nodule sensitivities with software use: 84.1% (116/138) compared with 71.7% (99/138) without software use ($p$ = 0.02). The average false-positive rate was 21% for both concurrent AI (0.49 false positive per scan) and unassisted reading (0.44 false positive per scan) ($p$ = 0.98).

## Assisted second-read AI versus unassisted reader (one study)

- **Screening population (one study)**

*Hsu et al. 2021,[53] Taiwan: ClearRead CT (Riverain Technologies)*

Hsu *et al.*'s study[53] comprised 150 consecutive cases with lung nodules ≤ 1 cm or no nodules (93 standard-dose CT images from clinical routine and 57 LDCTs from lung cancer screening). Six readers participated in the MRMC study: three residents in radiology (junior group) and three experienced chest radiologists (senior group). They first read the CT images unaided and then used the reading performed by the software to make a final decision (assisted 2nd-read mode). Accuracy results were reported separately for the 57 LDCTs obtained for screening purposes. For all readers, the mean per-nodule sensitivity was significantly higher with software use: 80% (95% CI 77% to 83%) compared with 63% (95% CI 59% to 66%) without software use ($p$ < 0.001). The mean specificity was 82% (95% CI 79% to 84%) with 2nd-read AI and 77% (95% CI 74% to 80%) without software ($p$ = 0.360).

In the junior group, the mean per-nodule sensitivity increased significantly from 52% (95% CI 47% to 57%) without software support to 76% (95% CI 72% to 80%) with 2nd-read AI use ($p$ < 0.001). The mean specificity was 76% (95% CI 72% to 80%) with and 68% (95% CI 64% to 73%) without software support ($p$ = 0.333). For the senior group, the mean per-nodule sensitivity improved from 73% (95% CI 69% to 77%) without software support to 84% (95% CI 80% to 87%) with 2nd-read software use ($p$ = 0.001). The mean specificity was 88% (95% CI 85% to 91%) with versus 86% (95% CI 83% to 90%) without software support ($p$ = 0.795).

- **Mixed population (one study)**

*Hsu et al. 2021,[53] Taiwan: ClearRead CT (Riverain Technologies)*

Hsu *et al.*'s reader study[53] retrospectively analysed data from consecutive cases with lung nodules ≤ 1 cm or no nodules (93 standard dose CT images from clinical routine and 57 LDCTs from lung cancer screening) from a hospital in Taiwan. In assisted 2nd-read AI mode, the six readers first read the CT images without AI and then combined the displays of the AI results to make the final decision. The mean per-nodule sensitivity for all six readers was increased from 64% (95% CI 62% to 66%) without software use to 82% (95% CI 80% to 84%) with 2nd-read AI ($p$ < 0.001). The mean specificity was 84% (95% CI 82% to 85%) using 2nd-read AI compared with 80% (95% CI 78% to 81%) with unaided reading ($p$ = 0.177).

For the three junior readers, the mean per-nodule sensitivity was 79% (95% CI 76% to 81%) with and 52% (95% CI 49% to 55%) without software use ($p$ < 0.001). Their mean specificity was 79% (95% CI 77% to 82%) with and 71% (95% CI 69% to 74%) without software use ($p$ = 0.088). For the three senior readers, the mean per-nodule sensitivity was 85% (95% CI 83% to 87%) with and 74% (95% CI 72% to 77%) without software use ($p$ < 0.001). Their mean specificity was 88% (95% CI 87% to 90%) with and 87% (95% CI 85% to 89%) without software use ($p$ = 0.729).

### *Accuracy for detecting actionable nodules*

### Concurrent AI versus unassisted reader (five studies)

- **Screening population (three studies)**

*Singh et al. 2021,[56] USA: ClearRead CT (Riverain Technologies)*

Singh *et al.*[56] selected 150 LDCT from the US-based NLST: the first 125 patients with mixed attenuation or ground-glass nodules and the first 25 patients with no nodules. Two radiologists (with 5 and 10 years of thoracic CT experience) participated in a MRMC study to detect nodules ≥ 6 mm on vessel-suppressed CT images as well as on standard CT images. The evaluated software did not have a nodule detection function. For ground-glass nodules, the pooled per-nodule sensitivity was 67% (209/312) on vessel-suppressed CT images and 66% (207/312) on standard CT images. The average specificity was 78.5% on vessel-suppressed images and 84% on standard CT images. For part-solid nodules, the pooled per-nodule sensitivity was 80% (245/308) versus 70% (216/308), and the average specificity was 85% versus 76% in vessel-suppressed versus standard CT images, respectively. For all subsolid nodules, the pooled per-nodule sensitivity was 73% (453/620) versus 68% (423/620), and the mean specificity was 74% versus 78% on vessel-suppressed versus standard CT images.

*Lo et al. 2018,[54] USA: ClearRead CT (Riverain Technologies)*

Lo *et al.*'s study[54] included 324 LDCTs from the US-based NLST and two US hospitals; images with nodules (5–44 mm) and without nodules were selected in a ratio of 2 : 1. Twelve general radiologists certified by the American Board of Radiology (with 6–26 years of experience) participated in a MRMC study. Concurrent software use increased the mean per-nodule sensitivity by 12.4% (95% CI 6.2% to 18.6%) from 60.1 ± 3.3% to 72.5 ± 3.3% ($p$ < 0.001) and decreased the mean specificity by 5.5% (95% CI −9.0% to −1.9%) from 89.9 ± 2.0% to 84.4 ± 2.0% ($p$ = 0.0025). The average false-positive rate increased slightly from 0.17 false-positive nodules/scan to 0.28 false-positive nodules/scan ($p$ < 0.01) with software use.

*Hall et al. 2022,[27] UK: Veolity (MeVis)*

Hall *et al.*'s study[27] included all 770 LDCTs from the London-based LSUT study. In a MRMC study, two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCT with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm). The comparator were the experienced study radiologists (5–28 years of experience; 95% of scans read by single readers and 5% by double reading) who had read the CT images in clinical practice without software use. The reference standard comprised all nodules identified by study radiologists without software, plus consensus radiologist confirmed additional nodules identified by the software-assisted radiographers. At the 5-mm threshold, the per-subject sensitivity was 68.0% (102/150) and 73.7% (115/156) for AI-assisted radiographer 1 and 2, respectively. Specificity was 92.1% (490/532) and

92.7% (510/550) for reader 1 and 2, respectively. The average false-positive rate was 7.9% (42/532) and 7.3% (40/550) for reader 1 and 2, respectively, using concurrent AI. The sensitivity was 91.1% (144/158) for the unaided experienced radiologists, and the specificity for unaided reading was by definition of the reference standard 100%. However, 19 scans were excluded from the reference standard that were recalled by the original radiologists but contained nodules below the size threshold in the BTS guidelines for warranting surveillance.[12] Therefore, the specificity of the unaided radiologists in identifying people without actionable nodules was 96.7% (558/577).

- **Symptomatic population (one study)**

*Kozuka et al. 2020,[59] Japan: InferRead CT Lung (Infervision)*

Kozuka *et al.*[59] randomly selected 120 chest CT images (117 cases included in analysis) from cases with suspected lung cancer. They performed a MRMC study with two less experienced radiologists (1 and 5 years of experience). The pooled per-nodule sensitivity for the detection of nodules ≥ 6 mm was 51.9% (219/422) with versus 38.9% (164/422) without software support (calculated by reviewers; no level of significance reported).

- **Mixed population (one study)**

*Murchison et al. 2022,[33] UK: Veye Lung Nodules (Aidence)*

Murchison *et al.*'s study[33] used CT studies from a routine clinical population in a single academic hospital (Royal Infirmary of Edinburgh, Edinburgh, UK), between January 2008 and December 2009. Two thoracic radiologists (≥ 9 years' experience) participated in a MRMC study. Two data sets were created from the 337 CT scans: one set with AI results and one set without AI results. Reader 1 reviewed all the CT scans, but half of the CT scans with the AI results and the other half without AI results. For reader 2 this was vice versa. Hence, each CT scan was reviewed twice, once by one reader with the AI results and once by the other reader without the use of AI. The sensitivity for detecting actionable nodules (5–30 mm) was 80.3% (95% CI 75.2% to 85.0%) with and 71.7% (95% CI 66.0% to 77.0%) without software use ($p < 0.01$), with an average number of false-positive detections per image of 0.16 and 0.11, respectively.

### Assisted second-read AI versus unassisted reader (no study)

No data available.

### *Accuracy for detecting malignant nodules*

Three comparative studies[54,57,67] evaluated the *accuracy for detecting malignant nodules*. Of these, two included a screening population[54,67] and one included a mixed population.[57]

### Concurrent AI versus unassisted reader (three studies)

- **Screening population (two studies)**

*Lo et al. 2018,[54] USA: ClearRead CT (Riverain Technologies)*

The study by Lo *et al.*[54] included 324 LDCTs (including 95 lung cancer cases) from the US-based NLST and two US hospitals; images with nodules (5–44 mm) and without nodules were selected in a ratio of 2 : 1. Twelve general radiologists certified by the American Board of Radiology (with 6–26 years of experience) participated in a MRMC study. The study found 15.4% (95% CI 8.2% to 22.5%; $p = 2.50 \times 10^{-5}$) higher sensitivity (80.0 ± 3.9% vs. 64.7 ± 3.9%) and −5.5% (95% CI −9.0% to −1.9%; $p = 0.0025$) lower specificity (84.4 ± 2.0% vs. 89.9 ± 2.0%) in the detection of malignant nodules with concurrent AI compared with unaided reading. The number of false detections per image increased from 0.22 with unaided reading to 0.39 with concurrent AI use ($p < 0.01$).

*Park et al. 2022,[67] USA, Republic of Korea: VUNO Med-LungCT AI (VUNO)*

Park *et al.*[67] included a nodule- and cancer-enriched screening population (200 baseline LDCT; 31 cancer cases) selected from the US-based NLST data set. Five readers participated in the MRMC study. They consisted of one fourth-year radiology resident and four board-certified radiologists with 1, 4, 8 and 20 years of experience in chest radiology from the Asan Medical Center in Seoul (Republic of Korea). The pooled sensitivity to detect malignant nodules was 91.6% (95% CI 81.7% to 96.4%) with and 85.2% (95% CI 74.2% to 92.0%) without software use ($p$ = 0.004).

- **Mixed population (one study)**

*Takaishi et al. 2021,[57] Japan: ClearRead CT (Riverain Technologies)*

Takaishi *et al.*[57] performed a retrospective analysis of 61 thoracic or thoracic-abdominal unenhanced CT images (including one cancer case) produced at Konan Kosei hospital during September 2019. The MRMC study assessed the nodule detection accuracy of three radiologists (8, 6 and 2 years' experience, respectively) with and without software support. The sensitivity for detecting malignant nodules was 100% (1/1) for both AI-assisted and unassisted readers. The positive predictive value was 2.4% (1/42) without and 2.0% (1/49) with software use (average of 3 readers) (no level of significance reported).

### Assisted second-read AI versus unassisted reader (no study)

No data available.

## Descriptions of evidence from individual studies providing data on characteristics of detected nodules (summarised in *Characteristics of detected nodules*)

### *All detected nodules (true positive and false positive) (six studies)*

a. Comparative results: reader with and without software (two studies)

*Mixed population: Veye Chest (Aidence) (one study)*

Hempel *et al.* selected 50 chest CT scans with incidentally detected nodules (35 with and 10 without prior imaging) or no nodules ($n$ = 5) from one hospital in the Netherlands.[34] For this MRMC study, two experienced radiologists independently assessed the CT images twice to determine nodule management recommendation grade based on the 2015 BTS guidelines:[12] first unaided and then aided by Veye Chest software (Aidence). The readers were tasked with reporting the relevant pulmonary nodules that contributed to their management decision. A summary of the nodule types and sizes is reported in *Table 49*. Both radiologists reported fewer actionable nodules with concurrent software use, most likely because the software provided the radiologist with a list of nodules, and therefore there was no need to personally keep track of all findings. With software use, the proportion of detected nodules that were solid was lower (87.1%) than without software use (90.6%) (no level of significance reported).

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

In a before-and-after study, Hwang *et al.* included 6487 consecutive participants of the K-LUCAS project: 1821 participants were screened before the AVIEW Lungscreen software was implemented and 4666 participants were screened after.[51] The study observed a significantly larger number of detected nodules per participant (0.76 vs. 1.07; $p$ < 0.001) and a higher proportion of solid nodules (90.2% vs. 93.9%; $p$ < 0.001) in participants screened after software implementation (*Table 50*). No significant difference in nodule size was observed when nodules were measured on transverse planes after software implementation ($p$ = 0.441), but nodules were significantly larger when they were measured on any maximum plane ($p$ < 0.001) or maximum orthogonal plane ($p$ = 0.021). The significance of these findings needs to be treated with caution, however, as the study did not use a fully paired design, but different CT images were analysed by different readers before and after software implementation.

b. Comparative results: stand-alone AI versus unaided reader (one study)

**TABLE 49** Nodule number, type and size in patients with incidentally detected nodules on CT, with and without concurrent use of Veye Chest[34]

| | Unaided | | Aided | |
|---|---|---|---|---|
| | Reader 1 | Reader 2 | Reader 1 | Reader 2 |
| Number of nodules reported (*n*) | 64 | 63 | 41 | 44 |
| Patients with nodules, *n* (%) | 41/50 (82.0%) | 44/50 (88.0%) | 41/50 (82.0%) | 40/50 (80.0%) |
| ***Nodule type, n (%)*** | | | | |
| Solid | 58/64 (90.1%) | 57/63 (90.5%) | 36/41 (87.8%) | 38/44 (86.4%) |
| Part-solid | 5/64 (7.8%) | 4/63 (6.3%) | 4/41 (9.8%) | 4/44 (9.1%) |
| GGO | 1/64 (1.6%) | 2/63 (3.2%) | 1/41 (2.4%) | 2/44 (4.5%) |
| ***Nodule size (mean ± SD)*** | | | | |
| Volume (mm$^3$) | 567.2 ± 626.8 (*n* = 29) | 613.9 ± 791.3 (*n* = 35) | 736.3 ± 835.0 (*n* = 40) | 632.0 ± 720.0 (*n* = 42) |
| Diameter (mm) | 10.8 ± 5.7 (*n* = 35) | 10.0 ± 3.5 (*n* = 28) | 27.0 ±NA (*n* = 1) | 17.8 ± 8.6 (*n* = 2) |

GGO, ground-glass opacities.

One study reported the size of nodules detected by stand-alone AI as well as by an expert unaided reader in a mixed population.[47]

*Mixed population: AI-Rad Companion Chest CT (Siemens Healthineers) (one study)*

Abadia *et al.* included 103 patients with at least one lung condition and one suspicious lung nodule (≥ 4 mm) on radiology report and 40 patients with one lung condition and no lung nodule on radiology report from random LDCT images taken at a single US hospital for various reasons.[47] The nodule 2D axial size of all 312 nodules detected by stand-alone software (AI-Rad Companion CT Chest prototype) and of all 366 nodules detected by an unaided expert chest radiologist are reported in *Table 51*.

Three studies reported the characteristics of nodules detected by software-assisted readers[50,52] and stand-alone software,[66] respectively, without a comparator. These non-comparative results are reported in *Appendix 6*.

### *True-positive nodules (seven studies)*

a. Comparative results: reader with and without software (two studies)

Two studies compared the characteristics of true-positive nodules in readers assessing the same CT images with and without software use (InferRead CT lung, Infervision).[59,61]

*Symptomatic population: InferRead CT Lung (Infervision) (one study)*

Kozuka *et al.*[59] randomly selected 120 chest CT images from cases of suspected lung cancer at a single hospital in Japan. In a MRMC study, two less experienced radiologists assessed the CT images first without software for nodule detection and then with software (InferRead CT Lung, Infervision). The distribution of size and type of the 743 nodules ≥ 3 mm that were detected by the reference standard (majority reading of three experienced radiologists) as well as nodule type and size of correctly detected lung nodules of readers with and without software support are reported in *Table 52*. An additional 254 true-positive nodules were identified by the two readers with software use. The additional nodules had

TABLE 50 Characteristics of detected nodules (true and false positives) in consecutive screening populations from the Republic of Korea (three studies)

| Reference and country | Technology/reading details for detection | Number of nodules or participants | Nodule type | Nodule size | Lung-RADS category | |
|---|---|---|---|---|---|---|
| Hwang *et al.* 2021,[51] Republic of Korea | Unaided reader | 1,391 nodules | Solid 90.2%; part-solid 3.7%; pure GGN 6.0% | Transverse plane (all nodules); mean 4.5 mm, SD 3.8 mm | Per-nodule: Transverse plane<br>2 – 84.8%<br>3 – 9.1%<br>4A – 3.2% | 4B – 1.2%<br>4X – 1.7% |
| | AVIEW Lungscreen (Coreline Soft); 2nd-read mode | 4,990 nodules | Solid 93.9%; part-solid 1.6%; pure GGN 4.5% | Transverse plane (all nodules); mean 4.4 mm, SD 3.5 mm | 2 – 89.2%<br>3 – 6.5%<br>4A – 2.5% | 4B – 1.1%<br>4X – 0.7% |
| | Unaided reader | 1,821 participants | NR | NR | Per-participant: Transverse plane<br>1 – 58.6%<br>2 – 31.5%<br>3 – 5.3% | 4A – 2.3%<br>4B – 0.7%<br>4X – 1.5% |
| | AVIEW Lungscreen (Coreline Soft); 2nd-read mode | 4,660 participants | NR | NR | 1 – 51.5%<br>2 – 37.6%<br>3 – 6.1% | 4A – 2.9%<br>4B – 1.2%<br>4X – 0.8% |
| Hwang *et al.* 2021,[50] Republic of Korea | AVIEW Lungscreen (Coreline Soft); 2nd-read mode | 10,080 nodules | Solid 93.9%; part-solid 1.6%; pure GGN 4.5% | Average transverse diameter<br>Solid: median 3.6 mm; < 5 mm: 75.1%; 5–6 mm: 8.1%; 6–8 mm: 6.0%; ≥ 8 mm: 4.6%<br>Part-solid: median 11.9 mm; < 5 mm: 0.008%; ≥ 5 mm: 1.5%<br>Pure GGN: median 5.8 mm; < 5 mm: 1.7%; ≥ 5 mm: 2.8% | NR | |
| | AVIEW Lungscreen (Coreline Soft); 2nd-read mode | 4,642 risk-dominant nodules | Average transverse diameter: solid < 6 mm: 76.9%; solid 6–7 mm: 6.5%; solid 7–8 mm: 2.8%; solid 8–9 mm: 1.8% | Solid 9–10 mm: 1.1%<br>Solid ≥ 10 mm: 4.7%<br>Part-solid: 2.7%<br>Pure GGN: 3.4% | NR | |

continued

**TABLE 50** Characteristics of detected nodules (true and false positives) in consecutive screening populations from the Republic of Korea (three studies) (*continued*)

| Reference and country | Technology/reading details for detection | Number of nodules or participants | Nodule type | Nodule size | Lung-RADS category | |
|---|---|---|---|---|---|---|
| | AVIEW Lungscreen (Coreline Soft); assisted 2nd-read mode | 10,424 participants | NR | NR | 1 – 53.0%<br>2 – 26.9%<br>3 – 11.7% | 4A – 4.3%<br>4B/X – 4.1% |
| Hwang *et al.* 2021,[52] Republic of Korea | AVIEW Lungscreen (Coreline Soft); 2nd-read mode (original institutional reading) | 3,452 nodules | Solid 94.1%; part-solid 1.5%; pure GGN 4.3% | Solid: median 5 mm; part-solid: median 12 mm; pure GGN: median 6 mm | NR | |
| | AVIEW Lungscreen (Coreline Soft); 2nd-read mode (original institutional reading) | 3,353 participants | NR | NR | 1 – 53.0%<br>2 – 26.9%<br>3 – 11.7% | 4A – 4.3%<br>4B/X – 4.1% |

GGN, ground-glass nodules; NR, not reported.

**TABLE 51** Nodule 2D axial diameter in all detected nodules in patients with complex lung disease[46]

| | Stand-alone software | | Unaided expert chest radiologist | |
|---|---|---|---|---|
| Lung condition | Number of nodules detected | Nodule size (mm), median (IQR) | Number of nodules detected | Nodule size (mm), median (IQR) |
| All | 312 | 8.4 (6.3–11.6) | 366 | 7.1 (5.3–10.5) |
| Interstitial lung disease | 59 | 8.4 (6.9–11.5) | 76 | 6.9 (5.5–10.2) |
| Chronic obstructive lung disease | 70 | 7.7 (6.0–10.7) | 68 | 6.0 (4.9–8.1) |
| Respiratory bronchiolitis | 59 | 7.6 (5.4–10.2) | 58 | 7.1 (4.9–9.2) |
| Oedema | 46 | 10.4 (7.2–13.8) | 63 | 8.4 (5.8–10.3) |
| Pulmonary embolism | 78 | 9.1 (6.5–13.9) | 101 | 8.2 (5.5–18.6) |

IQR, interquartile range.

the following composition: 57% solid, 14% part-solid, 15% ground glass and 14% calcified. Seventy-eight per cent of additional nodules had a diameter of 3–6 mm, and 22% were ≥ 6 mm in diameter.

*Screening population: InferRead CT Lung (Infervision) (one study)*

Zhang *et al.*[61] included 860 consecutive patients who had undergone lung cancer screening at one Chinese hospital as part of the NELCIN-B3 project. In the real-world radiologist observation, one of 14 residents drafted the diagnostic report, and one of 15 board-certified radiologists supervised the final version. In a MRMC study, one resident and one radiologist re-evaluated all CT images with the assistance of the InferRead CT Lung software to locate and measure the detected lung nodules. Consensus reading of two experienced radiologists detected at least one nodule in 43.5% (374/860) of participants, of which 66.8% (250/374) had solid nodules, 3.5% (13/374) had part-solid nodules and 29.8% (111/374) had ground-glass nodules. The size and type of the correctly detected nodules with and without software support as well as of the nodules detected by the reference standard are reported in *Table 53*. AI-assisted reading resulted in the correct detection of nodules in an additional 208 participants: 56% had solid nodules, 5% had part-solid nodules and 39% had ground-glass nodules. Of 126 additional participants with solid or part-solid nodules, 67% had a nodule diameter of ≤ 5 mm, and 33% had nodules that were ≥ 6 mm in diameter. The 82 additional participants with pure ground-glass nodules had nodules with a diameter < 20 mm.

b. Comparative results: stand-alone software versus unaided reader (one study)

One study[60] reported the proportions of detected nodules by size and type for unaided radiologists and stand-alone software as well as consensus expert reading.

*Mixed population: InferRead Lung CT (Infervision) (one study)*

Liu *et al.*[60] included a test set consisting of 1129 CT images (screening and inpatients) from more than 10 hospitals in China using convenience sampling. The chest CT images were retrospectively assessed by stand-alone software (InferRead Lung CT) as well by two experienced radiologists without software use. *Table 54* reports the proportions of detected nodules by size and type for unaided radiologists, stand-alone software and consensus expert reading (two experienced radiologists, reference standard).

Four studies reported on the characteristics of true-positive nodules detected by stand-alone software,[51,66] by software-assisted readers[56] and/or by the reference standard[32,56,66] without a comparator. These non-comparative results are reported in *Appendix 6*.

TABLE 52 Nodule type and size in a random symptomatic population from Japan[59]

| Detection details | Number of nodules | Nodule type, % (*n*) | Nodule size, % (*n*) |
|---|---|---|---|
| Reference standard | All 743 nodules | Solid 69.7% (518); part-solid 8.7% (65); calcified 10.0% (74); GGN 11.6% (86) | 3–6 mm 71.6% (532); 6–10 mm 19.4% (144); 19–15 mm 6.2% (46); 15–20 mm 1.9% (14); ≥ 20 mm 0.9% (7) |
| InferRead CT Lung (Infervision); concurrent mode | 564 true-positive nodules (reader A + reader B) | Solid 59.9% (338); part-solid 13.5% (76); calcified 14.4% (81); GGN 12.2% (69) | 3–6 mm 61.2% (345); 6–10 mm 24.3% (137); 10–15 mm 9.6% (54); 15–20 mm 3.0% (17); ≥ 20 mm 2.0% (11) |
| Unaided reader | 310 true-positive nodules (reader A + reader B) | Solid 62.3% (193); part-solid 13.2% (41); calcified 14.5% (45); GGN 10.0% (31) | 3–6 mm 47.1% (146); 6–10 mm 31.0% (96); 10–15 mm 15.2% (47); 15–20 mm 4.2% (13); ≥ 20 mm 2.6% (8) |
| InferRead CT Lung (Infervision); concurrent mode | 922 false-negative nodules (reader A + reader B) | Solid 75.7% (698); part-solid 5.9% (54); calcified 7.3% (67); GGN 11.2% (103) | 3–6 mm 78.0% (719); 6–10 mm 16.4% (151); 10–15 mm 4.1% (38); 15–20 mm 1.2% (11); ≥ 20 mm 0.3% (3) |
| Unaided reader | 1,176 false-negative nodules (reader A + reader B) | Solid 71.7% (843); part-solid 7.6% (89); calcified 8.8% (103); GGN 12.0% (141) | 3–6 mm 78.1% (918); 6–10 mm 16.3% (192); 10–15 mm 3.8% (45); 15–20 mm 1.3% (15); ≥ 20 mm 0.5% (6) |

GGN, ground-glass nodules.

TABLE 53 Nodule characteristics of participants with at least one nodule in a consecutive screening population from China, by mode of detection[61]

| Nodule type | Nodule diameter category | Participants with ≥ 1 nodule | | | Consensus reading (reference standard) |
|---|---|---|---|---|---|
| | | Unaided | AI assisted | Difference in numbers detected (%) | |
| All | All | 162 | 370 | + 128 | 374 |
| Solid | ≤ 5 mm | 65.4% (106/162) | 50.3% (186/370)[a] | + 75 | 50.3% (188/374) |
| | 6–7 mm | 9.9% (16/162) | 11.1% (41/370)[a] | + 156 | 11.2% (42/374) |
| | 8–14 mm | 4.9% (8/162) | 5.1% (19/370)[a] | + 138 | 5.1% (19/374) |
| | ≥ 15 mm | 0.6% (1/162) | 0.3% (1/370) | 0 | 0.3% (1/374) |
| | All | 80.9% (131/162) | 66.8% (247/370)[a] | + 89 | 66.8% (250/374) |
| Part-solid | ≤ 5 mm | 1.9% (3/162) | 2.1% (8/370)[a] | + 167 | 2.1% (8/374) |
| | ≥ 6 mm | 0 | 1.4% (5/370)[a] | N/A | 1.3% (5/374) |
| | All | 1.9% (3/162) | 3.5% (13/370)[a] | + 333 | 3.5% (13/374) |
| GGN | ≤ 19 mm | 17.3% (28/162) | 29.7% (110/370)[a] | + 293 | 29.7% (111/374) |
| | ≥ 20 mm | 0 | 0 | NA | 0 |
| | All | 17.3% (28/162) | 29.7% (110/370)[a] | + 293 | 29.7% (111/374) |

GGN, ground-glass nodules; N/A, not applicable.

[a] Indicates significant difference ($p < 0.001$) by the chi-squared test between unaided and AI-assisted reading.

**TABLE 54** Characteristics of correctly detected nodules in a mixed population from China obtained via convenience sampling[60]

| | | | Correctly detected nodules | | |
|---|---|---|---|---|---|
| **Nodule type** | **Nodule size** | **Reference standard** | **Stand-alone software** | **Reader 1: unaided** | **Reader 2: unaided** |
| Total | All | 6,363 | 4,484 | 2,562 | 3,617 |
| Solid | ≤ 6 mm | 53.4% | 50.0% | 49.5% | 47.1% |
| | > 6 mm | 4.1% | 5.1% | 8.1% | 5.1% |
| | All | 57.5% | 55.1% | 57.6% | 52.3% |
| Subsolid | ≤ 5 mm | 20.8% | 19.6% | 13.1% | 20.8% |
| | > 5 mm | 6.8% | 7.9% | 10.0% | 10.1% |
| | All | 27.6% | 27.5% | 23.1% | 30.9% |
| Calcified | NR | 5.1% | 6.6% | 6.0% | 5.1% |
| Pleural | NR | 9.8% | 10.7% | 13.3% | 11.7% |

NR, not reported.

**TABLE 55** Characteristics of all detected nodules and true-positive, false-positive and false-negative nodules: stand-alone software in a random mixed population[66]

| **Reference and country** | **Detection details** | **Number of nodules** | **Nodule type** | **Nodule size (mm), mean (± SD)** |
|---|---|---|---|---|
| Martins Jarnalo *et al*. 2020,[66] the Netherlands | Reference standard | 93 nodules | Solid 80%<br>Subsolid 18%<br>Mixed solid/subsolid 2% | 7.0 ± 4.1 |
| | Veye Chest (Aidence); stand-alone | 130 detected nodules (TP and FP) | Solid 85%<br>Subsolid 14%<br>Mixed solid/subsolid 1% | 9.0 ± 7.1 |
| | | 80 TP nodules | Solid 81%<br>Subsolid 16%<br>Mixed solid/subsolid 3% | 7.0 ± 3.8 |
| | | 50 FP nodules | Solid 90%<br>Subsolid 10%<br>Mixed solid/subsolid 0% | 11.8 ± 10.0 |
| | | 11 FN nodules | Solid, 4 mm: *n* = 5<br>Solid, calcified, 4 mm: *n* = 3<br>Subsolid, 4 mm: *n* = 1<br>Subsolid, 18 mm: *n* = 1<br>Subsolid, 20 mm: *n* = 1 | 6.7 ± 6.1 |

FP, false positive; SD, standard deviation; TP, true positive.

### *False-positive nodules*

No comparative results were available. Non-comparative results are reported in *Appendix 6*.

## False-negative (missed) nodules (nine studies)

a. Comparative results: reader with and without software (two studies)

*Symptomatic population: InferRead CT Lung (Infervision) (one study)*

Kozuka *et al.*[59] randomly selected 120 chest CT images from cases of suspected lung cancer at a single hospital in Japan. In a MRMC study, two less experienced radiologists assessed the CT images first without software (InferRead CT Lung, Infervision) for nodule detection and then with software. The distribution of size and type of missed lung nodules of readers with and without software support are reported in *Table 56*. With software use, the two readers missed fewer nodules (922 vs. 1176; −22%); false negatives were reduced by 145 (−17.2%) for solid, by 35 (−39.3%) for part-solid, by 36 (−35.0%) for calcified and by 38 (−27.0%) for ground-glass nodules compared with unaided reading.

*Screening population: InferRead CT Lung (Infervision) (one study)*

Zhang *et al.*[61] included 860 consecutive patients who had undergone lung cancer screening at one Chinese hospital as part of the NELCIN-B3 project. In the real-world radiologist observation, one of 14 residents drafted the diagnostic report, and one of 15 board-certified radiologists supervised the final version. In a MRMC study, one resident and one radiologist re-evaluated all subjects with the assistance of the InferRead CT Lung software to locate and measure the detected lung nodules. Of the 212 participants with nodules that were missed by unaided readers in clinical practice, 56.1% had solid nodules, 4.7% had part-solid nodules and 39.2% had ground-glass nodules (*Table 56*). Missed nodules were solid and > 5 mm in 17.5%, part-solid and > 5 mm in 2.4% and ground-glass nodules < 20 mm in 39.2%. In the reader study, AI-assisted readers missed four participants with at least one nodule. Of these, two (50%) had

**TABLE 56** Characteristics of missed nodules in a consecutive screening population from China[61]

| | | Missed subjects with ≥ 1 nodule | | |
|---|---|---|---|---|
| **Nodule type** | **Nodule diameter category** | **Unaided (clinical practice)** | **AI assisted (MRMC study)** | **Difference in numbers missed, (%)** |
| All | All | 212 | 4 | 208 (−98.1%) |
| Solid | ≤ 5 mm | 38.7% (82/212) | 50.0% (2/4) | −80 (−97.6%) |
| | 6–7 mm | 12.3% (26/212) | 25.0% (1/4) | −25 (−96.2%) |
| | 8–14 mm | 5.2% (11/212) | 0 | −11 (−100.0%) |
| | ≥ 15 mm | 0 | 0 | 0 |
| | All | 56.1% (119/212) | 75.0% (3/4) | −116 (−97.5%) |
| Part-solid | ≤ 5 mm | 2.4% (5/212) | 0 | −5 (−100.0%) |
| | ≥ 6 mm | 2.4% (5/212) | 0 | −5 (−100.0%) |
| | All | 4.7% (10/212) | 0 | −10 (−100.0%) |
| GGN | ≤ 19 mm | 39.2% (83/212) | 25.0% (1/4) | −82 (−98.8%) |
| | ≥ 20 mm | 0 | 0 | 0 |
| | All | 39.2% (83/212) | 25.0% (1/4) | −82 (−98.8%) |

GGN, ground-glass nodules.

solid nodules ≤ 5 mm, one (25%) had solid nodules > 5 mm and the remaining participant had a ground-glass nodule < 20 mm. The absolute reduction in missed nodules with software use was largest for ground-glass nodules ≤ 19 mm and for solid nodules ≤ 5 mm (an additional 82 and 80 nodules detected with concurrent software use, respectively). Relative reduction was slightly higher for part-solid (−100.0%) and ground-glass nodules (−98.8%) than for solid nodules (−97.5%).

b. Comparative results: stand-alone AI versus unaided reader (two studies)

*Mixed population: AI-Rad Companion Chest CT (Siemens Healthineers) (one study)*

Abadia *et al.*[47] included 103 patients with at least one lung condition and one suspicious lung nodule (≥ 4 mm) on radiology report and 40 patients with one lung condition and no lung nodule on radiology report from random LDCT images taken at a single US hospital for various reasons. The median 2-D axial size of the 29.3% (129/441) nodules missed by stand-alone software (AI-Rad Companion CT Chest prototype, Siemens Healthineers) was 8.9 mm (IQR 5.7–14.4 mm), whereas the unaided expert chest radiologist missed 8.4% (37/441) of nodules with a median size of 6.1 mm (IQR 5.1–9.2 mm). Most of the nodules missed by the nodule detection software were near the pleura; occasionally, hilar and basilar nodules were also missed.

*Mixed population: InferRead Lung CT (Infervision) (one study)*

Liu *et al.*[60] included a test set consisting of 1129 CT images (screening and inpatients) from more than 10 hospitals in China using convenience sampling. The chest CT images were retrospectively assessed by stand-alone software (InferRead Lung CT) as well by two experienced radiologists without software use. *Table 57* reports the proportions of missed nodules by size and type for the stand-alone software as well as for the unaided radiologists.

Non-comparative results (five studies) are reported in *Appendix 6*.

**TABLE 57** Characteristics of missed nodules in a mixed population from China obtained via convenience sampling[60]

| | | Missed nodules | | |
|---|---|---|---|---|
| **Nodule type** | **Nodule size** | **Stand-alone software** | **Reader 1: unaided** | **Reader 2: unaided** |
| Total | All | 1879 | 3,801 | 2,746 |
| Solid | ≤ 6 mm | 61.5% | 56.1% | 61.7% |
| | > 6 mm | 1.6% | 1.4% | 2.7% |
| | All | 63.1% | 57.4% | 64.4% |
| Subsolid | ≤ 5 mm | 23.7% | 26.1% | 20.9% |
| | > 5 mm | 4.2% | 4.6% | 2.4% |
| | All | 27.9% | 30.7% | 23.3% |
| Calcified | NR | 1.5% | 4.4% | 5.0% |
| Pleural | NR | 7.6% | 7.4% | 7.3% |

NR, not reported.

## Descriptions of evidence from individual studies providing data on proportion of detected nodules that are malignant [summarised in *Proportion of detected nodules that are malignant (three studies)*]

a. Comparative results: reader with and without software (two studies)

*Screening population: Veolity (MeVis) (one study)*

The study by Hall *et al.*[27] was performed in London (UK) and is a substudy of the LSUT trial. It comprised all 770 patients who received LDCT for lung cancer screening. In a MRMC study, two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCT with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm). The study compared the findings with the numbers of nodules ≥ 5 mm detected by the original unaided reading (single expert thoracic radiologists, with 5% of CT images checked by a second radiologist). In the original, unaided reading, 21.3% (33/155) of all detected actionable nodules were malignant: 60.0% (18/30) of all actionable nodules with direct referral to a MDT ('suspicious lesions') and 12.0% (15/125) of all actionable nodules referred for CT surveillance ('intermediate nodules'). Of the actionable nodules detected by radiographer 1 with concurrent software use, 16.7% (24/144) were malignant; of those detected by radiographer 2, the proportion of malignant nodules was 19.4% (30/155).

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

In a before-and-after study, Hwang *et al.* included 6487 consecutive participants of the K-LUCAS project: 1821 participants were screened before the AVIEW Lungscreen software was implemented and 4666 participants were screened after.[51] A whole read (nodule detection and classification based on nodule type and size) was performed by a single experienced thoracic radiologist with or without concurrent software use (AVIEW Lungscreen, Coreline Soft) in a clinical setting. Positivity was based on Lung-RADS category ≥ 3, and cases of lung cancer were identified by medical record review. The proportion of all detected nodules (Lung-RADS category ≥ 2) that were later diagnosed as lung cancer was 1.2% (16/1391) before the implementation of the software and 0.6% (31/4990) after. Of the screen-positive (actionable) nodules (Lung-RADS category ≥ 3), 6.6% (14/212) and 5.2% (28/538) were malignant before and software implementation, respectively.

b. Non-comparative results (one study)

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

The other study by Hwang *et al.*[50] comprised 10,424 concurrent baseline LDCT scans obtained after the AVIEW Lungscreen software was implemented as part of the Korean K-LUCAS project. The number of lung cancers (within 1 year after LDCT and any lung cancers after LDCT) by nodule type and size of the risk-dominant nodule is reported in *Appendix 5*, *Table 58*. In all 4642 risk-dominant nodules, 1.1% (52/4642) were diagnosed as lung cancer within 1 year after LDCT, and 1.2% (58/4642) were diagnosed with any lung cancer after LDCT. The highest proportion of malignant nodules was found among solid nodules ≥ 10 mm (14%) and among part-solid nodules (13%).

## Descriptions of evidence from individual studies providing data on the impact of test result on clinical decision-making [summarised in *Impact of test result on clinical decision-making (six studies)*]

a. Comparative results: reader with and without software (six studies)

*Screening population: MeVis (two studies)*

The study by Jacobs *et al.*[64] comprised a nodule-enriched screening population. One-hundred and sixty LDCT images were selected from the US-based NLST data set stratified by Lung-RADS category (*n* = 40 Lung-RADS 1 or 2, *n* = 40 Lung-RADS 3, *n* = 40 Lung-RADS 4A and *n* = 40 Lung-RADS 4B, with half being baseline scans and half being 1-year

DOI: 10.3310/JYTW8921

TABLE 58 Proportion of detected risk-dominant nodules that are malignant, by nodule type and size, in a consecutive screening population from the Republic of Korea[50]

| | Solid nodules | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | < 6 mm | 6–7 mm | 7–8 mm | 8–9 mm | 9–10 mm | ≥ 10 mm | Part-solid nodules | Non-solid nodules | Total |
| ***Average transverse diameter*** | | | | | | | | | |
| Risk-dominant nodule (*n*) | 3,570 | 304 | 130 | 83 | 53 | 217 | 125 | 160 | 4,642 |
| Lung cancer diagnosed within 1 year after LDCT, *n* (%) | 2 (0.06) | 0 (0) | 1 (0.77) | 0 (0) | 4 (7.55) | 30 (13.8) | 15 (12.00) | 0 (0) | 52 (1.12) |
| Any lung cancer diagnosed after LDCT, *n* (%) | 5 (0.14) | 1 (0.33) | 1 (0.77) | 0 (0) | 5 (9.43) | 30 (13.8) | 16 (12.80) | 0 (0) | 58 (1.25) |
| ***Effective diameter*** | | | | | | | | | |
| Risk-dominant nodule (*n*) | 3,574 | 301 | 131 | 80 | 53 | 217 | 126 | 160 | 4,642 |
| Lung cancer diagnosed within 1 year after LDCT, *n* (%) | 2 (0.06) | 1 (0.33) | 0 (0) | 0 (0) | 4 (7.55) | 30 (13.8) | 15 (11.90) | 0 (0) | 52 (1.12) |
| Any lung cancer diagnosed after LDCT, *n* (%) | 5 (0.14) | 1 (0.33) | 1 (0.77) | 0 (0) | 5 (9.43) | 30 (13.8) | 16 (12.70) | 0 (0) | 58 (1.25) |

LDCT, Low-dose computed tomography.

follow-up scans). Seven readers participated in the MRMC study detecting nodules ≥ 3 mm and classifying the Lung-RADS category of the risk-dominant nodule based on its type and size with and without concurrent use of the software Veolity (MeVis) (*Table 59*).

**TABLE 59** Lung-RADS category with and without concurrent software use in a nodule-enriched screening population[64]

| Lung-RADS category | Seven readers with concurrent software use (*n* = 160 LDCT scans each) | Seven readers without concurrent software use (*n* = 160 LDCT scans each) |
|---|---|---|
| 1 or 2 (negative) | 34% (377/1,120) | 47% (521/1,120) |
| 3 | 21% (232/1,120) | 18% (199/1,120) |
| 4A | 23% (252/1,120) | 15% (166/1,120) |
| 4B | 23% (259/1,120) | 21% (234/1,120) |

LDCT, Low-dose computed tomography.

Jacobs *et al.* found that the proportion of scans with a Lung-RADS category of 1 or 2 (negative screening result) was substantially reduced from 47% to 34% when the dedicated CT lung screening viewer with software support was used, whereas the total number of positive screening results (Lung-RADS category 3, 4A or 4B) increased from 53% to 66%. The spread of Lung-RADS results for readers with concurrent software use was more in line with how the cases were selected from the NLST database (25% in each category).

The study by Hall *et al.*[27] was performed in London (UK) and is a substudy of the LSUT trial. It comprised all 770 patients who received LDCT for lung cancer screening. In a MRMC study, two radiographers without previous experience of thoracic CT reporting independently read all 770 LDCT with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm) and common incidental findings and had to make patient management recommendations. The study reports on the concordance of management decisions against BTS guidelines[12] for the software-assisted radiographer as well as for the original unaided reading (single expert thoracic radiologists, with 5% of CT images checked by a second radiologist). For radiographer 1, the management recommendations for 39.7% (52/131) of CT scans were concordant with the BTS guidelines (15 cancers), for 19.8% (26/131) a more active follow-up was recommended (one cancer) and for 40.5% (53/131) a less active follow-up was recommended (three cancers). For radiographer 2, the management recommendations for 60.7% (91/150) of CT scans were concordant with the BTS guidelines (22 cancers), for 23.3% (35/150) a more active follow-up was recommended (four cancers) and for 16.0% (24/150) a less active follow-up was recommended (one cancer). For the original unaided radiologists, the management recommendations for 71.6% (111/155) of CT scans were concordant with the BTS guidelines (28 cancers), for 14.2% (22/155) a more active follow-up was recommended (three cancers) and for 12.9% (20/155) a less active follow-up was recommended (one cancer).

*Screening population: VUNO Med-Lung CT AI (VUNO) (one study)*

Park *et al.*[67] included a nodule- and cancer-enriched screening population (200 baseline LDCT), selected from the US-based NLST data set. In a MRMC study, five readers with varying levels of experience assessed the LDCT images with and without concurrent software use (VUNO Med-Lung CT AI, VUNO). The readers reported 71.5% negative screening results (Lung-RADS categories 1 and 2) without software use and 65.8% negative screening results with software use (*Table 60*).

In the majority of cases, the Lung-RADS categories remained unchanged between the two sessions for all readers [74.5% (149/200)–91.0% (182/200)]. With software use, the readers tended to upstage (average 12.3%) rather than downstage Lung-RADS categories (average 4.4%) compared with unaided reading, with most of the changes occurring between two contiguous categories. An upstage from screen-negative (Lung-RADS category 1 or 2) to screen-positive (Lung-RADS category ≥ 3) occurred in 6 out of 200 (3%) to 26 out of 200 (13%) of CT images that were assessed with software use. Between 0 and 18 out of 200 (9%) of CT images were downstaged by the five readers with software use compared with unaided reading.

**TABLE 60** Lung-RADS category based on stand-alone software and readers with and without concurrent software use in a nodule-enriched screening population[67]

| Lung-RADS category | Stand-alone software (*n* = 200 LDCTs) | Five readers with concurrent software use (*n* = 200 LDCT scans each) | Five readers without concurrent software use (*n* = 200 LDCT scans each) |
|---|---|---|---|
| 1 or 2 (negative) | 53.0% (106/200) | 65.8% (658/1,000) | 71.5% (715/1,000) |
| 3 | 15.5% (31/200) | 11.1% (111/1,000) | 9.0% (90/1,000) |
| 4A | 14.0% (28/200) | 10.5% (105/1,000) | 9.3% (93/1,000) |
| 4B | 17.5% (35/200) | 12.6% (126/1,000) | 10.2% (102/1,000) |

LDCT, Low-dose computed tomography.

With regard to patient management, the mean follow-up periods determined by the five unaided readers were 9.4 (range 9.1–9.8 months) and 8.9 months with concurrent software use (range 8.7–9.3 months). Although all readers gave a shorter mean follow-up interval with software use, the change was minor, being an average of 0.5 months (range 0.3–0.7 months).

For the 31 cancer-positive cases in the data set, substantial management discrepancies between the 310 reader pairs (Lung-RADS category 1/2 vs. 4A/B) were reduced in half with application of the software (32/310 to 16/310).

*Screening population: ClearRead CT (Riverain Technologies) (one study)*

Singh *et al.*[56] included 150 patients who underwent LDCT of the chest as part of the NLST: the first 125 patients with subsolid nodules (154 part-solid or 156 ground-glass nodules between 6 and 30 mm) and the first 25 patients with no nodules detected. As part of a MRMC study, two experienced chest radiologists sequentially interpreted the unprocessed CT images alone and then together with the vessel-suppressed (ClearRead CT, Riverain Technologies) CT image without washout period. Using vessel-suppressed images, both radiologists detected solid components in five part-solid nodules, which they had deemed as ground-glass nodules on the standard CT images. The Lung-RADS category of these five nodules changed from 2 to 4A, which would impact the management of these patients.

*Surveillance population with applicability concerns: Veolity (MeVis) (one study)*

Kim *et al.*[63] included 89 patients with subsolid nodules who underwent preoperative non-enhanced CT and subsequent surgical resection at the Seoul National University Hospital. In a MRMC study, nodule classification based on diameter measurements of 102 subsolid nodules obtained by two experienced radiologists were compared with and without concurrent use of Veolity (MeVis). The subsolid nodules were categorised according to Fleischner Society guidelines[70] into (1) pure ground glass, (2) part-solid with a solid component ≥ 5 mm or (3) part-solid with a solid component < 5 mm. Based on the solid component size (5-mm cut-off), the management recommendations for part-solid nodules by the Fleischner Society suggest surveillance CT or invasive procedures (biopsy or surgical resection). With software use for semiautomatic nodule measurement, both reader 1 and reader 2 classed more part-solid nodules as having a solid portion > 5 mm than did manual measurement (59.8% vs. 43.1% for reader 1; 58.8% vs. 55.9% for reader 2-1; 61.8% vs. 53.9% for reader 2-2; *Table 61*), which would suggest that, with software use, more people would receive invasive procedures and fewer people would receive CT surveillance.

*Unclear indication for CT scan: ClearRead CT (Riverain Technologies) (one study)*

This MRMC study by Milanese *et al.*[55] included 93 consecutive patients referred to University Hospital Zurich (Switzerland) for clinical non-enhanced, low-dose chest CT between August 2014 and February 2015 (unclear indication for the chest CT scan). One radiologist with 3 years of experience in chest CT and a radiology resident independently performed semiautomatic volume measurements of 65 solid nodules using the software 'MM Oncology' by Siemens Healthcare on vessel-suppressed (ClearRead CT, Riverain Technologies) as well as on standard CT images.

**TABLE 61** Subsolid nodule classification of the two readers with and without software use in patients with previously detected nodules[63]

| | | Reader 1 | Reader 2-1 | Reader 2-2 |
|---|---|---|---|---|
| With software for semi-automatic measurement | Pure ground glass | 21 (20.6%) | 19 (18.6%) | 16 (15.7%) |
| | Part-solid with solid portion < 5 mm | 20 (19.6%) | 23 (22.5%) | 23 (22.5%) |
| | Part-solid with solid portion ≥ 5 mm | 61 (59.8%) | 60 (58.8%) | 63 (61.8%) |
| Manual measurement | Pure ground glass | 19 (18.6%) | 15 (14.7%) | 18 (17.6%) |
| | Part-solid with solid portion < 5 mm | 39 (38.2%) | 30 (29.4%) | 29 (28.4%) |
| | Part-solid with solid portion ≥ 5 mm | 44 (43.1%) | 57 (55.9%) | 55 (53.9%) |

They categorised nodules according to Fleischner Society guidelines into < 100 mm$^3$, 100–250 mm$^3$ and > 250 mm$^3$.[68] With vessel suppression, reader 1 changed the nodule category from 100 to 250 mm$^3$ to < 100 mm$^3$ for two nodules, whereas reader 2 changed the nodule category for two nodules from the 100 to 250 mm$^3$ category to < 100 mm$^3$ and > 250 mm$^3$, respectively (*Table 62*).

**TABLE 62** Risk classification based on semiautomatic volume measurement using standard CT images and vessel-suppressed CT images in consecutive LDCT with unclear indication[55]

| Type of images | Size of nodules | Reader 1 (*n* = 65 solid nodules) | Reader 2 (*n* = 65 solid nodules) | Total (*n* = 130 solid nodules) |
|---|---|---|---|---|
| Standard CT | < 100 mm$^3$ | 48 (73.8%) | 48 (73.8%) | 96 (73.8%) |
| | 100–250 mm$^3$ | 11 (16.9%) | 11 (16.9%) | 22 (16.9%) |
| | > 250 mm$^3$ | 6 (9.2%) | 6 (9.2%) | 12 (9.2%) |
| Vessel-suppressed CT | < 100 mm$^3$ | 50 (76.9%) | 49 (75.4%) | 99 (76.2%) |
| | 100–250 mm$^3$ | 9 (13.8%) | 9 (13.8%) | 18 (13.8%) |
| | > 250 mm$^3$ | 6 (9.2%) | 7 (10.8%) | 13 (10.0%) |

## Descriptions of evidence from individual studies providing data on the number of people having computed tomography surveillance [summarised in *Number of people having computed tomography surveillance (five studies)*]

a. Comparative results: reader with and without software (two studies)

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

In a before-and-after study, Hwang *et al.*[51] included 6487 consecutive participants of the K-LUCAS project: 1821 participants were screened before the AVIEW Lungscreen software was implemented and 4666 participants were screened after. Before software implementation, unaided single expert chest radiologists manually measured the transverse plane of the risk-dominant nodules and classed 7.6% (139/1821) of participants as Lung-RADS categories 3 or 4A. After software implementation, single expert chest radiologists classed 9.0% (418/4666) of participants as Lung-RADS categories 3 or 4A based on transverse planes. Among these people with intermediate-risk lung nodules, 2.9% (4/139) and 0.7% (3/418) were diagnosed with lung cancer before and after software implementation, respectively. This suggests that around 93% (135/139) and 99% (415/418), respectively, would have received unnecessary CT surveillance.

*Screening population: Veolity (MeVis) (one study)*

The study by Jacobs *et al.*[64] included a nodule-enriched screening population. One hundred and sixty LDCT images were selected from the US-based NLST data set stratified by Lung-RADS category (*n* = 40 Lung-RADS 1 or 2, *n* = 40 Lung-RADS 3, *n* = 40 Lung-RADS 4A and *n* = 40 Lung-RADS 4B, with half being baseline scans and half being 1-year follow-up scans). Seven readers participated in the MRMC study detecting nodules ≥ 3 mm and classifying the Lung-RADS category of the risk-dominant nodule based on its type and size with and without concurrent use of the software Veolity (MeVis). Without software use, the seven readers classed 32.6% (365/1120) as Lungs-RADS categories 3 or 4A. By contrast, 43.2% (484/1120) were classed as Lung-RADS categories 3 or 4A with concurrent software use.

Non-comparative results (three studies) are reported in *Appendix 6*.

## Descriptions of evidence from individual studies providing data on the number of people having a biopsy or excision [summarised in *Number of people having a biopsy or excision (five studies)*]

a. Comparative results: reader with and without software (two studies)

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

In a before-and-after study, Hwang *et al.*[51] included 6487 consecutive participants of the K-LUCAS project. Before software implementation, unaided single expert chest radiologists manually measured the transverse plane of the risk-dominant nodules and classed 2.3% (41/1821) of participants as Lung-RADS categories 4B or 4X. After software implementation, a single expert chest radiologist classed 2.0% (93/4666) of participants as Lung-RADS categories 4B or 4X based on transverse planes. Among these people with highly suspicious lung nodules, 26.8% (11/41) and 26.9% (25/93) were diagnosed with lung cancer before and after software implementation, respectively. This suggest that around 73% (30/41 and 68/93, respectively) might have received unnecessary follow-up investigations.

*Screening population: Veolity (MeVis) (one study)*

The study by Jacobs *et al.*[64] included a nodule-enriched screening population. One hundred and sixty LDCT images were selected from the US-based NLST data set based on Lung-RADS category (*n* = 40 Lung-RADS 1 or 2, *n* = 40 Lung-RADS 3, *n* = 40 Lung-RADS 4A and *n* = 40 Lung-RADS 4B, with half being baseline scans and half being 1-year follow-up scans). Seven readers participated in the reader study detecting nodules ≥ 3 mm and classifying the Lung-RADS category of the risk-dominant nodule based on its nodule type and size with and without concurrent use of the software Veolity (MeVis). Without software use, the seven readers classed 21% (234/1120) as Lungs-RADS category 4B. With concurrent software use, the seven readers classed 23% (259/1120) CT images as Lung-RADS categories 4B.

Non-comparative results (three studies) are reported in *Appendix 6*.

# **Appendix 6** Additional evidence on test accuracy of stand-alone artificial intelligence and other evidence from non-comparative studies

## Accuracy for detecting any nodules

### *Stand-alone artificial intelligence versus unassisted reader (four studies)*

- Symptomatic population (one study)

*Kozuka et al. 2020,*[59] *Japan: InferRead CT Lung (Infervision)*

Kozuka *et al.*[59] randomly selected 120 chest CT images (117 cases included in analysis) from cases with suspected lung cancer. Two less experienced radiologists assessed the CT images with and without software use; stand-alone software performance was also reported. Per-patient sensitivity was 95.5% (95% CI 89.9% to 98.5%) for stand-alone AI and 68.0% (95% CI 61.45% to 74.1%) for the pooled unaided readers. Per-patient specificity was 83.3% (95% CI 35.9% to 99.6%) for stand-alone AI and 91.7% (95% CI 61.5% to 99.8%) for the pooled unaided readers. Per-nodule sensitivity was 70.3% (95% CI 66.8% to 73.5%) for stand-alone AI and 20.9% (95% CI 18.8% to 23.0%) for the pooled unaided readers. Stand-alone AI had a positive predictive value of 57.9% (95% CI 54.6 to 61.1%), and the pooled unaided readers' positive predictive value was 70.5% (95% CI 66.0% to 74.7%).

- Incidental population (one study)

*Rückel et al. 2021,*[49] *Germany: AI-Rad Companion (Siemens Healthineers)*

Rückel *et al.*[49] reported data from 105 consecutive patients who received a whole-body CT scan in the emergency department (shock room) at the LMU University Hospital (Munich, Germany) from January to November 2019. An on-premises prototype not yet commercially available has been used in this work. The reference standard was the original radiology report [reading by single board-certified radiologist alone (17%) or commonly reported by a radiology resident and a board-certified radiologist (83%)], with additional software-detected nodules verified by an expert. The per-nodule sensitivity was 96.7% (29/30) for stand-alone AI and 90.0% (27/30) for the original unaided reading, with an average 0.74 false positive per image (78/105) detected by the software. Per-patient sensitivity was 92.9% (13/14) for stand-alone AI and 85.7% (12/14) for the original unaided reading. The positive predictive value of stand-alone AI was 20.0% (13/65).

- Mixed population (two studies)

*Abadia et al. 2021,*[47] *USA: AI-Rad Companion (Siemens Healthineers)*

Abadia *et al.*[47] performed a retrospective test accuracy and MRMC study using a case–control data set (103 patients with at least one lung condition and one suspicious lung nodule on radiology report; 40 patients with one lung condition and no lung nodule on radiology report) from a single centre. One of five expert chest radiologists analysed the CT images in clinical practice (original radiology reports). The reference standard consisted of nodules in the radiology report plus additional nodules detected by stand-alone AI and validated by a single expert. The AI-Rad prototype had a sensitivity to detect the (up to) three largest nodules per patient of 89.4% (186/208). The original radiologist report correctly detected 76.9% (160/208) of the (up to) three largest nodules per patient.

Additionally, one expert chest radiologist with 15 years of experience assessed all 103 CT images with nodules as part of a MRMC study. The reference standard consisted of all radiologist-detected nodules plus additional nodules detected by stand-alone software and assessed by the radiologist as true positives. Stand-alone software had a per-nodule sensitivity of 67.7% (270/399; four nodules with wrong location seemed to have been excluded from the analysis) with an average 0.37 false-positive detections per image (38/103). The unaided expert reader correctly detected 90.8% (366/403) nodules with no false-positive detections as per definition of the reference standard.

*Liu et al. 2019,*[60] *China: InferRead CT Lung (Infervision)*

Liu *et al.*[60] included 1129 chest CT scans from multiple hospitals in China with convenience sampling. Two experienced radiologists assessed the CT images unaided under laboratory conditions. The per-nodule sensitivity was 70.4% (4481/6363) for stand-alone AI and 48.6% (6179/12,726) for the two pooled unaided readers. The false-positive rate for stand-alone AI was 46.5% (3894 false positive/8375 detected nodules) and an average 3.4 per scan (3894 false positives per 1129 scans), respectively. Using a free-response receiver operating characteristic curve, the performance of stand-alone AI was demonstrated: at an average of one false-positive detection per scan, the per-nodule sensitivity was 74%. Sensitivity reached a maximum of 86% with an average of eight false-positive detections per scan.

### *Non-comparative results (six studies)*

Six studies[30,47,51,58,65,66] evaluated *accuracy for detecting any nodules by stand-alone AI* without a comparator (*Figure 3*). Of these, one included a screening population,[51] and five included mixed populations.[30,47,58,65,66] The key characteristics and findings of studies with non-comparative outcomes are shown in *Table 4*.

- Screening population (one study)

*Hwang et al. 2021,*[51] *Republic of Korea: AVIEW LCS+ (Coreline Soft)*

Hwang *et al.*[51] included 4666 participants who had undergone lung cancer screening as part of the K-LUCAS project after the implementation of the software AVIEW Lungscreen (Coreline Soft). They reported a per-nodule sensitivity of 50.2% (2147/4280; 95% CI 48.7% to 51.7%) for the stand-alone software. The reference standard was the original reader decision (25 different, single experienced chest radiologists with 5–28 years of experience) with assisted 2nd-read software use. The original radiologist rejected 73.6% (5981/8128) of software-detected nodules as false positives (average 1.51 false-positive detections per image).

- Mixed population (five studies)

Information from a study identified in a report submitted by Aidence was considered confidential and has been removed from this section.

*Wan et al. 2020,*[58] *Taiwan: ClearRead CT (Riverain Technologies)*

Wan *et al.*[58] performed a retrospective analysis in 50 patients with 75 pathologically proven (benign or malignant) nodules ≤ 2 cm from hospitals in Taiwan. The stand-alone software had 81.3% (61/75) per-nodule sensitivity. The false-positive rate was not reported.

*Abadia et al. 2021,*[47] *USA: AI-Rad Companion (Siemens Healthineers)*

Abadia *et al.*[47] performed a retrospective test accuracy and MRMC study using a case–control data set (103 patients with at least one lung condition and one suspicious lung nodule on radiology report, 40 patients with one lung condition and no lung nodule on radiology report) from a single centre. The AI-Rad prototype assessment of the control population showed 82.5% (33/40) specificity. When tasked with classifying each of the 143 patients into nodule present or absent, the stand-alone software had a specificity of 77.5% (31/40) and a sensitivity of 96.1% (99/103).

*Blazis et al. 2021,*[65] *Netherlands: Veye Lung Nodules (Aidence)*

Blazis *et al.*[65] evaluated the performance of the stand-alone software with different reconstruction algorithms and reconstruction settings by retrospectively analysing 384 CT reconstructions from 24 patients from a hospital in the Netherlands. At a software sensitivity threshold of 0.86, the observed per-nodule sensitivity ranged from 57% to 96% depending on the reconstruction setting, with the average false positive per image ranging from 0.25 to 1.16. On the clinically preferred Thorax CT reconstructions (Br54f3 and I50f3) at 1.0 mm slice thickness, the per-nodule sensitivity was 83%.

*Martins Jarnalo et al. 2021,*[66] *the Netherlands: Veye Lung Nodules (Aidence)*

Martins Jarnalo *et al.*[66] randomly selected 145 patients with 145 CT images from a large teaching hospital in the Netherlands. CT examinations had been performed for various indications, ranging from ruling out metastases, follow-up of nodules and follow-up of other pulmonary abnormalities, to other miscellaneous indications. The per-nodule sensitivity of the stand-alone software was 87.9% (80/91) for all nodules, with 89.0% (65/73) of solid nodules, 81.3% (13/16) of subsolid nodules and 100.0% (2/2) of mixed (solid/subsolid) nodules correctly detected. The false-positive rate for the detection of all nodules was 38.5% (average 1.04 false positives per scan).

## Accuracy for detecting actionable nodules

### *Stand-alone AI versus unassisted reader (two studies)*

- Symptomatic population (one study)

*Kozuka et al. 2020,*[59] *Japan: InferRead CT Lung (Infervision)*

Kozuka *et al.*[59] randomly selected 120 chest CT images (117 cases included in analysis) from cases with lung cancer suspicion. They performed a MRMC study with two less experienced radiologists (1 year and 5 years of experience). Stand-alone AI had a per-nodule sensitivity of 61.1% (129/211), whereas the pooled unaided readers correctly detected 38.9% (164/422) of nodules ≥ 6 mm (no level of significance reported). The false-positive rate was not reported.

- Mixed population (one study)

*Liu et al. 2019,*[60] *China: InferRead CT Lung (Infervision)*

Liu *et al.*[60] included 1129 chest CT scans from multiple hospitals in China with convenience sampling. Two experienced radiologists assessed the CT images unaided under laboratory conditions. The per-nodule sensitivity for the detection of solid nodules > 6 mm and subsolid nodules > 5 mm combined was 84.1% (581/691) for stand-alone AI and 73.4% (1015/1382) for the pooled unassisted readers (no level of significance reported).

### *Non-comparative results (two studies)*

Two studies[30,48] evaluated the *accuracy for detecting actionable nodules by stand-alone AI*. Of these, one included a screening population[48] and one included a mixed population.[30]

- Screening population (one study)

*Chamberlin et al. 2021,*[48] *USA: AI-Rad Companion (Siemens Healthineers)*

Chamberlin *et al.*[48] evaluated 117 randomly selected LDCT studies that were performed for routine lung cancer screening between January 2018 and July 2019 in one US hospital. For stand-alone software, the study found 100% per-nodule sensitivity (132/132) and 100% per-patient sensitivity (69/69). The specificity was 70.8% (34/48) by patient and 37.8% (34/90) by nodule. A false-positive rate of 12.0% (14/117) per patient and 25.2% (56/222) per nodule (0.48 false positive/scan) was observed.

- Mixed population (one study)

Information from a study identified in a report submitted by Aidence was considered confidential and has been removed from this section.

## Accuracy for detecting malignant nodules

### *Stand-alone artificial intelligence versus unassisted reader (no study)*

No data available.

### *Non-comparative results (three studies)*

Three studies[27,51,58] evaluated *accuracy for detecting malignant nodules by stand-alone AI*[51,58] *or with concurrent software use*.[27] Of these, two included a screening population[27,51] and one included a mixed population.[58]

- Screening population (two studies)

*Hwang et al. 2021,*[51] *Republic of Korea: AVIEW LCS+ (Coreline Soft)*

Hwang *et al.*[51] included 4666 participants who had undergone lung cancer screening as part of the K-LUCAS project after the software AVIEW Lungscreen (Coreline Soft) was implemented. Stand-alone software correctly detected 70.4% (19/27; 95% CI 49.8% to 86.2%) confirmed cancer nodules.

*Hall et al. 2022,*[27] *UK: Veolity (MeVis)*

Hall *et al.*'s study[27] comprised all 770 LDCT from the London-based LSUT trial. In a MRMC study, two radiographers without prior experience in thoracic CT reporting independently read all CT images with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm). At the 5-mm threshold, the per-subject sensitivity for confirmed cancers was 77.4% (24/31) and 93.8% (30/32) for AI-assisted radiographer 1 and 2, respectively. Specificity and false-positive rate were not reported.

- Mixed population (one study)

*Wan et al. 2020,*[58] *Taiwan: ClearRead CT (Riverain Technologies)*

Wan *et al.*[58] performed a retrospective analysis of 75 pathology-proven nodules (≤ 2 cm: benign, *n* = 28; malignant, *n* = 47) in 50 patients from hospitals in Taiwan. The study reported a sensitivity of 93.6% (44/47; 95% CI 82.5% to 98.7%) for detecting of malignant nodules by stand-alone AI. The specificity was 39.3% (11/28; 95% CI 21.5% to 59.4%).

## Nodule type determination

### *Accuracy for nodule type determination*

Two studies[33,66] evaluated the accuracy of stand-alone AI-based software (Veye Chest, Aidence) to determine nodule type. The indication for the chest CT scan was mixed in both studies. The overall accuracy of the composition algorithm for distinguishing subsolid from solid nodules was 94.2–95.0% (*Table 63*). Additional information from a report submitted by Aidence was considered confidential and was removed from this section.

a. Non-comparative results (two studies)

*Mixed population: Veye Chest (Aidence) (two studies)*

Both studies used the software Veye Chest from Aidence in stand-alone mode and compared the findings with a reference standard of consensus reading of two radiologists, with discrepancies resolved by a third radiologist (majority consensus).

Murchison *et al.*[33] used two composition classes (solid or subsolid) and found that the sensitivity and specificity of the Veye Chest software to determine the composition of solid nodules was 98.8% and 68.4%, respectively (see *Table 63*). Accordingly, the sensitivity and specificity to determine the composition of subsolid nodules was 68.4% and 98.8%, respectively. The overall accuracy for determining the composition (solid or subsolid) of a pulmonary nodule was 94.2% (360/382), and the kappa was 0.77.

Martins Jarnalo *et al.*[66] stated that the agreement on classification between the software results and the reference standard was 95%; two cases were determined solid by Veye Chest software and subsolid by the radiologists, whereas another two were determined solid by the software and mixed solid/subsolid by the radiologists. Using three composition classes (solid, subsolid, mixture of both) the sensitivity and specificity of Veye Chest software to determine the composition of solid nodules was 100.0% and 73.3% and to determine the composition of subsolid nodules was 84.6% and 100.0%, respectively (see *Table 63*). The composition of the two mixed (solid and subsolid) nodules could not be correctly detected by the software as its composition algorithm can only allocate one composition class (solid or subsolid) to a nodule.

**TABLE 63** Accuracy of stand-alone software to determine nodule type (two studies)

| Reference and country | Target population/ nodule characteristics | Reference standard | Nodule type to be determined | Sensitivity, % | Specificity, % | TP | FP | FN | TN |
|---|---|---|---|---|---|---|---|---|---|
| ***Veye Chest (Aidence): stand-alone mode*** | | | | | | | | | |
| Martins Jarnalo et *al.* 2021,[66] the Netherlands | Mixed indication; 65 solid, 13 subsolid, 2 mixture of solid and subsolid, 4–30 mm | Consensus reading of two radiologists, with discrepancies resolved by a third radiologist | Solid | 100.0 | 73.3 | 65 | 4 | 0 | 11 |
| | | | Subsolid | 84.6 | 100.0 | 11 | 0 | 2 | 67 |
| | | | Mixture solid/ subsolid | 0 | 100.0 | 0 | 0 | 2 | 78 |
| Murchison *et al.* 2022,[33] UK | Mixed indication; 325 solid, 57 subsolid; 3–30 mm? | Consensus reading of two radiologists, with discrepancies resolved by a third radiologist | Solid | 98.8 | 68.4 | 321 | 18 | 4 | 39 |
| | | | Subsolid | 68.4 | 98.8 | 39 | 4 | 18 | 321 |

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

## Whole read

### *Accuracy for lung cancer detection based on whole read*

a. Non-comparative results (one study)

*Screening population: AVIEW Lungscreen (one study)*

A second analysis of the K-LUCAS project by Hwang *et al.*[50] comprised 10,424 consecutive participants who underwent baseline LDCT after the implementation of the AVIEW Lungscreen software. The LDCTs were assessed in clinical practice by single expert thoracic radiologists with concurrent software use. Using the Lung-RADS (version 1.1) diameter threshold of 6 mm for solid nodules and part-solid nodules, respectively, and 30 mm for non-solid nodules, the study compared the performance of average transverse and effective nodule diameters for lung cancer diagnosis within 1 year from LDCT as well as any lung cancer diagnosis after LDCT. The reference standard was based on medical record review, with 52 participants diagnosed with lung cancer within 1 year from LDCT and six participants diagnosed after 1 year from LDCT. Using the average transverse diameter (2-D measurement), the sensitivity for lung cancer within 1 year was 96.2% (50/52) and the specificity was 91.7% (9515/10,372; 95% CI 91.2% to 92.3%). Using the effective nodule diameter (based on volumetric measurement), the sensitivity for lung cancer within 1 year was also 96.2% (50/52) and the specificity was slightly lower, at 90.9% (9433/10,372; 95% CI 90.4% to 91.5%). For the detection of any lung cancer after LDCT, the average transverse diameter had a sensitivity of 91.4% (53/58) and a specificity of 91.8% (9512/10,366; 95% CI 91.2% to 92.3%). When using the effective diameter, the sensitivity was again 91.4% (53/58), with a specificity of 91.0% (9430/10,366; 95% CI 90.4% to 91.5%).

## Nodule registration and growth assessment

### *Nodule registration*

a. Non-comparative results (one study)

*Mixed population: Veye Chest (Aidence) (one study)*

Murchison *et al.*[33] included a routine cohort of current or ex-smokers and/or those with radiological evidence of pulmonary emphysema between 55 and 74 years (to mimic a screening population) who underwent chest CT for non-screening purposes at a single centre in Edinburgh (UK). Forty-six CT scans from 23 patients undergoing CT surveillance of a pulmonary nodules (baseline CT scans, $n$ = 23; follow-up CT scans, $n$ = 23) were included in the analysis of nodule registration and growth rate assessment. The study used the software Veye Chest (Aidence) in stand-alone mode for nodule registration and compared the findings with a reference standard of majority consensus (consensus reading of two radiologists, with discrepancies resolved by a third radiologist).

According to Murchison *et al.*, the total number of nodule-pairs in baseline and follow-up CT scans was 23, and all nodule pairs were successfully identified by the Veye Chest software. The sensitivity for detecting nodule pairs of the stand-alone software was 100.0% (23/23), and the average number of false-positive pairs was 0.0.[33]

### *Nodule growth assessment*

#### Stand-alone AI versus unaided reader

*Mixed population: Veye Chest (Aidence) (one study)*

The study mentioned above[33] also compared nodule growth rate assessment (relative volume difference between a nodule visible on the baseline scan and visible on a follow-up CT scan) for 23 nodule pairs between stand-alone AI and two unaided radiologists. The mean growth percentage difference was similar with readers and stand-alone software: 1.30 (95% CI 1.02 to 2.21) between radiologists and 1.35 (95% CI 1.01 to 4.99) between the stand-alone AI and radiologists, which was not significantly different. However, because of a single incorrect segmentation of the stand-alone software, the upper end of its CI is twice as high as that of readers, illustrating that visual verification of the nodule segmentation by human readers is still advised.

## Practical implications: additional results

### *Other outcomes (not prespecified in the protocol)*

**Radiologist workload reduction when using AI-based software as pre-screen (one study)**
*Screening population: AVIEW LCS (Coreline Soft) (one study)*

One study[32] was identified that reported on the simulated radiologist workload reduction when stand-alone AI-based software would be used as pre-screen to rule out CT images with no or only benign nodules. Lancaster *et al.* included 283 patients undergoing baseline screening between February 2017 and February 2018 in the Moscow Lung Cancer Screening programme with at least one solid nodule present on ultra-LDCT images. They used the stand-alone software AVIEW LCS from Coreline Soft to automatically detect, measure and classify nodules based on a volume threshold of 100 mm$^3$ in accordance with the NELSONplus/EUPS protocol.[91,92] Lancaster *et al.* simulated the use of stand-alone AI software as pre-screen in a general lung cancer screening population based on the results of this study. When radiologists would only read CT scans where nodules ≥ 100 mm$^3$ are present in order to determine the follow-up strategy, instead of reading all scans, a workload reduction between 77.4% (lower limit) and 86.7% (upper limit) could be expected.

## Impact on patient management: additional results

### *Characteristics of detected nodules*

a. Non-comparative results (three studies)

Three studies reported characteristics of nodules detected by software-assisted readers[50,52] and stand-alone software,[66] respectively, without a comparator.

*Mixed population: Veye Chest (Aidence) (one study)*

Martins Jarnalo *et al.*[66] randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital. The average size of all 130 (true positive, *n* = 80; false positive, *n* = 50) nodules between 4 and 30 mm detected by stand-alone software (Veye Chest, Aidence) was 9.0 mm (SD 7.1 mm); 85% were solid, 14% were subsolid and 1% were mixed solid/subsolid.

*Screening population: AVIEW Lungscreen (Coreline Soft) (two studies)*

The two prospective studies by Hwang *et al.*[50,52] are both based on the K-LUCAS project and possibly have overlapping patients and CT images. The software AVIEW Lungscreen from Coreline Soft was used in assisted 2nd-read mode by experienced thoracic radiologists to detect nodules. The characteristics (type, size, Lung-RADS category) of all nodules as well as the risk-dominant nodules detected with software use in screening practice are reported in *Table 64*.

### *Characteristics of true-positive nodules*

a. Non-comparative results (four studies)

Four studies reported the characteristics of true-positive nodules detected by stand-alone software,[51,66] by software-assisted readers[56] and/or by the reference standard.[32,56,66]

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

Hwang *et al.*[51] included 4666 participants who had undergone lung cancer screening as part of the K-LUCAS project after the implementation of the software AVIEW Lungscreen (Coreline Soft). Stand-alone software correctly detected 2147 nodules, of which 96.6% (2075/2147) were solid, 1.6% (34/2147) were part-solid and 1.8% (38/2147) were ground-glass nodules. The Lung-RADS categories of the correctly detected nodules are reported in *Table 64*.

*Screening population: ClearRead CT (Riverain Technologies) (one study)*

Singh *et al.*[56] included 150 patients who underwent LDCT of the chest as part of the NLST: the first 125 patients with subsolid nodules (154 part-solid or 156 ground-glass nodules between 6 and 30 mm) and the first 25 patients with no nodules detected. As part of a reader study, two experienced chest radiologists sequentially interpreted of the unprocessed CT images alone and then together with the vessel-suppressed (ClearRead CT, Riverain Technologies) CT images without washout period. According to the reference standard of consensus expert reading with a third radiologist resolving discrepancies, the average diameter of the risk-dominant part-solid nodules was 15.7 ± 7.0 mm and 12.7 ± 5.0 mm for the risk-dominant ground-glass nodules. The average size of nodules correctly identified by the readers on vessel-suppressed CT images was 15 ± 7 mm for part-solid nodules and 12 ± 5 mm for ground-glass nodules.

*Mixed population: Veye Chest (Aidence) (one study)*

Martins Jarnalo *et al.*[66] randomly selected from 145 different patients 145 chest CT scans that were performed for various indications at a single Dutch hospital. Ninety-one nodules with sizes between 4 and 30 mm were detected by the reference standard (consensus reading of an experienced chest radiologist and a resident radiologist, with discrepancies resolved by a third experienced chest radiologist). The mean nodule size was 7.0 mm (SD 4.1 mm); 73 (80%) nodules were solid, 16 (18%) were subsolid, and two (2%) were a mixture of solid and subsolid (see *Appendix 5*, *Table 55*). The 80 nodules correctly detected by stand-alone software (Veye Chest, Aidence) had an average size of 7.3 mm (SD 3.8 mm); 81% were solid, 16% were subsolid and 3% were a mixture of both.

*Screening population: reference standard only (one study)*

Lancaster *et al.*[32] included 283 patients undergoing baseline screening between February 2017 and February 2018 in the Moscow Lung Cancer Screening programme with at least one solid nodule present on ultra-LDCT images. According to the

**TABLE 64** Characteristics of correctly detected and missed nodules of stand-alone software in a consecutive screening population in the Republic of Korea[51]

| | Stand-alone software | |
|---|---|---|
| **Lung-RADS category** | **Correctly detected** | **Missed** |
| Total | 2,147 | 2,133 |
| Solid | 96.6% (2,075/2,147) | 91.7% (1,957/2,133) |
| Part-solid | 1.6% (34/2,147) | 1.7% (36/2,133) |
| Ground glass | 1.8% (38/2,147) | 6.6% (140/2,133) |
| Lung-RADS 2 | 86.5% (1,857/2,147) | 92.6% (1,975/2,133) |
| Lung-RADS 3 | 8.2% (1,75/2,147) | 4.6% (98/2,133) |
| Lung-RADS 4A | 3.4% (73/2,147) | 1.5% (33/2,133) |
| Lung-RADS 4B | 1.1% (24/2,147) | 0.6% (14/2,133) |
| Lung-RADS 4X | 0.8% (18/2,147) | 0.6% (13/2,133) |
| Confirmed cancer nodules | 1.3% (27/2,147) | 0.4% (8/2,133) |

consensus read of three experienced radiologists and an experienced IT technologist, 71% of the 283 risk-dominant solid nodules were < 100 mm$^3$ and 29% were ≥ 100 mm$^3$.

## Characteristics of false-positive nodules

*Incidental population: AI-RAD Companion Chest (Siemens Healthineers) (one study)*

The study by Rückel *et al.*[49] comprised 105 consecutive patients who received a whole-body CT scan in the emergency department of a single German hospital. Nodules were detected retrospectively by stand-alone software (AI-RAD Companion Chest CT prototype, Siemens Healthineers) and compared with the original radiologist report (17% of CT scans were originally reported by a board-certified radiologist alone, and the other 83% CT scans were commonly reported by a radiology resident and a board-certified radiologist). Of 81 additional nodules detected by the stand-alone software, three were true positive. The remaining 78 false-positive nodules were classed as trauma-associated (27%), scarred/post-inflammatory (38%), perifissural lymph nodes (6%), granuloma (6%) or not able to be confirmed visually (22%).

*Screening population: AI-RAD Companion CT Chest (Siemens Healthineers) (one study)*

Chamberlin *et al.*[48] included a random 117 patients who underwent LDCT for lung cancer screening at a single US hospital and evaluated the stand-alone performance of an AI-RAD Companion Chest CT prototype (Siemens Healthineers) to detect nodules > 6 mm. The software detected 56 false-positive nodules out of a total of 222 detected nodules. False positives were identified as atelectasis (23%), extrapleural fat (16%), infection (7%), protruding osteophytes from thoracic vertebral bodies (7%), bowel (7%), blood vessel (7%), pleura (5%), rib (4%), hilum (4%), scarring (2%) and perifissural lymph nodes (2%). Nine false positives (16%) were uncategorisable by the panel of radiologists.

*Mixed population: AI-RAD Companion Chest CT (Siemens Healthineers) (one study)*

Abadia *et al.*[47] included 103 patients with at least one lung condition and one suspicious lung nodule (≥ 4 mm) on radiology report and 40 patients with one lung condition and no lung nodule on radiology report from random LDCT images taken at a single US hospital for various reasons. The percentage of false-positive nodules detected by the AI-RAD Companion CT Chest prototype (Siemens Healthineers) was 8.6%, with a median size of 10.0 mm (IQR 7.5–17.2 mm). If the nodule was near a blood vessel, an overestimation of nodule size was occasionally observed. A few false positives were also caused by incorrect lung segmentation.

*Mixed population: Veye Chest (Aidence) (one study)*

Martins Jarnalo *et al.*[66] randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital. There were 50 false-positive nodules detected by stand-alone software (Veye Chest, Aidence) with an average size of 11.8 mm (SD 10.0 mm); 90% were solid and 10% were subsolid (*Table 55*). The average size of the false-positive findings was larger than the size of the true-positive nodules (7.3 ± 3.8 mm). Nineteen (38%) false-positive nodules showed considerable atelectasis, 12 (24%) were found to be fibrosis and 10 (20%) were not rounded. The atelectasis and fibrosis cases also had a non-round shape. The remaining nine (18%) cases were found to be false positive for various reasons, for example a gland, bronchiectasis or a large consolidation.

## Characteristics of false-negative (missed) nodules

a. Non-comparative results (five studies)

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

Hwang *et al.*[51] included 4666 participants who had undergone lung cancer screening as part of the K-LUCAS project after the implementation of the software AVIEW Lungscreen. Stand-alone software nodule detection results were available in 3972 (85.1%) of participants. Out of 2133 nodules missed by the stand-alone software, 91.7% (1957/2133) were solid, 1.7% (36/2133) were part-solid and 6.6% (140/2133) were ground-glass nodules. The Lung-RADS categories of missed nodules are reported in *Table 64*. Around 0.4% (8/2133) of missed nodules were confirmed cancer nodules.

*Screening population: Veolity (MeVis) (one study)*

The study by Hall *et al.*[27] was performed in London (UK) and is a substudy of the LSUT trial. It comprised all 770 patients who received LDCT for lung cancer screening. In a MRMC study, two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCT with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm). For radiographer 1 and radiographer 2, 14.6% (7/48) and 4.9% (2/41) of missed nodules, respectively, were malignant.

*Screening population: ClearRead CT (Riverain Technologies) (one study)*

Singh *et al.*[56] included 150 patients who underwent LDCT of the chest as part of the NLST: the first 125 patients with subsolid nodules (154 part-solid or 156 ground-glass nodules between 6 and 30 mm) and the first 25 patients with no nodules detected. As part of a MRMC study, two experienced chest radiologists sequentially interpreted the unprocessed CT images alone and then together with the vessel-suppressed (ClearRead CT, Riverain Technologies) CT image without washout period. The average size of nodules missed by the readers on vessel-suppressed images was 9 ± 2 mm for ground-glass nodules and 8 ± 2 mm for part-solid nodules.

*Mixed population: ClearRead CT (Riverain Technologies) (one study)*

Wan *et al.*[58] included LDCT images from 50 Taiwanese patients with mixed indications whose nodule(s) were subsequently excised. Of 75 nodules ≤ 2 cm, the stand-alone software (ClearRead CT, Riverain Technologies) missed 14 : 11 were benign and three were malignant (one adenocarcinoma, one minimally invasive adenocarcinoma and one adenocarcinoma in situ, measuring 5.7, 6.4 and 6.8 mm in diameter, respectively). All three malignant nodules were ground-glass nodules. Of the 11 missed benign nodules, seven were ground-glass nodules, two were solid and two were part-solid. The stand-alone software ignored three (6.4%) of the 47 malignant nodules and 11 (39.3%) of the 28 benign lesions, with a statistically significant difference ($p$ = 0.001).

*Mixed population: Veye Chest (Aidence) (one study)*

Martins Jarnalo *et al.*[66] randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital. The nodules missed by stand-alone software (Veye Chest, Aidence) were an average size of 6.7 mm (SD 6.1 mm). Eight missed nodules were solid with a size of 4 mm, three were solid/calcified with a size of 4 mm and the remaining three were subsolid (4 mm, 18 mm and 20 mm).

### *Number of people undergoing computed tomography surveillance*

a. Non-comparative results (three studies)

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

The study by Hwang *et al.*[52] comprised 3353 consecutive CT images from the K-LUCAS lung cancer screening project in the Republic of Korea. Based on the original reading by single experienced thoracic radiologist with concurrent use of the AVIEW Lungscreen (Coreline Soft) software, 16.0% (535/3353) were classed as Lung-RADS category 3 or 4A and 21.6% (723/3353) were classed as 'intermediate' according to NELSON criteria, respectively.

*Screening population: Veolity (MeVis) (one study)*

The study by Hall *et al.*[27] comprised all 770 patients from the UK-based LSUT trial who received LDCT for lung cancer screening. In a MRMC study, two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCTs with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm). The study also reports the management decisions of the original unaided readers (single expert thoracic radiologists, with 5% of CT images checked by a second radiologist): 17.3% (133/770) of people were referred for CT surveillance, among whom eight people were later discounted after comparison with previous imaging, leaving 16.2% (125/770) receiving CT surveillance.

*Symptomatic population: InferRead CT Lung (Infervision) (one study)*

Kozuka *et al.*[59] randomly selected 120 chest CT images from people with suspected lung cancer who underwent CT examination at a single hospital. Of 743 nodules ≥ 3 mm that were detected by the reference standard (majority reading of three experienced radiologists), 92.5% (687/743) were followed up as nodules suspected benign.

### *Number of people having biopsy or excision*

a. Non-comparative results (three studies)

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

The study by Hwang *et al.*[52] included 3353 consecutive CT images from the K-LUCAS project in the Republic of Korea. In the original reading by single experienced thoracic radiologist with concurrent use of the AVIEW Lungscreen (Coreline Soft) software, 4.1% (137/3353) were positive on the narrow definition of Lung-RADS (i.e. Lung-RADS category 4B or 4X) and 1.6% (52/3353) were positive according to NELSON criteria.

*Screening population: Veolity (MeVis) (one study)*

The study by Hall *et al.*[27] was performed in London (UK) and is a substudy of the LSUT trial. It included all 770 patients who received LDCT for lung cancer screening. In a reader study, two radiographers without prior experience in thoracic CT reporting independently read all 770 LDCT with concurrent software use (Veolity, MeVis) and reported on the presence of clinically significant nodules (≥ 5 mm). The study reports the management decisions of the original unaided readers (single expert thoracic radiologists with 5% of CT images checked by a second radiologist); 3.9% (30/770) were directly referred to MDT because of 'suspicious nodules'.

*Symptomatic population: InferRead CT Lung (Infervision) (one study)*

Kozuka *et al.*[59] randomly selected 120 chest CT images from cases of suspected lung cancer in patients aged ≥ 20 years who underwent CT examination at a single hospital in Japan between November and December 2018. Of all 743 nodules ≥ 3 mm that were detected by the reference standard (majority reading of three experienced radiologists), 12 (1.6%) were diagnosed as malignant and 44 (5.9%) were followed up as nodules of suspected lung cancer.

### *Other outcomes (not prespecified in the protocol)*

#### Positivity rate (Lung-RADS category ≥ 3) (three studies)

Three studies based on consecutive participants from the K-LUCAS project (with possibly overlapping populations) reported on the positivity rate (proportion of people with Lung-RADS category ≥ 3) of LDCT images taken and assessed in screening practice with and without the use of the AVIEW Lungscreen software (Coreline Soft).[50–52] The only comparative study[51] found no significant differences in the positivity rate before and after software implementation when nodules were measured on transverse planes. With software use, the measurement of nodule diameter on maximum orthogonal planes or any maximum planes significantly increased the positivity rate compared with measurement on transverse planes.

a. Comparative results: reader with and without software (one study)

*Screening population: AVIEW Lungscreen (Coreline Soft) (one study)*

In a before-and-after study, Hwang *et al.*[51] included 6487 consecutive participants of the K-LUCAS project: 1821 participants were screened before the implementation of the AVIEW Lungscreen software, and 4666 participants were screened after. The LDCT images were read by single experienced chest radiologists in clinical practice, and patients with Lung-RADS category ≥ 3 were classed as positive and referred for additional follow-up CTs or diagnostic procedures. The study found that, when nodules were measured on transverse planes, the per-participant positive rates did not significantly differ between LDCT images analysed before software implementation (9.9%, 180/1821) and images interpreted after software implementation (11.0%, 511/4666; $p = 0.211$). With software use, the per-participant positive rate was significantly increased though when nodules were measured on maximum orthogonal planes (14.1%, 657/4666; $p < 0.001$) or any maximum planes (17.4%, 812/4666; $p < 0.001$) compared with measurement on transverse planes.

b. Non-comparative results (two studies)

*Screening population: AVIEW Lungscreen (Coreline Soft) (two studies)*

In 10,424 LDCT images that were interpreted using concurrent software, the positivity rate was 8.7% (907/10,424) when using the average transverse diameter and 9.5% (989/10,424) when using the effective diameter.[50] Discrepancies in screening positivity between average transverse diameters and effective diameters occurred in 214 (2.1%) of participants.

The third analysis based on the K-LUCAS project comprised 3353 consecutive LDCT images that were read in screening practice by 20 different expert chest radiologists with concurrent software use. Using Lung-RADS, the positivity rate was 20.0% (672/3353).[52]

# **Appendix 7** Growth model and its development process

## Introduction

Assessing the impact of AI assistance during CT surveillance necessitates modelling the pathways that people with lung nodules would take between repeated CT scans based on the findings of the earlier CT scan. During the time between CT scans, the nodule may grow, and this needs to be considered when assessing the impact of AI assistance at follow-up CT scans. Thus, we need to know the natural history of lung cancer in the form of growth in malignant nodules and quantify it using a malignant nodule growth model. To facilitate this, we first identified studies that include such models and then obtained information from relevant studies to develop a growth model that can be incorporated into our decision modelling as described below.

## Methods

We undertook a targeted search for studies that explicitly modelled disease progression of lung cancer based on tumour growth. We searched electronic databases (e.g. MEDLINE and EMBASE) for potentially relevant studies. The titles and abstracts of records were screened by Peter Auguste and Hesam Ghiasvand. Articles that were considered appropriate were read in full. No quality appraisal or data extraction was undertaken. Full details of the search strategy can be found in *Appendix 3*.

## Results

We screened 750 titles and abstracts, of which 15 were potentially relevant and were read in full. From these, four studies[71,85,93,94] that modelled disease progression based on tumour growth were considered useful and discussed below. Details of these studies can be found in *Appendix 7*, *Table 65*.

The underlying growth model used by Edelsberg *et al.* and Sutton *et al.* was obtained from Gould *et al.*[93] Briefly, Gould *et al.* undertook an economic analysis that compared management strategies (including or excluding FDG-PET) for the diagnosis of pulmonary nodules by using a model with two components: a decision tree and a Markov model. The Markov component was used to model and estimate the long-term costs and health outcomes associated with managing people with benign and malignant lung nodules. Before clinical presentation, people with malignant lung nodules who were managed through watchful waiting were at risk of progressing from local → regional → distant/metastatic lung cancer during the observation period. At the time of diagnosis/clinical presentation, people would move/progress from a pre-clinical health state to a clinical health state (benign, local or regional).

To determine the probability of disease progression during watchful waiting, Gould *et al.* used information obtained from Steele and Buell.[79] In this study, data were collected from the Veterans Administration-Armed Forces Cooperative Study on Asymptomatic Solitary Nodules involving Veterans Administration across 13 participating military hospitals. The growth rate of lung nodules was based on the VDT measured in 67 cases of people with asymptomatic nodules measuring < 6 cm. Nodule size was routinely collected using chest films based on incidental findings.

Edelsberg *et al.* assessed the cost-effectiveness of autoantibody test compared with CT surveillance alone in people with an indeterminate risk of lung cancer. The authors fitted an exponential model to the observed data from Steele and Buell[79] to derive monthly transition probabilities. Sutton *et al.*[85] undertook a similar economic analysis, which estimated the cost-effectiveness of an autoantibody test, EarlyCDT-Lung, in the diagnosis of lung cancer among people with an indeterminate pulmonary nodule as an adjunct to CT surveillance compared with CT surveillance alone. The authors used the same approach to derive monthly transition probabilities. We noted similarities and differences in the assumptions made with regard to the growth models: Gould *et al.*[93] assumed that if there was no evidence of growth, nodules were considered benign, and transition probabilities for progressing from local to regional and from regional to distant disease were the same. Edelsberg *et al.*[94] assumed that after three CT scans with no evidence of the nodule doubling, the nodule was considered benign. The authors further assumed that malignant nodules not diagnosed at model entry increased in size and progressed during CT surveillance. Sutton *et al.*[85] assumed that the transition probability of progressing from local to regional is the same as that of progressing from regional to distant disease, and people undergoing CT surveillance all received three CT scans.

In general, these assumptions made were considered feasible; however, we query the usefulness of the underlying study[79] to model our growth model. We considered that this study may not be generalisable to our subpopulations of interest as study participants were male and all had lung nodules < 6 cm. Additionally, the study is dated, and the characteristics of patients are likely to be different from those of a more contemporary cohort. Furthermore, the techniques used to model the growth have improved based on the knowledge about how lung nodules grow. It is understood that the growth of lung nodules is better modelled using a Gompertz function than an exponential. Moreover, evidence of VDT was collected using routine chest films in the original study, but this is now done through CT scans.

Given these limitations, other alternative studies with a more contemporary cohort were pursued. One such study was undertaken by Treskova *et al.* These authors investigated the effects of the eligibility criteria and nodule management on the benefits, harms and cost-effectiveness of lung cancer screening with LDCT by using a microsimulation model. The model was populated with 10% of the German population aged ≥ 40. Data on smoking behaviour were obtained from the German Health Update (GEDA) survey (years 2009–12), and the demographic structure of 2012 was obtained from the German statistical office. The growth model also uses the data from US NLST and the NELSON lung cancer screening trials. The NLST algorithm assessed the nodule diameter, and depending on the size it recommends three categories of screening results: negative, positive intermediate, and positive. Conversely, NELSON assessed the nodule volume, and, depending on an individual's result, that person could be recommended to undergo further screening (people with negative results), a follow-up examination (people with indeterminate results) or an immediate diagnostic work-up (people with positive results).

Treskova *et al*. assumed that the threshold tumour volumes at the stages of nodal involvement, and distant disease and clinical diagnoses were randomly drawn from log-normal distributions. Lung cancer progression was described via tumour growth, lymph nodes involvement and metastases, and the growth of malignant nodules is defined by a Gompertz function.[71] The model included a natural history of a biological two-stage clonal expansion of the disease incorporating the nodule growth (in terms of the rate and time). The two-stage clonal expansion model considers the age of individuals at the first presentation of a malignant lung nodule, which was categorised as adenocarcinoma, large-cell carcinoma, small-cell carcinoma and squamous cell carcinoma.

Researchers identified the harms as incurred costs, false positives and overdiagnosis due to a lung cancer screening. Benefits included reduction in mortality, the number of deaths averted due to earlier detection of lung cancer, and subsequently the life-years gained. They assumed that there was a balance between the harms and benefits that can result in efficiency. They adopted a model that traced the efficiency and effectiveness of the lung cancer screening programme from the initial development of the nodule through to its turning into lung cancer. The screening module of their model included eligibility assessment, screening detection, nodule management (including follow-up), diagnostic work-up and lung cancer survival. This created a screening schedule for each person based on US NLST and the NELSON trials.

Treskova *et al.*[71] used VDT, an indicator used in the BTS guidelines[12] for managing people with lung nodules. The authors were transparent in their modelling methodology by providing details of their approaches, including their functions, parameters and assumptions. Given the advantages of this study over others identified, we used this as the basis of our growth model for solid malignant nodules.

## Growth/progression of malignant nodules

To the simulated nodule diameter measurements at baseline CT scan, we applied growth curves and simulated how nodules grew over 2 years of CT surveillance for solid nodules, and 4 years of CT surveillance for subsolid nodules. Growth curves were simulated for the reference standard, AI-assisted radiologist reading of CT scan and unaided radiologist reading.

We used the growth model developed by Treskova *et al.* to track malignant nodules' growth over time from baseline. Treskova *et al.* suggest a Gompertz function with a log-normal distribution for the scale and shape parameters of a malignant nodule's growth over a person's lifetime. In the proposed growth model, the disease progression is characterised by the nodule's volume, location, and metastatic probability. They assumed that if a person's threshold volume exceeds the calculated maximum expected (*Vmax*), the corresponding cancer stage will not be reached during that patient's lifetime.

A spherical volume measurement for computing the volume of the nodule was provided for four histological types along with threshold values. We selected the threshold parameters for adenocarcinoma to simulate malignant tumour growth. This histological class was chosen because it accounts for the majority (87%) of the lung cancers diagnosed in the UK.

Nodule volume was calculated from the baseline nodule diameter. Then, the growth function was applied to calculate nodule volume at subsequent time points. Nodule diameter was calculated by rearranging the formula for the sphere volume. Using the newly calculated diameters, VDT was calculated for each person with a lung nodule that showed no clear features of being benign.

The following formulae were used for both solid and subsolid nodules (only the growth function differs between solid and subsolid nodules):

$$sphere\ volume = \frac{\pi}{6} \times (Diameter)^3 \tag{10}$$

$$sphere\ diameter = \sqrt[3]{\frac{6 \times (sphere\ volume)}{\pi}} \tag{11}$$

$$volume\ doubling\ time\ (VDT) = time \times \frac{\log(2)}{\log(\frac{sphere\ volume\ at\ time_{t=i+1}}{sphere\ volume\ at\ time_{t=i}})} \tag{12}$$

Solid nodules:

$$Gompertz\ growth\ function = Volume_{\max} \times \frac{Volume_{t=0}}{Volume_{\max}}^{-time \times alpha} \tag{13}$$

Where $Volume_{\max} = 141137.17$ and $alpha \sim Normal\ distribution(-7.765, 0.5504)$ .

Subsolid nodules:

$$Linear\ growth\ function = Volume_{t=i} + 2 \times \frac{time}{alpha} \tag{14}$$

Where $alpha \sim \log(Normal\ distribution(3.6316,\ 1.5279))$.

# Appendix 8 Methods of simulation

Given that improved measurement consistency is one of the main purported advantages of AI-assisted image analysis, the EAG carried out two linked simulations to estimate the potential impact of different measurement consistency (magnitude of random measurement errors) and measurement accuracy (systematic bias) between AI-assisted reading and unaided radiologist reading on subsequent nodule management according to the BTS guidelines,[12] which then links to patient outcomes and costs through the EAG's model. The first simulation (baseline measurement simulation) was carried out to evaluate the potential impact of differential measurement performance on the classification of patients/nodules into appropriate risk categories based on nodule sizes measured by either AI-assisted reading or unaided radiologists. The second simulation (nodule growth monitoring simulation) was conducted to evaluate the potential impact of differential measurement performance on the classification of patients/nodules into appropriate risk groups based on estimated VDT using nodule size/volume measurements made at two CT scans in the context of surveillance, taking into account nodule growth between the scans. The procedures of the two simulations are described in detail in the following two sections.

## Simulation for nodule sizes at baseline (baseline measurement simulation)

We first generated a cohort of risk-dominant nodules (the largest nodule or the one most suspicious of being malignant) in people with at least one 'true' nodule (≥ 3 mm and ≤ 30 mm) at the time of their initial (baseline) CT scan. The size distribution of the cohort of nodules was based on data reported in a large population screening study[49] and served as the reference standard. We generated the values from a log-normal distribution that matched the reported median and IQR. For ease of interpretation, we also conceptualised nodule sizes estimated from this cohort as consensus reading, which frequently serves as the reference standard in studies of nodule detection and measurement and refer to these reference standard nodule sizes as being obtained by reader 1 (R1) as a shorthand. Acknowledging that the reference standard established by consensus is itself subject to limitations associated with measurement by human, we additionally created a set of nodule sizes that reflect the unobservable 'true' nodule sizes (denoted as reader 0, R0; details described below), based on which the growth of nodules between consecutive CT scans is estimated in the subsequent nodule growth monitoring simulation using the growth model described in *Appendix 7*, *Table 65*, and text.

Based on R1, we then created three sets of nodule size estimates representing the nodule sizes that would be obtained by stand-alone AI (designated as reader 2, R2), a radiologist with concurrent AI (reader 3, R3) and an unaided radiologist (reader 4, R4), respectively, if they were to measure the same cohort of nodules. Parameters for these sets of nodule size estimates (including the median and IQR of the true nodule sizes and the proportion of solid and subsolid nodules for R1, and the systematic bias and random errors of measurements for R2, R3 and R4) were determined using data from studies included in our test accuracy review or from additional studies identified from the literature, with different values used for different population of interest where data were available.

By using the simulated distribution of measured nodule sizes between R1, R3 and R4, we can estimate the proportion of nodules correctly or misclassified into different management pathways by concurrent AI (R3) or unaided radiologist (R4) compared with perfect classification (R1) according to the size threshold specified in the BTS guidelines (< 5, ≥ 5 to < 6, ≥ 6 to < 8, or ≥ 8 mm for solid nodules; < 5 or ≥ 5 mm for subsolid nodules).[12] Based on size-specific cancer risk estimated from the NELSON lung cancer screening trial,[4] we could then estimate the proportion of true malignant nodules that go through individual nodule management pathways (e.g. discharge, surveillance, definitive management) and subsequently are detected or missed. These outputs could then be used as parameter inputs for our model to compare downstream impacts.

### *Reader 1: consensus reading (reference standard)*

Data from Hwang *et al.*[50] were used as a reference for R1 for the *screening* population as this study included a large (*n* = 10,424) consecutive screening population and reported the distribution of nodules sizes separately for solid, part-solid and non-solid nodules. The median (IQR) average transverse diameter was 3.6 mm (1.9 mm) for solid nodules,

**TABLE 65** Characteristics of studies that included a growth model

| Authors, year | Type of study | Aim(s)/objective(s) | Data underpinning growth model | Assumptions | Pros | Cons |
|---|---|---|---|---|---|---|
| Gould *et al.*, 2003[93] | Economic evaluation | To evaluate the cost-effectiveness of strategies for pulmonary nodule diagnosis and to specifically compare strategies that did and did not include FDG-PET | Data obtained from the study undertaken by Steele *et al.* (1963) – male veterans administration armed forces cooperative study on asymptomatic pulmonary nodules | If there was no evidence of growth observed by 24 months, it was assumed that the nodule was benign<br>Assumed that pulmonary nodules measured 2 cm in diameter<br>12.5% of people with malignant nodules had regional lymph node involvement<br>Monthly probabilities for disease progression depended on VDT, a measure of tumour growth<br>Tumour starts from a single cell that measures 10 microns in diameter that doubles in volume at a constant rate<br>Death occurs after 40 doublings for a tumour of size 10cm<br>Untreated lung cancer progresses from local → regional → distant → dead<br>Transition probabilities for progressing from local → regional → distant disease are equal<br>Growth would be detected when the nodule doubles once in volume | Used in several economic analyses<br>Doubling time by cell type (squamous cell, adenocarcinoma, bronchiolar, adenosquamous and undifferentiated) | Based on dated information that included males only<br>Appears to be solitary nodules only<br>Unclear about definitions used for lung nodules (TP, TN, FP, FN)<br>Historical data in males with asymptomatic nodules measuring < 6cm. Evidence of VDT is collected using routine chest films |
| Sutton *et al.*, 2020[85] | Economic evaluation | *To examine the cost-effectiveness of autoantibody test (AABT), EarlyCDT–Lung, in the diagnosis of lung cancer amongst patients with IPNs applied in the addition to CT surveillance, compared to CT surveillance alone as specified in the British Thoracic Society guidelines in which patients are offered surveillance through repeat CT scanning* | Progression rates in people with undiagnosed malignant nodules were based on observed VDT obtained from Gould *et al.*, which were originally obtained from Steele *et al.* (1963). Exponential model was fitted to the observed data to derive monthly transition probabilities | It appears that malignant lung nodules were initially diagnosed at local (87.5%) or regional stage (12.5%) People undergoing surveillance received CT scans at 3, 12 and 24 months.<br>People with a negative test continued to undergo surveillance<br>Probability is the same for progression from undiagnosed local to regional disease and from regional to distant disease<br>Not explicitly stated but once locally diagnosed there is no progression to distant disease. However, if diagnosed regional there is a possibility of progressing to distant disease<br>100% compliance with CT surveillance | The model includes both detection and treatment phases<br>Included a probability associated with growth of a benign nodule at the first month and subsequent probability of growth<br>Transition probabilities reported for the natural history model | Unclear about stage shift<br>Not revealed natural history for the growth rate<br>Not including VDT for measuring the growth of the lung nodules using information obtained from Gould *et al.* study, which is a dated database (1973) |

TABLE 65 Characteristics of studies that included a growth model (*continued*)

| Authors, year | Type of study | Aim(s)/objective(s) | Data underpinning growth model | Assumptions | Pros | Cons |
|---|---|---|---|---|---|---|
| Edelsberg *et al.*, 2018[94] | Cost-effectiveness analysis | To assess if the cost-effectiveness of autoantibody test compared with CT surveillance alone could improve outcomes for people at intermediate risk of lung cancer | Based on information reported in Gould *et al.* (2003) | People have incidentally detected nodules that measure between 8 and 30 mm and have an estimated 5–60% risk of lung cancer<br>After three CT scans and there is no volume doubling, the nodule is assumed to be benign<br>Malignant nodules are diagnosed at biopsy.<br>If not diagnosed at time of model entry, then nodules were assumed to increase size and progress during the 24-month follow-up and are assumed to be diagnosed soon after CT scan following volume doubling<br>Patients whose nodules are benign who had tested positive would receive a biopsy that would confirm no malignancy | Using the VDT for identifying the lung cancer progression over time, targeting quality of life as the main outcome | Using data from Gould *et al.* study, which is related to 1973 (dated database), focused only on malignant nodules, natural history is based on the VDT, but not elaborated |
| Chen *et al.*, 2014[95] | To model the natural history of an individual from birth to lung cancer initiation, progression, detection and death | Several models (carcinogenesis, tumour growth and metastasis, and cancer detection) were used to address the research question. Our focus is on the model used to measure tumour growth | Simulation and validated using the SEER data set | Several assumptions were made for the tumour growth and metastasis modelling:<br>The primary tumour grows from a single cell, with an assumed volume of $1 \times 10^{-9}$ cm$^{3}$.<br>The growth rate λ is related to the tumour doubling time and is determined when first detected and is assumed to remain constant over time<br>Growth rate follows a gamma distribution<br>Metastases are defined as nodal or distant. Different rates for each type of metastases | Provided tumour size frequency distribution for local, regional and distant disease<br>Incorporating the smoking behaviour in the natural history<br>Yearly mean growth rate by stage and VDT by stage (days) | The study focuses on developing and validating a predicting model for lung cancer based on demographical and smoking characteristics, thus the study does not provide a clear lung nodules growth pattern over time<br>The study seems more suitable for predicting the lung cancer probability due to smoking and then for non-smoker population probably not applicable |

continued

**TABLE 65** Characteristics of studies that included a growth model (*continued*)

| Authors, year | Type of study | Aim(s)/objective(s) | Data underpinning growth model | Assumptions | Pros | Cons |
|---|---|---|---|---|---|---|
| Treskova *et al.*, 2017[71] | A stochastic modular microsimulation model that simulated individual life histories focusing on the development of lung cancer and its progression from the onset of the first malignant cell to death from lung cancer | The study aimed to investigate the effects of the eligibility criteria and nodule management on the benefits, harms and cost-effectiveness of lung screening with LDCT in a population-based setting | The model was populated with 10% of the German population aged ≥ 40 years. Data on smoking behaviour were obtained from the German Health Update (GEDA) survey (years 2009–12), and the demographic structure of 2012 was obtained from the German statistical office. The model also uses the data from US NLST and NELSON as lung cancer screening trials | The module uses the age at the onset of the first malignant cell<br>Threshold tumour volumes at the stages of nodal involvement, distant metastases and clinical diagnosis are randomly drawn from log-normal distributions<br>Threshold tumour volumes at the stages of nodal involvement, distant metastases and clinical diagnosis are randomly drawn from log-normal distributions<br>The clinical detection module determines the stage of lung cancer (I, II, III, IV) according to the TNM staging system based on the tumour volume and spread (local, nodal involvement, distant metastasis) at the age of diagnosis<br>Lung cancer survival is modelled as long-term survival, which lets the individual live until death from other causes, and short-term survival in years, which follows the Weibull distribution<br>The parameters vary over the histological classes and stages at the time of diagnosis<br>Two nodule management algorithms were designed based on those used in the NELSON and NLST trials<br>The tumour is staged according to TNM classification based on the volume and spread<br>Individuals with screen-detected lung cancer live at least if they would in the no screening scenario<br>In the screening module lung cancer, survival component alters the age of death from lung cancer for the persons with a screen-detected lung cancer at stages I and II: if they die from lung cancer in the no-screening scenario, they receive 40% probability of long-term survival in the screening scenario<br>The tumour growth rate is based Gompertz model | The natural history module contains a biological two-stage clonal expansion model and a tumour growth component and simulates a complete flow of events in the development of lung cancer<br>The model has space for smoking and its impacts<br>The probabilities of overdiagnosis, by using data from both NLST, and NELSON<br>The survival probabilities are based on the histological staging of lung cancer, size specific sensitivity of LDCT<br>Rate of cases at stage II as an earlier stage of lung cancer<br>The complication rates at work-up by the diameter of the malignant nodule and for benign nodule<br>Developing a two steps calibration: for each lung cancer type mean and SD of the log-normal distributed threshold volumes of lymph nodes involvement (regional), distant metastases (distant) and clinical diagnosis were simultaneously calibrated to fit the German UICC data on diseases stage at time of diagnosis. Second, we simultaneously calibrated the age- and cancer type-dependent malignant conversion rates and age boundaries of the survival functions (the Nelder–Mead simplex method) in R package 'FME' | The model is only focused on the screening population<br>No cost per QALYs analysis (only cost per life-year gained)<br>The total cost of screening is not included for lifetime lung cancer treatment costs and the costs for pharmaceuticals, because of partial German database in this regard The calibration has not been done for all parameters because of limitations in the data set |

**TABLE 65** Characteristics of studies that included a growth model (*continued*)

| Authors, year | Type of study | Aim(s)/objective(s) | Data underpinning growth model | Assumptions | Pros | Cons |
|---|---|---|---|---|---|---|
| | | | | | To obtain the costs for people with early-stage cancer in our model we applied ratio of costs between III and I stages used to define a base-case scenario<br>The simulated parameters for proportion of all detected cancers and by its histological stages are consistent with data from NLST. The VDT figures by either NLST or NELSON | |
| Lin *et al.*, 2012[96] | A natural history model of cancer to estimate the probability of disease-specific cure as a function of tumour size, the TVDT and disease-specific mortality reduction achievable by screening | To estimate the impact of early detection of cancer, knowledge of how quickly primary tumours grow and at what size they shed lethal metastases is critical | Model parameter estimates were based on Surveillance Epidemiology and End Results (SEER) cancer registry data sets and validated on screening trials | Primary tumour volume grows exponentially<br>The tumour has a constant TVDT<br>The 'treatment cure threshold' of cancer as the primary tumour volume at which the disease transitions from being curable to incurable, assuming standard of care following detection<br>The patient would never die from their specific disease if it was detected and treated at or before the treatment cure threshold<br>The lethal metastatic burden starts increasing at the treatment cure threshold; therefore, we are implicitly excluding metastasis that may be eradicated or controlled by systemic treatment when treated before the onset of the lethal metastatic burden<br>The lethal metastatic burden grows in proportion (*f*) to the growth of the primary tumour, and continues to grow even after the primary tumour is detected and removed<br>If the patient is not diagnosed and treated before the treatment cure threshold, the lethal metastatic burden becomes the cause of death at the maximal lethal metastatic burden<br>Disease is symptomatically detected either due to the primary tumour or the lethal metastatic burden, dependent on which presents with symptoms first | The model has been evaluated by using simulation of data from different databases<br>The model is not only for screening population, and it seems to be helpful for considering other route of diagnosis of lung cancer<br>The study has a good explained natural history-based VDT and the parameters that have been defined and explained very well<br>The model outputs have some parameters, including the distribution of tumour by size, the proportion of advancement/progression of the lung cancer cells by tumour size and survival rates | The analysis was limited to Caucasians because this is the largest ethnic group of lung cancer patients<br>The analysis was limited to males because the external validation data set from the Mayo Lung Project (described below) was limited to males only |

continued

**TABLE 65** Characteristics of studies that included a growth model (*continued*)

| Authors, year | Type of study | Aim(s)/objective(s) | Data underpinning growth model | Assumptions | Pros | Cons |
|---|---|---|---|---|---|---|
| | | | | Patients are clinically staged with advanced disease if lethal metastatic burden is detected at symptomatic detection<br>The size of the primary tumour at detection *VP* and the growth rate of tumour volume *r* are assumed to have bivariate lognormal distribution with mean (*μ1*, *μ2*), variance (*σ1*, *σ2*), and correlation coefficient *ρ*<br>The treatment cure threshold *VC* is assumed to have a Weibull distribution with shape parameter *c1* and scale parameter *c2*; and the ratio *BD*/*f* is assumed to have a Weibull distribution with shape parameter *b1* and scale parameter *b2* | | |
| Heuvelmans *et al.*, 2017[97] | Solid lung nodules found at ≥ 3 CT examinations before lung cancer diagnosis were included. Lung cancer volume (*V*) growth curves were fitted with a single exponential, expressed as *V* = *V*1exp(t/_), with *t* time from baseline (days), *V*1 estimated baseline volume ($mm^3$), and _ estimated time constant. The $R^2$ coefficient of determination was used to evaluate goodness of fit. Overall volume-doubling time for the individual lung cancer is given by _ * log(2) | To evaluate and quantify growth patterns of lung cancers detected in the Dutch-Belgian low-dose CT lung cancer screening trial (NELSON), to elucidate the development and progression of early lung cancer | Eligible sample of participants from the NELSON lung cancer screening clinical trial | The nodule growth rate has an exponential pattern | The study has a good explanation from the model and how to calculate the VDT<br>The study has used the NELSON trial database<br>The study has some findings in terms of VDT (the number of cancers by VDT groups)<br>*Figure 5* reports the VDT in days for 46 lung cancers from the NELSON trial | The study assumptions have not been stated<br>The study natural history model has not been elaborated well<br>The VDTs have not been compared with different growth models (e.g. Gompertz, linear or log-linear)<br>Growth patterns for slow-growing lung cancers were evaluated in this study. Faster growing lung cancers did not receive at least three CT scans |

CT, computed tomography; NELSON, Nederlands–Leuvens Longkanker Screenings Onderzoek; SCLC, small-cell lung cancer; SEER, The Surveillance, Epidemiology, and End Results; TVDT; tumour volume doubling time; VDT, volume doubling time.

11.9 mm (11.1 mm) for part-solid nodules and 5.8 mm (IQR 4.7 mm) for non-solid nodules. The part-solid and non-solid nodules were combined in a ratio of 4 : 5 to create the simulated subsolid nodules population. Moreover, a log-normal distribution was used to simulate nodule sizes for R1 as nodule sizes were heavily skewed.

Data reported by Kozuka *et al.*[59] were used as the input for R1 for the *symptomatic* population, as this study was the only one identified that reported nodule type and size in people suspected of having lung cancer. The median nodule size was reported as 4.7 mm. The IQR was estimated using *Table 1* of this paper and assumed to be equal between nodule types due to a lack of available data. As with the screening population, a log-normal distribution was used. The majority of nodules in this paper were solid (70%), so the median solid nodule size was assumed to be 4.7 mm. As the nodule sizes by nodule type were not presented, we made the following assumption for subsolid nodules based on the screening population.

The median nodule size was 3.6 mm for solid nodules and 8.5 mm for subsolid nodules, a factor of 2.36.

This was applied to the 4.7 mm from reported by Kozuka *et al.*,[59] resulting in an assumed median subsolid nodule size of 11.1 mm.

As we are simulating nodule sizes from the following three readers based on R1, we assume a dependency between R1 and the other readers. Therefore, the nodule sizes simulated for R2–4 were normally distributed around the R1 nodule. Other assumptions are as follows. These assumptions were the same for both the screening population, and the symptomatic populations. Only the R1 inputs differed. Furthermore, the screening and incidental populations were assumed to be equivalent in the simulation.

### *Reader 0: the unobservable 'true' nodule size*

Reader 0 was the assumed 'true' nodule size that was simulated using the values from R1. We expected consensus reading to be very close to the true size of the nodule, and so we applied a SD of 0.1 to the R1 values to allow the true size to deviate slightly from the size as measured by the reference reader.

Based on their true size (R0), we assumed that nodules had a probability of being malignant. These lung cancer probabilities were derived from Horeweg *et al.*,[4] who used 9681 non-calcified nodules detected by CT screening in 7155 participants in the screening group of the NELSON trial. For solid nodules, this was estimated to be 0.009 for nodules between 5 and < 6 mm, 0.011 for nodules between 6 and < 8 mm and 0.094 for nodules ≥ 8 mm. We also assumed that 10% of detected nodules had clear features of being benign, which would be identified by each reader without error. The 10% estimate seemed to be consistent between the symptomatic population,[59] screening population[98] and incidental population.[87]

### *Reader 2: stand-alone AI*

Although in current practice all CT scans will still be checked by a radiologist even if AI software is used for automatic nodule detection and analysis, we included the 'stand-alone AI reading' option in the simulation as these were the only data reported in some of the included studies, and it is generally recommended that, unless there are clear issues related to nodule segmentation, size/volume measurements obtained by AI should not be manually adjusted in order to preserve the consistency afforded by AI measurements.[18]

The base-case simulation for R2 was based on the discrepancies between nodule size measurements by stand-alone AI and by a panel of three radiologists as reported by Martins Jarnalo *et al.*[66] This study was chosen as it was the only identified study that reported individual measurement discrepancies of stand-alone AI compared with a reference standard for each of the 77 nodules (*Table 66*). The mean (SD) of these discrepancies was 0.234 (0.771) mm, so the mean size (mm) of R2 simulated nodules was R1 + 0.234, with an SD of 0.771, for both solid and subsolid nodules (see *Table 69*).

Scenario analysis 1 also used data by Martins Jarnalo *et al.*[66] where stand-alone AI and majority reading of three radiologists agreed on 67.5% (54/80) of measurements (same millimetre). Therefore, the mean simulated nodule size for R2 was the same as for R1, only the SD was changed so that the agreement between R1 and R2 was approximately 67.5% (see *Table 69*).

Scenario analysis 2 was based on a phantom study by Wu *et al.*,[99] in which the relative volume error of AI-based measurement (AI software C) was 0.69 (0.27, 1.35) for ground-glass nodules and 0.91 (0.49, 1.30) for solid nodules. Assuming a cubic relationship between volume and diameter, the mean (SD) simulated nodule size for solid nodules was R1 + 0.969 (0.249), and R1 + 0.884 (0.411) for subsolid nodules (see *Table 69*).

### *Reader 3: concurrent artificial intelligence*

The base-case simulation for R3 was similar to that for R2, using the discrepancies reported by Martins Jarnalo *et al.*[66] The difference between R3 and R2 is that the assumption was made that the radiologist will manually correct the 4 mm measurement discrepancy of the stand-alone software measurement (*Table 67*). Therefore, the mean size of R3 simulated nodules was R1 + 0.182 mm, with a SD of 0.639, for both solid and subsolid nodules (see *Table 69*).

In a scenario analysis (scenario analysis 3), we further assumed that the radiologist would manually correct the ± 2 mm discrepancies of stand-alone software measurement (see *Table 67*). Thus, the mean size of R3 simulated nodules in scenario analysis 3 was R1 + 0.182 mm with a SD of 0.448 mm (see *Table 69*).

### *Reader 4: unaided radiologist*

Inputs for the accuracy of manual nodule size measurement using electronic callipers were based on the phantom study by Xie *et al.*[100] This study was chosen as the base case as it observed an underestimation of nodule size, whereas the second identified study[37] reported an overestimation. This DAR observed that 'the studies found similar[58,63] or significantly larger[47] nodule diameters with semiautomatic measurements compared to manual measurements' (see *Nodule diameter measurement*); we therefore rated the underestimation observed by Xie *et al.*[100] as more plausible and used it as the base case. This study found that the overall underestimation of diameter was 9.2 ± 6.0% for nodules of any density and 10.1 ± 6.9% for solid nodules.

In the simulation, the mean size of solid nodules was based on that of R1 minus 10.1%, and for subsolid nodules, the mean size was based on R1 minus 9.2%. When calculating the SD for the distribution of nodule sizes from Xie *et al.*,[100]

**TABLE 66** Nodule size measurement discrepancies of stand-alone AI compared with the reference standard as reported by Martins Jarnalo *et al.*[66]

| Size discrepancy (mm) | Number of nodules (R2 base case) |
|---|---|
| −2 | 2 |
| −1 | 2 |
| 0 | 54 |
| 1 | 16 |
| 2 | 2 |
| 4 | 1 |

**TABLE 67** Discrepancies of concurrent AI diameter measurements, estimated from Martins Jarnalo *et al.*[66]

| Size discrepancy (mm) | Number of nodules (R3 base case) | Number of nodules (scenario 3) |
|---|---|---|
| −2 | 2 | 0 (Corrected manually) |
| −1 | 2 | 2 |
| 0 | 54 | 54 |
| 1 | 16 | 16 |
| 2 | 2 | 0 (Corrected manually) |
| 4 | 0 (Corrected manually) | 0 (Corrected manually) |

we got a SD of 0.52. However, we expect the error for a manual diameter measurement to be greater than the error of the concurrent AI (R3), and therefore the SD was fixed at 1.5 × SD of R3 (1.5 × 0.639) (see *Table 69*).

In a scenario analysis (scenario analysis 4), inputs based on results from Cohen *et al.*[37] were used. This study observed that the manual measurements of the entire nodule were larger than the tumour size on pathology after resection, by a mean difference of + 2.38 mm. For both solid and subsolid nodules, mean nodule size (mm) was R1 + 2.38, with a SD of 0.50 and 0.46, respectively, for the screening population, and 0.47 and 0.41, respectively, for the symptomatic population. This was to keep the SD consistent with that in scenario analysis 1 (see *Table 69*).

A final scenario analysis, scenario analysis 5, was performed for both the screening and symptomatic populations, in which the following assumptions were made for the SDs of simulated nodule sizes; the mean for each reader was based on that of R1 (*Table 68*):

- R1: SD kept the same.
- R2 (stand-alone AI): we assumed that AI alone would perform worse than R3 and R4 (SD multiplied by 2).
- R3 (concurrent AI): we assumed that this reader would measure more accurately than R2 and R4 (SD multiplied by 0.5).
- R4 (unaided radiologist): we assumed that this reader would measure more accurately than R2 but worse than R3 (SD multiplied by 1.5).

### Other assumptions

Nodule type distribution was different for the screening and symptomatic populations.

### Running the simulation

The simulation followed these steps:

1. 1,000,000 observations are created, which are the simulated nodules.
2. We randomly assign a percentage of these nodules as either solid or subsolid.
3. We simulate R1's nodule size measurements using a log-normal distribution with the following parameters:
   a. number of nodules = 1,000,000
   b. \Mu = log(median nodule size − 3)
   c. $\sigma$ = the solution to rearranged quantile functions of the log-normal distribution populated using the reported IQR to calculate $\sigma$.

4. The measurements for the other three readers are simulated.
5. Summary statistics are produced.

The simulation was carried out using R version 4.1.0 (The R Foundation for Statistical Computing, Vienna, Austria).

**TABLE 68** Inputs for scenario analysis 5

| | | Screening population | | Symptomatic population | |
|---|---|---|---|---|---|
| Reader | SD multiple | SD (solid) | SD (subsolid) | SD (solid) | SD (subsolid) |
| R1 | 1 | 5.82 | 5.56 | 3.89 | 6.00 |
| R2 | 2 | 11.64 | 11.12 | 7.78 | 12.00 |
| R3 | 0.5 | 2.91 | 2.78 | 1.95 | 3.00 |
| R4 | 1.5 | 8.73 | 8.34 | 5.84 | 9.00 |

SD, standard deviation.

TABLE 69 Mean nodule size simulation inputs

| Population | Screening | Symptomatic | Both | Both | Both | Screening | Symptomatic | Both | Both | Both |
|---|---|---|---|---|---|---|---|---|---|---|
| **Reader** | **R1** | **R1** | **R2** | **R3** | **R4** | **R2** | **R2** | **R2** | **R3** | **R4** |
| **Distribution** | **Log-normal** | **Log-normal** | **Normal** | **Normal** | **Normal** | **Normal** | **Normal** | **Normal** | **Normal** | **Normal** |
| ***Solid*** | | | | | | | | | | |
| Mean | 3.6[a] | 4.7[a] | R1 + 0.234 | R1 + 0.182 | R1–10.1% | R1 | R1 | R1 + 0.969 | R1 + 0.182 | R1 + 2.38 |
| SD | 2.1[a] | 1.3[a] | 0.771 | 0.639 | 0.639*1.5 | 2.60 | 0.63 | 0.249 | 0.448 | 0.50 |
| ***Subsolid*** | | | | | | | | | | |
| Mean | 11.9[a] | 11.1[a] | R1 + 0.234 | R1 + 0.182 | R1–9.2% | R1 | R1 | R1 + 0.884 | R1 + 0.182 | R1 + 2.38 |
| SD | 11.1[a] | 1.3[a] | 0.771 | 0.639 | R1 * 6.0% | 0.54 | 0.53 | 0.411 | 0.448 | 0.46 |
| Base case | Base case | Base case | Base case | Base case | Base case | | | | | |
| Scenario | | | | | | 1 | 1 | 2 | 3 | 4 |

[a] Median/IQR.

## Simulation for nodule growth monitoring

We used the nodules simulated using the base-case assumptions for R0 and applied the different growth curves (for both solid and subsolid nodules) to calculate the true nodule growth at each subsequent time point (3, 12, 24 and 48 months) for malignant nodules. For non-malignant nodules, we did not model any change or growth from their starting size. Then we back-calculated the 'true' diameter from the volume at each time point.

Using these 'true' diameter values at each time point, we applied the same transformations to R0 that we applied at baseline for R3 and R4 and calculated the corresponding estimated nodule volumes and VDTs.

To track the solid nodules' growth over time from the baseline to turning into cancerous nodules, we used the model developed by Treskova *et al.*[71] Treskova *et al.* suggest a Gompertz function with a log-normal distribution for the scale and shape parameters of the nodule growth over the patient's lifetime.

The study used a spherical volume measurement to compute the volume of the nodule and provided the VDT for four common histological lung cancer types:

1. small-cell carcinoma
2. large-cell carcinoma
3. squamous cell carcinoma
4. adeno/AIS carcinoma.

The threshold values for each type of this carcinoma were provided at four stages of cancer: regional stage, distant stage, diagnosis before the regional stage, and diagnosis after the regional stage. Then they followed a NELSON trial nodules algorithm management, which means that based on the assessed volume (*V*) the screening-detected nodule is classified as a negative ($V < V_{fup}$), positive ($V \geq V_{cut}$) or indeterminate result ($V_{fup} \leq V < V_{cut}$). More details on the Treskova *et al.*[71] study can be found in *Appendix 7*, *Table 65*, and text.

For subsolid nodules, a linear growth over time was assumed, as reported by Kakinuma *et al.*[101]

Treskova *et al.*[71] is used in this simulation as follows.

For R0, nodule volume was calculated from the baseline nodule diameter. The growth function was then applied to calculate nodule volume at subsequent time points. Then nodule diameter was calculated by rearranging the formula for the sphere volume. Using the newly calculated diameters, VDT was calculated.

Using the diameters, volume and VDT that were calculated for R1, R3 and R4, we calculated the probabilities for the model structure. The formulae used for the calculation have been described in *Appendix 7*.

# **Appendix 9** Findings of probabilistic sensitivity analyses for the cost-effectiveness analyses from the full model

## Symptomatic population

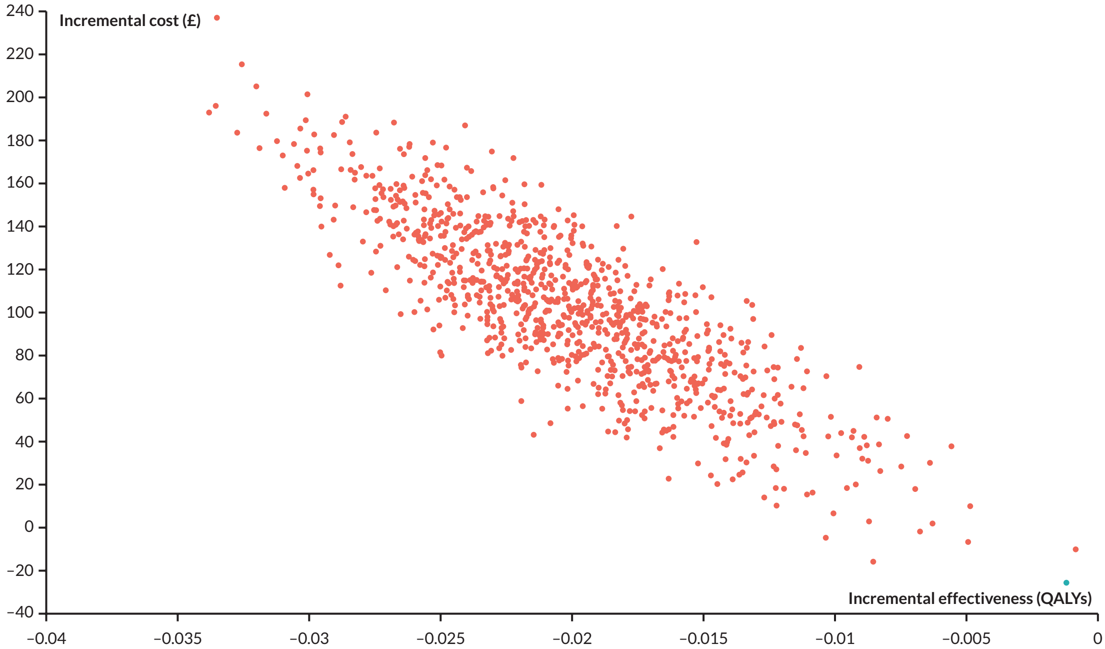


**FIGURE 24** Incremental cost-effectiveness scatterplot for the comparison of AI-assisted radiologist reading with unaided radiologist reading (symptomatic population).

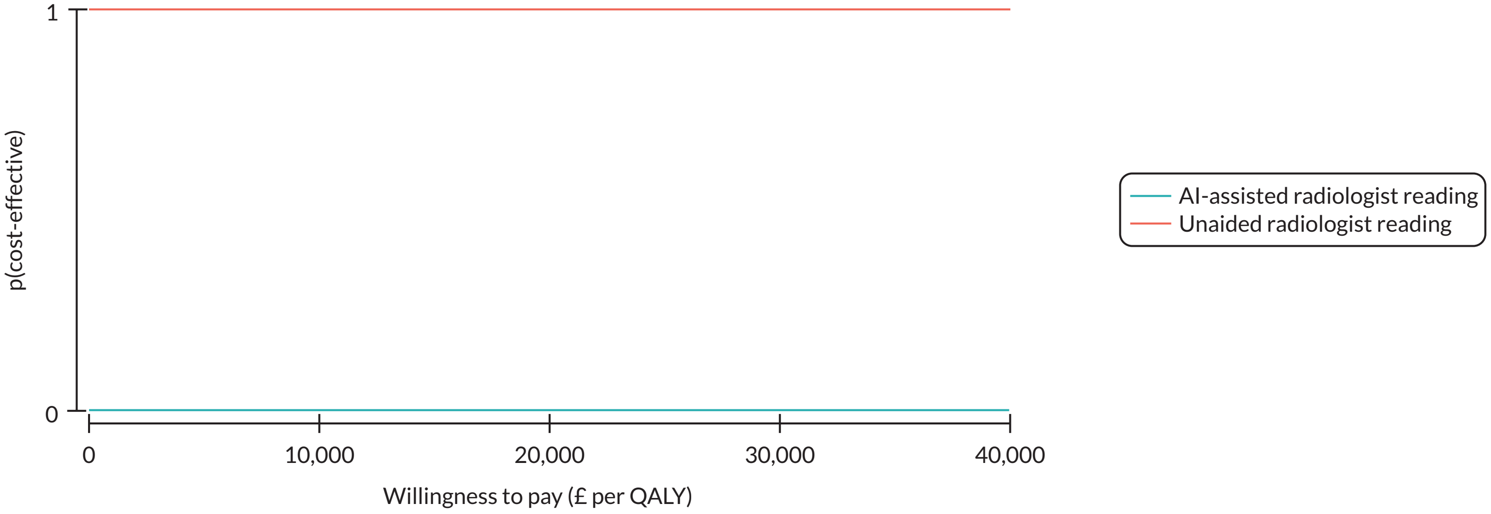


**FIGURE 25** Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (symptomatic population).

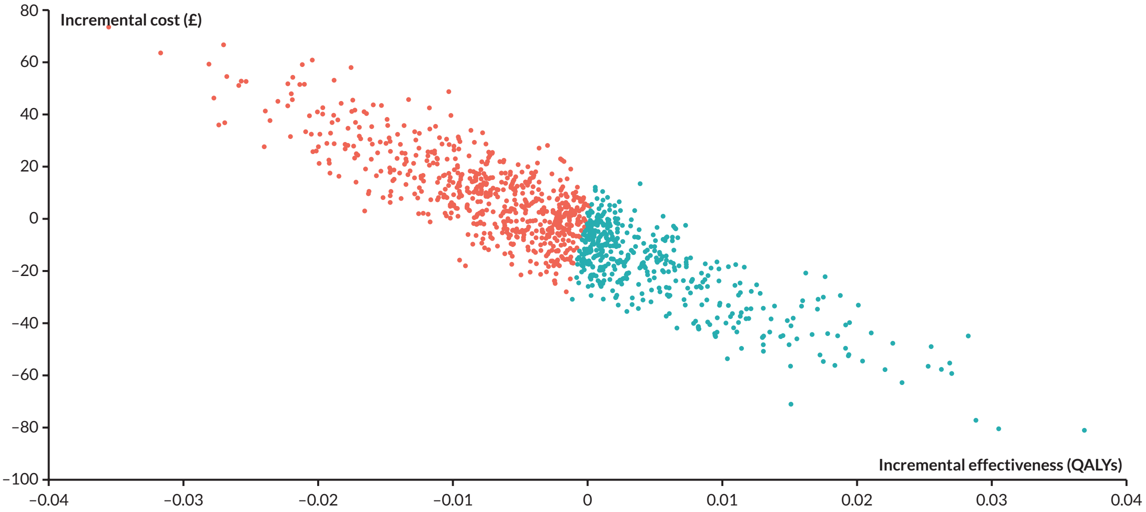


**FIGURE 26** Incremental cost-effectiveness scatterplot for the comparison of AI-assisted radiologist reading with unaided radiologist reading (incidental population).

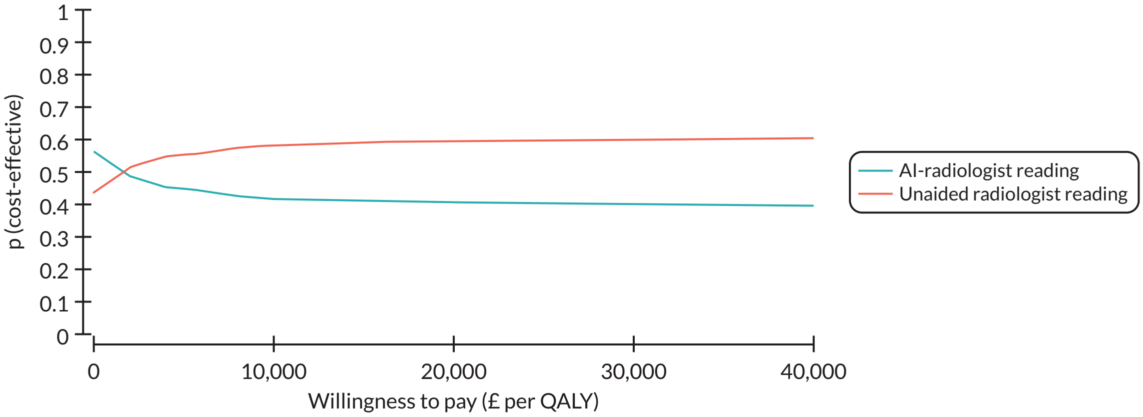


**FIGURE 27** Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (incidental population).

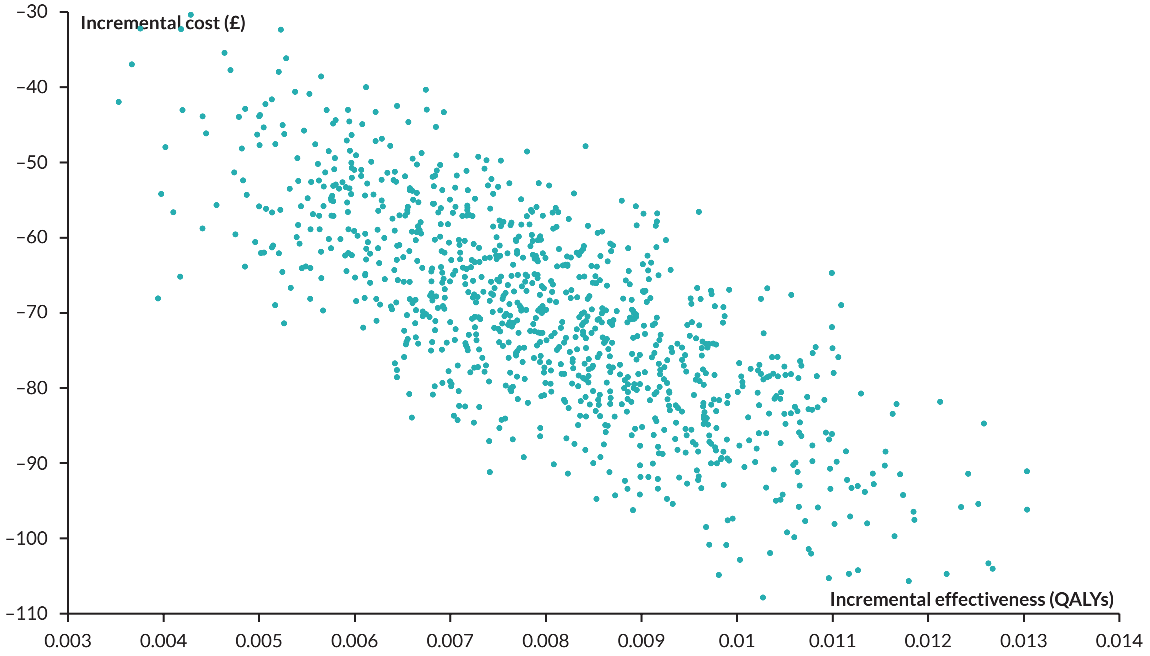


**FIGURE 28** Incremental cost-effectiveness scatterplot for the comparison of AI-assisted radiologist reading with unaided radiologist reading (screening population).

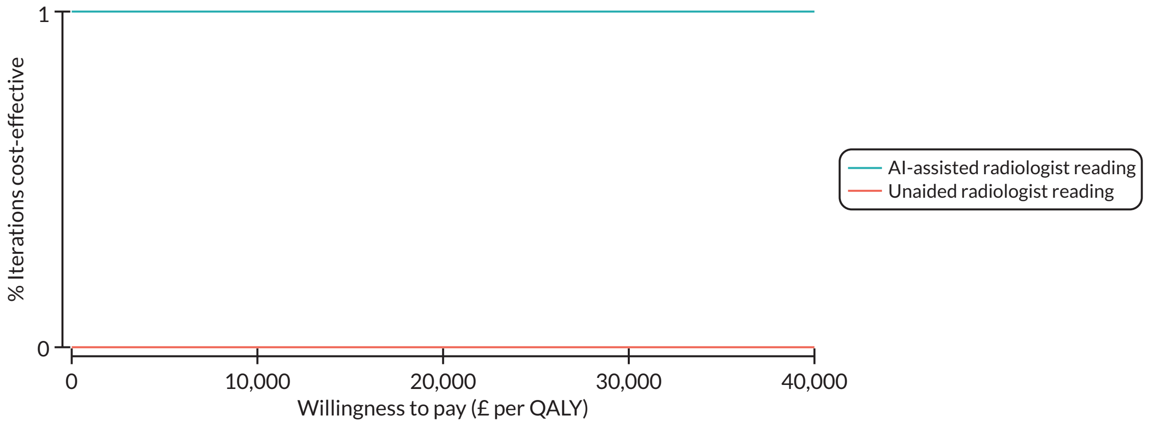


**FIGURE 29** Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (screening population).

EME
HSDR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

*This report presents independent research funded by the National Institute for Health and Care Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care*

***Published by the NIHR Journals Library***