Technology Assessment Report commissioned by the NIHR Evidence Synthesis Programme on behalf of the National Institute for Health and Care Excellence – Diagnostics Assessment Report

Title: Software with artificial intelligence derived algorithms for automated detection and analysis of lung nodules from CT scan images [DAP060]

Produced by	Warwick Evidence, Warwick Medical School, University of Warwick, [redacted]
Authors	Julia Geppert, Research Fellow, Warwick Evidence and Warwick Screening Peter Auguste, Assistant Professor, Warwick Evidence Asra Asgharzadeh, Research Fellow, Warwick Evidence Hesam Ghiasvand, Research Fellow, Warwick Evidence Mubarak Patel, Research Associate, Warwick Evidence Anna Brown, Information Specialist, Warwick Evidence Surangi Jayakody, Assistant Professor, Warwick Evidence Emma Helm, Consultant Radiologist, University Hospitals Coventry and Warwickshire Dan Todkill, Associate Clinical Professor, Warwick Medical School Jason Madan, Professor, Warwick Medical School Chris Stinton, Senior Research Fellow, Warwick Evidence and Warwick Screening Daniel Gallacher, Assistant Professor, Warwick Evidence Sian Taylor-Phillips, Professor, Warwick Screening Yen-Fu Chen, Associate Professor, Warwick Evidence
Correspondence to	Dr Yen-Fu Chen Warwick Evidence, Division of Health Sciences Warwick Medical School, University of Warwick, [redacted] Email: [redacted]
Date completed	5th December 2022

Source of funding: This report was commissioned by the NIHR Evidence Synthesis Programme as project number 13/53/25.

Declared competing interests of the authors

None.

Acknowledgements

Prof. Ben Glocker (Reader in Machine Learning for Imaging, Imperial College London) provided expert technical advice to the EAG during the preparation of the report.

Prof. 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.

We are grateful to the Specialist Committee members for their advice during our preparation of the economic models.

We thank Sarah Abrahamson for her assistance in editing and proof-reading the report.

Rider on responsibility for report

The views expressed in this report are those of the authors and not necessarily those of the NIHR Evidence Synthesis Programme. Any errors are the responsibility of the authors.

This report should be referenced as follows:

Geppert J, Auguste P, Asgharzadeh A, Ghiasvand H, Patel M, Brown A, Jayakody S, Helm E, Todkill D, Madan J, Stinton C, Gallacher D, Taylor-Phillips S, Chen Y-F. Software with artificial intelligence derived algorithms for automated detection and analysis of lung nodules from CT scan images: A Diagnostics Assessment Report. Warwick EAG, 2022

Contributions of authors

Julia Geppert: Performed the clinical effectiveness review and wrote associated sections of this report.

Peter Auguste: Led the cost-effectiveness components of this project, undertook systematic review of the health economic literature, constructed health economic models and wrote economics sections of this report.

Asra Asgharzadeh: Performed the clinical effectiveness review and wrote associated sections of this report.

Hesam Ghiasvand: Performed the systematic review of the health economic literature, undertook health economic modelling and wrote economics sections of this report.

Mubarak Patel: Performed statistical analyses and wrote associated sections of the report.

Anna Brown: Developed the search strategies, undertook searches, managed references and wrote the search methods sections of this report.

Surangi Jayakody: Performed a review of the literature on overdiagnosis, supported the clinical effectiveness review and report writing of associated sections.

Emma Helm: Provided expert clinical advice and helped development of economic models.

Dan Todkill: Commented on earlier versions of the report and assisted in revising the report.

Jason Madan: Provided methodological advice on economic modelling and revised associated sections of the report.

Chris Stinton: Provided methodological advice and training for the clinical effectiveness team, acted as 3rd reviewer, and assisted in revising the report.

Daniel Gallacher: Provided statistical advice on simulation and assisted in revising the report.

Sian Taylor-Phillips: Contributed to study design and protocol, acted as senior advisor, provided methodological support, and assisted in revising the report.

Yen-Fu Chen: Led the project, its coordination and implementation, and write-up.

Please note that:

Sections highlighted in yellow and underlined are ‘academic in confidence’ (AIC). Sections highlighted in aqua and underlined are ‘commercial in confidence’ (CIC).

Copyright statement:

Copyright belongs to the University of Warwick.

Copyright is retained by British Thoracic Society guideline authors for **Figure 1**, **Figure 2** and **Figure 3**.

ABSTRACT

OBJECTIVE

To assess accuracy, clinical and cost-effectiveness of computed tomography (CT) image analysis assisted by software with artificial intelligence (AI) derived algorithms capable of automated detection and analysis of lung nodules, compared with unassisted analysis in people undergoing CT scans that include 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.

DESIGN

Systematic review and de-novo cost-effectiveness analysis (CEA).

METHODS

We performed systematic reviews including studies on 13 NICE specified technologies for CT image analysis with outcomes on nodule detection and measurement accuracy or reliability, practical implications, impact on patient management (key question 1, KQ1), clinical effectiveness (KQ2) and cost effectiveness (KQ3). We searched electronic databases and other sources from 2012 to January 2022. Company submissions were accepted until 31 August 2022.

Study quality was assessed by QUADAS-2 (and QUADAS-C, if applicable) and COSMIN Risk of bias tool. Outcomes were synthesised narratively.

We adopted two approaches to decision modelling, both used decision trees. One is a simple decision tree evaluating cost-effectiveness of AI-assisted image analysis for the detection of actionable nodules using test accuracy results. The other is a more extensive decision tree reflecting the full clinical pathways for people undergoing chest CT scan. Information on prevalence of lung nodules, sensitivity and specificity for nodule detection and reliability of nodule measurement was linked to the British Thoracic Society (BTS) guidelines through simulation, incorporating a further model to account for growth of malignant nodules during surveillance. The model estimates incremental cost-effectiveness ratios (ICERs) expressed as cost per quality-adjusted life year (QALY) (primary outcome). Secondary outcome measures (cost per correct detection of a person with an actionable nodule, cost per cancer detected and treated) were analysed. We undertook a series of scenario analyses and sensitivity analyses.

RESULTS

For KQ1, 27 studies evaluating eight of the 13 NICE-specified technologies were identified that reported outcomes of interest. All studies were at high risk of bias. No study directly compared radiologists assisted by different technologies of interest. Twenty-four studies used retrospective datasets, 17 of which compared the performance of readers with and without AI software (main comparison of interest). One study reported on prospective screening experiences before and after AI software implementation. The remaining studies either evaluated stand-alone AI (outside NICE scope) or only provided non-comparative evidence.

Accuracy / reliability

Nodule detection - AI-assisted reading generally improved sensitivity, with similar or lower specificity compared with unaided reading. Estimated sensitivity and specificity varied substantially between studies, possibly due to heterogeneity in patient population, reader speciality and experience, reading conditions, other study design features and risk of bias.

Nodule size measurement - Measured nodule diameters were similar or significantly larger with semi-automatic measurements compared to manual measurements. Intra-reader and inter-reader agreement in nodule size measurement and in risk classification based on clinical guidelines generally improved with AI-assistance or are comparable to unaided reading.

Practical implications

Segmentation failure or rejection of automated segmentation by radiologists ranged from 0% to 57% of nodules. Radiologist reading time generally decreased with AI assistance in research setting.

Impact on patient management

AI-assisted reading tended to upstage risk categories defined by clinical guidelines based on retrospective application of findings from AI-assisted reading.

Clinical and cost-effectiveness

For KQ2 and KQ3, no relevant studies were identified.

De-novo CEA

Due to the complete absence of clinical effectiveness evidence and major challenges in linking test accuracy evidence to clinical and economic outcomes, methods and findings presented here are highly uncertain and should serve as early indicators and frameworks for future assessment. Our CEA suggested for the symptomatic and incidental populations that AI-assisted CT image analysis dominates the 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 the unaided reader. This is driven by the costs and disutilities associated with false positive results and CT surveillance. AI was deemed cost effective for both symptomatic and incidental populations in the scenario analyses where disutility associated with false positive results and CT surveillance were removed. 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 estimates of improved test specificity for AI. Although there was more data available to populate the screening population model, there was very great uncertainty across all models.

LIMITATIONS

The identified evidence was of low quality and high applicability concerns. No study was performed prospectively in clinical or screening practice in the UK. Available evidence was very limited and heterogeneous, preventing meta-analyses, subgroup analyses and reliable CEA.

CONCLUSIONS

AI-assisted analysis of CT scan images may reduce variability and improve consistency in the measurement of lung nodules and in clinical management following current guidelines. AI-assisted analysis may increase the accuracy of nodule detection but may also increase the number of patients undergoing CT surveillance. Current evidence is largely collected from research settings and will need to be verified by evidence collected prospectively from clinical settings.

No direct comparative evidence between AI technologies of interest was found, and no study provided direct evidence on clinical outcomes and cost-effectiveness. We established a methodological framework for economic evaluation, which suggested AI-assisted image analysis may be cost-effective for the screening population but may be dominated by unaided analysis for the symptomatic population, but reliable estimates of cost-effectiveness cannot be obtained until more evidence becomes available.

TABLE OF CONTENTS

ABSTRACT 5

LIST OF TABLES AND LIST OF FIGURES 15

List of tables 15

List of figures 19

DEFINITION OF TERMS AND LIST OF ABBREVIATIONS 22

Glossary 22

List of abbreviations 25

SCIENTIFIC SUMMARY 28

PLAIN ENGLISH SUMMARY 38

1 BACKGROUND AND DEFINITION OF THE DECISION PROBLEM(S) 40

1.1 Lung nodules and lung cancer 40

1.2 Diagnostic and care pathway 41

1.2.1 Pathway to CT scan due to signs and symptoms suggestive of lung cancer 41

1.2.2 Lung cancer screening 41

1.2.3 Initial assessment and CT surveillance of lung nodules 42

1.2.4 Current methods of detecting nodules and measuring nodule volume and growth on CT scans 46

1.2.5 Diagnosis and staging of lung cancer 49

1.2.6 Treatment for lung cancer 49

1.3 Population and relevant subgroups 49

1.4 Description of technology(ies) under assessment 50

1.4.1 AI-Rad Companion Chest CT (Siemens Healthineers) 51

1.4.2 AVIEW LCS+ (Coreline Soft) 51

1.4.3 ClearRead CT (Riverain Technologies) 51

1.4.4 contextflow SEARCH Lung CT (contextflow) 52

1.4.5 InferRead CT Lung (Infervision) 52

1.4.6 JLD-01K (JLK Inc.) 52
1.4.7 Lung AI (Arterys) 53
1.4.8 Lung Nodule AI (Fujifilm) 53
1.4.9 qCT-Lung (Qure.ai) 53
1.4.10 SenseCare-Lung Pro (SenseTime) 53
1.4.11 Veolity (MeVis) 54
1.4.12 Veye Lung Nodules (Aidence) 54
1.4.13 VUNO Med-LungCT AI (VUNO) 54
1.5 Proposed position of the intervention in the diagnostic pathway 55
1.6 Comparators 57
1.7 Outcomes 58
1.8 Objectives 58
2 SYSTEMATIC REVIEW OF ASSESSING TEST ACCURACY, PRACTICAL IMPLICATIONS AND IMPACT ON PATIENT MANAGEMENT (KEY QUESTION 1) - METHODS 61
2.1 Identification and selection of studies 61
2.1.1 Search strategy 61
2.1.2 Study eligibility criteria 62
2.1.3 Study screening and selection 64
2.2 Data extraction and risk of bias assessment 65
2.2.1 Data extraction strategy 65
2.2.2 Assessment of study risk of bias 65
2.3 Methods of analysis/synthesis 65
3 SYSTEMATIC REVIEW OF ASSESSING TEST ACCURACY, PRACTICAL IMPLICATIONS AND IMPACT ON PATIENT MANAGEMENT (KEY QUESTION 1) - RESULTS 67
3.1 Description of the evidence 67
3.1.1 Results of literature search 67
3.1.2 Characteristics of included studies 69

3.2 Methodological quality of the evidence 80
3.2.1 Risk of bias and applicability concerns according to QUADAS-2 and QUADAS-C 80
3.2.2 Risk of bias for reliability and measurement error (COSMIN tool) 90
3.3 Use case 1: nodule detection and analysis in people with no known lung nodules 91
3.3.1 Nodule detection 91
3.3.2 Nodule type determination 120
3.3.3 Nodule diameter measurement 121
3.3.4 Nodule volume measurement 129
3.3.5 Classification into risk categories based on nodule type and size 133
3.3.6 Whole read 140
3.4 Use case 2: nodule growth monitoring in people with previously identified lung nodules 141
3.4.1 Detection of growing nodules (No study) 141
3.4.2 Nodule registration and growth assessment 141
3.5 Practical implications 143
3.5.1 Technical failure rate (12 studies) 143
3.5.2 Radiologist reading time (10 studies) 150
3.5.3 Radiology report turnaround time (No study) 158
3.5.4 Acceptability and experience of using the software (3 studies) 158
3.5.5 Other non-prespecified outcomes 159
3.5.6 Sub-questions 1 to 6. 159
3.6 Impact on patient management 159
3.6.1 Characteristics of detected nodules 159
3.6.2 Proportion of detected nodules that are malignant (3 studies) 174
3.6.3 Impact of test result on clinical decision-making (6 studies) 177
3.6.4 Number of people having CT surveillance (5 studies) 181
3.6.5 Number of CT scans taken as part of CT surveillance (No study) 182
3.6.6 Number of people having a biopsy or excision (5 studies) 182

3.6.7 Stage of cancer at detection (No study) 183
3.6.8 Time to diagnosis (1 study) 184
3.7 Ongoing and/or unpublished studies 184
4 SYSTEMATIC REVIEW OF CLINICAL EFFECTIVENESS (KEY QUESTION 2) – METHODS AND RESULTS 185
4.1 Methods 185
4.1.1 Identification and selection of studies 185
4.2 Results 186
5 SYSTEMATIC REVIEW OF COST-EFFECTIVENESS (KEY QUESTION 3) – METHODS AND RESULTS 187
5.1 Methods for systematic review of cost-effectiveness 187
5.1.1 Identification and selection of studies 187
5.1.2 Extraction and study quality 189
5.1.3 Methods of analysis/synthesis 190
5.2 Results for systematic review of cost-effectiveness 190
5.2.1 Results of literature search 190
5.2.2 Description of the evidence 192
6 PRELIMINARY MODEL – METHODS AND RESULTS 197
6.1 Developing the model structure 197
6.2 Strategies 197
6.3 Information required for the model 200
6.3.1 Prevalence 200
6.3.2 Test accuracy 200
6.3.3 Resource use and costs 201
6.3.4 Outcomes 202
6.3.5 Analysis 203
6.4 Results 203
6.4.1 Deterministic results 203

6.4.2 Sensitivity analysis results 204
6.4.3 Scenario analysis results 205
6.4.4 Discussion 206
7 DE NOVO COST-EFFECTIVENESS ANALYSIS (FULL MODEL) – METHODS 208
7.1 Developing the model structure 208
7.2 Strategies 208
7.3 Natural history 212
7.4 Information required for the model 212
7.4.1 EAG simulation of measurement accuracy and precision 212
7.4.2 Prevalence 217
7.4.3 Test accuracy 219
7.4.4 Effectiveness 223
7.4.5 Resource use and costs 223
7.4.6 Utility values 227
7.4.7 Mortality 228
7.4.8 Outcomes 228
7.4.9 Model assumptions 228
7.4.10 Analysis 229
7.4.11 Areas beyond the scope of the assessment 232
8 DE NOVO COST-EFFECTIVENESS ANALYSIS (FULL MODEL) - RESULTS 233
8.1 Base-case results 233
8.1.1 Symptomatic population 235
8.1.2 Incidental population 241
8.1.3 Screening population 247
8.1.4 Surveillance population 253
8.2 Discussion 255
8.2.1 Summary of key results 255

8.2.2 Generalisability of results 257
8.2.3 Strengths and limitations of analysis 258
9 ASSESSMENT OF FACTORS RELEVANT TO THE NHS AND OTHER PARTIES 260
10 DISCUSSION 262
10.1 Statement of principal findings 262
10.2 Strengths and limitations of the assessment 263
10.2.1 Strengths 263
10.2.2 Limitations 263
10.3 Uncertainties 267
10.4 Other relevant factors 268
11 CONCLUSIONS 269
11.1 Implications for service provision 269
11.2 Suggested research priorities 270
12 REFERENCES 272
13 APPENDICES 280
13.1 Appendix 1: Literature search strategies: systematic review of test accuracy and clinical effectiveness 280
13.2 Appendix 2: Table of excluded studies with rationale 297
13.3 Appendix 3: Data extraction tables 325
13.4 Appendix 4: Quality assessment 342
13.5 Appendix 5: Additional evidence on test accuracy of stand-alone AI and other evidence from non-comparative studies 363
13.5.1 Accuracy for detecting any nodules 363
13.5.2 Accuracy for detecting actionable nodules 366
13.5.3 Accuracy for detecting malignant nodules 368
13.5.4 Nodule type determination 369
13.5.5 Whole read 373

13.5.6 Nodule registration and growth assessment 373

13.5.7 Practical implications – Additional results 374

13.5.8 Impact on patient management - Additional results 375

13.6 Appendix 6: Literature search strategies: searches to inform the economic model 383

13.6.1 Searches for information on model structures, costs and utility values to inform the economic model 383

13.6.2 Searches for pulmonary nodule growth rates / volume doubling times 391

13.6.3 Searches for pulmonary nodule prevalence by size and type 397

13.7 Appendix 7: Growth model and its development process 400

13.8 Appendix 8: Methods for simulation 430

13.8.1 Simulation for nodule sizes at baseline (baseline measurement simulation) 430

13.8.2 Simulation for nodule growth monitoring 440

13.8.3 R code for the simulation 441

13.9 Appendix 9: Findings of probabilistic sensitivity analyses for the cost-effectiveness analyses from the full model 446

13.10 Appendix 10: Rationale for developing the Warwick Evidence (WE) model and comparison with the Exeter NAtural History-Based economic model of Lung cancer screening (ENaBL) model used by the National Screening Committee (NSC) 450

LIST OF TABLES AND LIST OF FIGURES

List of tables

Table 1. Summary of the included technologies (reproduced from final NICE scope) 55

Table 2. Characteristics of included studies (n=27) 71

Table 3. Outcomes – Nodule detection and analysis: Accuracy, concordance and variability 77

Table 4. Outcomes – Practical implications 79

Table 5. Outcomes - Impact on patient management 79

Table 6. Quality assessment results based on QUADAS-2 and QUADAS-C tools (22 studies) 81

Table 7. Quality of studies assessed by COSMIN Risk of Bias tool[24] (4 studies) 90

Table 8. Characteristics of included studies with comparative results for nodule detection accuracy, and their quality ratings (12 studies) 93

Table 9. Characteristics of included studies with non-comparative results for nodule detection accuracy and quality ratings (8 studies) 96

Table 10. Summary of evidence related to accuracy of AI-assisted reading and stand-alone AI for detecting malignant nodules (6 studies) 108

Table 11. Accuracy for the detection of any nodules in standard dose and low dose CT scans according to Hsu et al.[51] 111

Table 12. Effect of nodule type on nodule detection accuracy in screening populations - Concurrent AI vs unaided reader (2 studies) 114

Table 13. Effect of nodule type on nodule detection accuracy in a symptomatic population - Concurrent AI / stand-alone AI vs unaided reader (1 study) 114

Table 14. Effect of nodule type on nodule detection accuracy in screening population - Stand-alone AI (2 studies) 115

Table 15. Effect of nodule type on nodule detection accuracy in mixed populations - Stand-alone AI alone (1 study) or vs unaided readers (1 study) 117

Table 16. Main findings, risk of bias, applicability concerns and input into modelling 121

Table 17. Accuracy of readers with and without concurrent use of Veye Chest to identify patients with BTS grade A (no clinical follow-up recommended)[32] 133

Table 18. Risk categorisation using standard CT images and vessel-suppressed CT images for semi-automatic volume measurement (modified from Milanese et al.[53]) 135

Table 19. Technical failure rate of AI-based software for lung nodule detection and analysis, by target population and technology (12 studies) 147

Table 20. Effect of software use on radiologist reading time, by target population and technology (10 studies) 155

Table 21. Nodule number, type and size in patients with incidentally detected nodules on CT, with and without concurrent use of Veye Chest[32] 160

Table 22. Characteristics of detected nodules (true and false positives) in consecutive screening populations from Korea (3 studies) 162

Table 23. Nodule 2-D axial diameter in all detected nodules in patients with complex lung disease[45] 164

Table 24. Nodule type and size in a random symptomatic population from Japan[57] 166

Table 25. Nodule characteristics of subjects with at least 1 nodule in a consecutive screening population from China, by mode of detection[59] 167

Table 26. Characteristics of correctly detected nodules in a mixed population from China obtained via convenience sampling.[58] 168

Table 27. Characteristics of all detected nodules, true positive, false positive and false negative nodules – Stand-alone software in a random mixed population[64] 170

Table 28. Characteristics of missed nodules in a consecutive screening population from China[59] 172

Table 29. Characteristics of missed nodules in a mixed population from China obtained via convenience sampling.[58] 173

Table 30. Proportion of detected risk-dominant nodules that are malignant, by nodule type and size, in a consecutive screening population from Korea[48] 176

Table 31. Lung-RADS category with and without concurrent software use in a nodule-enriched screening population[62]) 177

Table 32. Lung-RADS category based on stand-alone software and readers with and without concurrent software use in a nodule-enriched screening population[65] 179

Table 33. Sub-solid nodule classification of the two readers with and without software use in patients with previously detected nodules[61] 180

Table 34. Risk classification based on semi-automatic volume measurement using standard CT images and vessel-suppressed CT images in consecutive LDCT with unclear indication[53] 181

Table 35. Test accuracy estimates for identifying actionable nodules by test strategy 200

Table 36. Resource use associated with reporting CT scans 202

Table 37. Costs inputs used in the model 202

Table 38. Deterministic results based on expected costs and expected identification of people with actionable lung nodules (screening population of 1,000 people undergoing CT scan) 204

Table 39. Scenario analysis results based on cost per person with an actionable lung nodule correctly identified (screening population) 206

Table 40. Prevalence of having at least one lung nodule by population of interest 217

Table 41. Proportion of detected risk-dominant nodules that are solid/sub-solid 218

Table 42. Prevalence of lung cancer in detected nodules, by population and nodule measurement 218

Table 43. Comparative studies reporting detection accuracy for any nodules that could be used as CEA model inputs and their advantages and disadvantages (3 studies) 220

Table 44. Test accuracy estimates to identify any lung nodule by reason for undergoing CT scan ... 222

Table 45. Technologies outlined in scope against our selection criteria for the base-case economic analysis 225

Table 46. Costs inputs used in the model 226

Table 47. Scenario analyses by changing the prevalence of any lung nodules detected at baseline CT scans in a screening population and incidental population, respectively 231

Table 48. Resource use associated with reading and reporting CT scans 231

Table 49. Summary of intermediate outcomes from the full model 234

Table 50. Deterministic results based on expected costs and expected correct identification of people with actionable lung nodules (symptomatic population of 1000 people undergoing CT scan) 235

Table 51. 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) 236

Table 52. Deterministic results based on expected costs and expected QALYs (symptomatic population of 1000 people undergoing CT scan) 237

Table 53. Scenario analysis results based on cost per QALY (symptomatic population) 239

Table 54. Deterministic results based on expected costs and expected cases appropriately identified (incidental population of 1,000 people undergoing CT scan) 242

Table 55. Deterministic results based on expected costs and expected cancer correctly detected and treated (incidental population of 1,000 undergoing CT scan) 242

Table 56. Deterministic results based on expected costs and expected QALYs (incidental population of 1000 undergoing CT scan) 243

Table 57. Scenario analysis results based on cost per QALY (incidental population) 245

Table 58. Deterministic results based on expected costs and expected correct identification of people with actionable nodules (screening population of 1,000 people undergoing CT scan) 247

Table 59. Deterministic results based on expected costs and expected identification of people with cancer detected and treated (screening population of 1000 people undergoing CT scan) 248

Table 60. Deterministic results based on expected costs and expected QALYs (screening population of 1000 undergoing CT scan) 248

Table 61. Scenario analysis results based on cost per QALY (screening population) 250

Table 62. Deterministic results based on expected costs and QALYs (screening population of 1,000 people undergoing CT surveillance) 254

Table 63. Deterministic results based on expected costs and QALYs (screening population of 1,000 people undergoing CT surveillance) 254

Table 64. Publications excluded after review of full-text articles – Electronic database searches (n=150) 297

Table 65. Publications excluded after review of full-text articles – Studies provided by companies (n=99) 309

Table 66. Study characteristics and main outcomes of records excluded on study population only (n=11) 316

Table 67. Characteristics of ongoing and/or unpublished studies (7 studies) 322

Table 68. Study level description of the 27 included studies for key question 1 326

Table 69. Accuracy of stand-alone software to determine nodule type (2 studies) 372

Table 70. Characteristics of correctly detected and missed nodules of stand-alone software in a consecutive screening population in Korea[49] 376

Table 71. Characteristics of studies that included a growth model 407

Table 72. Nodule size measurement discrepancies of stand-alone AI compared to the reference standard as reported by Martins Jarnalo et al. 2021[64] 433

Table 73. Discrepancies of concurrent AI diameter measurements, estimated from Martins Jarnalo et al.[64] 434

Table 74. Inputs for scenario analysis 5 435

Table 75. Mean nodule size simulation inputs 439

List of figures

Figure 1. Initial assessment of solid lung nodules (reproduced with permission from Callister et al. 2015)[11] 43

Figure 2. Sub-solid pulmonary nodules algorithm (reproduced with permission from Callister et al. 2015)[11] 44

Figure 3. CT surveillance of solid lung nodules (reproduced with permission from Callister et al. 2015)[11] 45

Figure 4. Points at which AI derived software may have an impact in the process of nodule detection and analysis and relevant evidence in this report 48

Figure 5. Proposed roles of the intervention in the process of diagnosing lung cancer 56

Figure 6. PRISMA diagram. Summary of publications included and excluded at each stage of the review 68

Figure 7. Findings of risk of bias assessment for all 22 studies as well as separately for comparative (QUADAS-C) and non-comparative (QUADAS-2) studies 86

Figure 8. Findings of applicability concern assessment (QUADAS-2) by index test 88

Figure 9. Visual map of included studies for detection accuracy based on population, comparison and reported outcomes (targets of detection) 92

Figure 10. Evidence on AI-assisted reading compared with unaided reading for accuracy of detecting any nodules (7 studies) 99

Figure 11. Comparative evidence for accuracy of detecting actionable nodules (6 studies) 104

Figure 12 An illustration of linked evidence approach adopted for this diagnostic assessment 186

Figure 13. PRISMA Flow diagram for economic evaluation of using the AI for detection of lung nodules 191

Figure 14. Illustrative model structure for the detection of actionable lung nodules 199

Figure 15. Tornado diagram of the impact to the cost per actionable lung nodule correctly identified by changing individual parameters (screening population) 205

Figure 16. Illustrative structure of the clinical pathways 209

Figure 17. Illustrative model structure for the detection of lung nodules 211

Figure 18. Abbreviated representation of the decision tree, required model parameters and data source (further parts shown in Figure 18 & 19) 214

Figure 19. Abbreviated representation of the solid nodule part of the decision tree, required model parameters and data source (continued from Figure 17) 215

Figure 20. Abbreviated representation of the sub-solid nodule part of the decision tree, required model parameters and data source (continued from Figure 17) 216

Figure 21. Tornado diagram of the impact to the cost per QALY by changing individual parameters (symptomatic population) 238

Figure 22. Tornado diagram of the impact to the cost per QALY Identified by changing individual parameters (incidental population) 244

Figure 23. Tornado diagram of the impact to the cost per QALY by changing individual parameters (screening population) 249

Figure 24. Incremental cost-effectiveness scatterplot for the comparison between AI-assisted radiologist reading versus unaided radiologist reading (symptomatic population) 446

Figure 25. Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (symptomatic population) 446

Figure 26. Incremental cost-effectiveness scatterplot for the comparison between AI-assisted radiologist reading versus unaided radiologist reading (incidental population) 447

Figure 27. Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (incidental population) 447

Figure 28. Incremental cost-effectiveness scatterplot for the comparison between AI-assisted radiologist reading versus unaided radiologist reading (screening population) 448

Figure 29. Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (screening population)..... 448

DEFINITION OF TERMS AND LIST OF ABBREVIATIONS

Technical terms and abbreviations are used throughout this report. The meaning is usually clear from the context, but a glossary is provided for the non-specialist reader.

Glossary

<table>
<tr><th>Term</th><th>Definition</th></tr>
<tr><td>2x2 contingency table</td><td>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 which is considered to reflect the true disease status in the following form:
<table>
<tr><td>Reference (gold) standard (→) / Index (new) test (↓)</td><td>Yes</td><td>No</td></tr>
<tr><td>Yes</td><td>a = TP</td><td>b = FP</td></tr>
<tr><td>No</td><td>c = FN</td><td>d = TN</td></tr>
</table></td></tr>
<tr><td>Cohen’s Kappa</td><td>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.</td></tr>
<tr><td>Concordance</td><td>The agreement between two variables.</td></tr>
<tr><td>Concurrent AI</td><td>In this report, concurrent AI refers to the use of AI software at the same time when a radiologist is reading and analysing the CT scan image. This is in contrast with second-read AI (see definition below).</td></tr>
<tr><td>Correlation</td><td>The degree which two continuous variables are linearly related.</td></tr>
<tr><td>Dice similarity coefficient (DSC) or Dice coefficient</td><td>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) to 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.
$$DSC = \frac{2a}{2a+b+c} = \frac{a}{a+\frac{1}{2}(b+c)} = \frac{2TP}{2TP+FP+FN}$$
Dice coefficient ranges between 0 and 1, with 1 signifying the greatest similarity between the two tests.</td></tr>
</table>

	Also known as the F score or the Sorensen-Dice coefficient.
False negative value	The number of cases in which the index test has wrongly suggested the patient as being disease-free. $$FN = c$$
False positive rate	The proportion of people who test positive for a disease amongst people who do not have the disease of interest. The ratio between the false positive value and (true negative value + false positive value). Equals to 1 – specificity This term is sometimes used in the literature to describe the 'number of false positive detections per image' (see definition below), which may cause confusion.
False positive value	The number of cases in which the index test has wrongly indicated the patient as having the disease. $$FP = b$$
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 which shows 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 (CCC)	Also denoted as ρ_c, or CCC, is 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	How accurate 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) is compared with the 'true' quantity, e.g. whether computer software tends to over-estimate 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 above) or when different people try to measure the

	same quantity (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. $$NPV = \frac{d}{c+d} = \frac{TN}{FN+TN}$$
Number of false positive detections per image	For 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 CT scan image. The number of false positive detections per image represents the total number of false positive findings across a set of CT scan images divided by the total number of CT scan images within this set. For example, if an overall of 15 false positive findings are recorded among 10 CT 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 indicates perfect negative correlation, 1 indicates perfect positive correlation and 0 indicates 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 CT scan image. As multiple nodules may be found in a CT 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. $$PPV = \frac{a}{a+b} = \frac{TP}{TP+FP}$$
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.
Receiver operating	A graph showing the sensitivity and specificity for every possible threshold of a test.

characteristic (ROC) curve	
Risk dominant nodule	The lung nodule that is judged to carry the highest risk (or probability) of being a malignant nodule and based on which the decision on clinical management is made for a patient with more than one nodule detected in the CT scan. It is usually the largest nodule without clearly benign features.
Second-read AI (2^{nd}-read AI)	In this report, second-read (2^{nd}-read) AI refers to radiologist reading and analysing the CT scan image independently first, then bringing up and considering findings produced with AI-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 in a CT scan that is likely to represent a lung nodule to be marked and separated out from the rest of the image.
Sensitivity	The proportion of people who test positive for a disease amongst people who have the disease of interest. The ratio between the true positive value and (true positive value + false negative value). $$Sensitivity = \frac{a}{a+c} = \frac{TP}{TP+FN}$$
Specificity	The proportion of people who test negative for a disease amongst people who do not have the disease of interest. The ratio between the true negative value and (true negative value + false positive value). $$Specificity = \frac{d}{b+d} = \frac{TN}{TN+FP}$$
True negative value	The number of cases in which the index test has correctly indicated the patient as being disease-free. $$TN = d$$
True positive value	The number of cases in which the index test has correctly indicated the patient as having the disease. $$TP = a$$

List of abbreviations

Abbreviation	**Full term**
A&E	Accident and emergency
AI	Artificial intelligence
AUC	Area under the receiver operating curve

BTS	British Thoracic Society
CAD	Computer-aided detection
CASP	Critical Appraisal Skills Programme
CCC	Lin's concordance correlation coefficient
CEA	Cost effectiveness analysis
CEAC	Cost-effectiveness acceptability curve
CHEERS	Consolidated Health Economic Evaluation Reporting Standards
CI	Confidence interval
CRUK	Cancer Research UK
CT	Computed tomography
CV	Coefficient of variation
CXR	Chest X-ray
DAC	Diagnostic Advisory Committee
DAR	Diagnostic assessment report
DL	Deep learning
EAG	External Assessment Group
EBUS-TBNA	Endobronchial ultrasound-guided transbronchial needle aspiration
EUS-FNA	Endoscopic ultrasound-guided fine-needle aspiration
FBP	Filtered back projection
FN	False negative
FP	False positive
GGN	Ground glass nodule
HR	Hazard ratio
HSROC	Hierarchical summary receiver operating characteristic
HTA	Health technology assessment
ICC	Intraclass correlation coefficient
ICER	Incremental cost-effectiveness ratio
IQR	Interquartile range
K-LUCAS	Korean Lung Cancer Screening Project
KQ	Key question
LDCT	Low-dose computed tomography
LoA	Limits of agreement
LSUT	Lung Screen Uptake Trial
Lung-RADS	Lung CT Screening Reporting And Data System

LY	Life-years
MBIR	Model-based iterative reconstruction
MDT	Multi-disciplinary team
MIP	Maximum intensity projection
MPR	Multiplanar reformations
MRMC	Multi-reader, multi-case study
MRI	Magnet resonance imaging
NA	Not applicable
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NLST	National Lung Screening Trial
NPV	Negative predictive value
NR	Not reported
PACS	Picture archiving and communication system
PET	Positron emission tomography
PSN	Part-solid nodules
PSS	Personal Social Services
PPSRU	Personal Social Services Research Unit
PPV	Positive predictive value
QALY	Quality-adjusted life-years
RCT	Randomised controlled trial
ROC	Receiver operating characteristic
SD	Standard deviation
SDCT	Standard-dose computed tomography
SEM	Standard error of the measurement
SSN	Sub-solid nodules
TLHC	Targeted Lung Health Check
TN	True negative
TP	True positive
UK NSC	UK National Screening Committee
VDT	Volume doubling time
WTP	Willingness-to-pay

SCIENTIFIC SUMMARY

Background

Lung nodules are small rounded or irregular shaped growths with a diameter of 3 cm or less found inside the lung. They vary in size, which is strongly associated with risk of malignancy but in a nonlinear fashion. Most lung nodules on a computed tomography (CT) scan appear as solid structures, but some are sub-solid. Sub-solid nodules have either 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, some may be malignant or may develop into lung cancer.

Lung nodules are found when people are (1) referred for a CT scan that includes the chest because of signs and symptoms suggestive of lung cancer (symptomatic population), (2) investigated for other conditions unrelated to lung cancer (incidental population), or (3) through lung cancer screening programmes (screening population). People with previously identified lung nodules can also have CT scans as part of surveillance (surveillance population) 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 in the lung.

This diagnostics assessment focuses on detection and analysis of lung nodules in CT scan images that include the chest, assisted by computer software with artificial intelligence (AI)-derived algorithms.

Objectives

Key question 1 (KQ1)

What is the accuracy of CT image analysis assisted by AI software for automated detection and analysis of lung nodules in CT scans that include the chest obtained from symptomatic, incidental, screening or surveillance populations, 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 to CT surveillance or having biopsy/excision)?

Sub-questions

1. Does the accuracy of CT image analysis assisted by AI 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; (3) of solid nodules and sub-solid nodules; (4) obtained from people of different ethnic groups; (5) read by general radiologists/health professionals and specialised thoracic radiologists/health professionals; (6) by reason for CT can (for the incidental population)?

2. a) What is the concordance between readers with and without software support to detect and/or measure lung nodules from CT images?

b) What is the concordance between readers using different software to detect and/or measure lung nodules from CT images?

c) Does the use of software-assisted CT image analysis impact on intra-observer and interobserver variability in lung nodule detection and measurement?

Key question 2 (KQ2)

What are the benefits and harms of detection and analysis of lung nodules assisted by AI software compared with unassisted reading in CT scans that include the chest obtained from symptomatic, incidental, screening or surveillance populations?

Sub-questions 1-2 (see KQ1) are adapted for KQ2.

Key question 3 (KQ3)

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

Sub-questions 1-2 (see KQ1) are adapted for KQ3.

Methods

Data sources

KQ1 and KQ2

Medline; Embase; Cochrane Database of Systematic Reviews; Cochrane CENTRAL; Health Technology Assessment (HTA) database (CRD); International HTA database (INAHTA); Science Citation Index

Expanded (Web of Science); Conference Proceedings - Science (Web of Science) from 1 January 2012 to January 2022.

MedRxiv preprint server; clinical trials registries (via clinicaltrials.gov and the WHO ICTRP portal); websites of the technologies and their manufacturers; websites of selected organisations and conferences of interest; reference lists of included studies and relevant systematic reviews identified via the database searches; forwards citation tracking from key publications.

Company submissions were accepted until 31 August 2022.

KQ3

Medline; Embase; National Health Service Economic Evaluation Database (NHS EED) (CRD); Health Technology Assessment (HTA) database (CRD); International HTA database (INAHTA); Cost-Effectiveness Analysis (CEA) registry (Tufts Medical Center); EconPapers (Research Papers in Economics (RePEc)); ScHARRHUD; targeted web searches (Google); selected organisations and conferences of interest; reference lists of selected highly relevant papers.

Company submissions were accepted until 31 August 2022.

Eligibility criteria

Population: KQ1-3) 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; or people having CT surveillance for a previously identified lung nodule.

Intervention: KQ1-3) At least one of the 13 NICE specified technologies, used as reader support (focus of this assessment) or stand-alone (not formally included in the assessment but providing supplementary evidence):

AI-Rad Companion Chest CT (Siemens Healthineers); AVIEW LCS+ (Coreline Soft); ClearRead CT (Riverain Technologies); contextflow SEARCH Lung CT (contextflow); InferRead CT Lung (Infervision); JLD-01K (JLK Inc.); Lung AI (Arterys); Lung Nodule AI (Fujifilm); qCT-Lung (Qure.ai); SenseCare-Lung Pro (SenseTime); Veolity (MeVis); Veye Lung Nodules (Aidence); VUNO Med-LungCT AI (VUNO).

Comparator: KQ1) CT image assessment without AI-based software support or no comparator. KQ2/3) CT image assessment without AI-based software support.

Outcomes: KQ1) Accuracy for detecting nodules; accuracy for measuring the diameter or volume of nodule or change in volume (when the technologies are used as part of CT surveillance);

characteristics of detected nodules; 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; number of CT scans taken as part of CT surveillance; number of people having a biopsy or excision; number of cancers detected; stage of cancer at detection; time to diagnosis; acceptability and experience of using the AI software.

Sub-question 2: Concordance between readers with and without AI software; concordance between readers using different AI software; concordance between different AI software without human involvement; inter-observer variability; repeatability/reproducibility.

KQ2) Morbidity; mortality; health-related quality of life; patients' acceptability of use of the software.

KQ3) Cost effectiveness (e.g., incremental costs, incremental benefits, incremental cost effectiveness ratio, quality adjusted life years).

Study selection, data extraction and quality appraisal

KQ1 and KQ2

Two reviewers independently assessed articles for inclusion. A single reviewer extracted data, which were checked by a second reviewer. Two reviewers independently assessed methodological quality using the QUADAS-2 and QUADAS-C tools for included studies reporting test accuracy outcomes, or the COSMIN Risk of bias tool to assess the quality of studies on reliability and measurement error of outcome measurement.

KQ3

Two reviewers independently screened all titles and abstracts for potentially relevant records. Studies meeting inclusion criteria would be independently assessed using the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) and Philips criteria.

Data synthesis

KQ1-3) Narrative data synthesis was performed.

De-novo cost effectiveness analysis (CEA)

We adopted two different approaches to CEA. Firstly, a simple decision tree (the preliminary model) was constructed for assessing the cost-effectiveness of AI-assisted detection of actionable nodule compared with unassisted detection. As this model did not allow assessment of the impact of AI-assisted CT image analysis on nodule and disease management following detection throughout the clinical pathway, a de novo decision analytical structure (the full model) comprising two stages was developed in TreeAge (TreeAge Software Inc., Williamstown, MA, USA) to compare the cost-effectiveness of AI-assisted radiologist reading versus unaided radiologist reading. The first stage followed current practice for identifying lung nodules and classifying them based on morphology, nodule type and size. The first stage aimed to predict the impact of AI software-assisted CT reading on nodule detection and classification. The second stage followed the British Thoracic Society (BTS) guidelines and showed the pathways for patients with lung nodules who required CT surveillance. Outputs from the first stage were used as inputs for the second stage, and the associated costs and health outcomes from the comparative strategies were estimated. Both stages utilise decision tree structures.

Information required to populate the models included prevalence of lung nodules, prevalence of lung cancer, sensitivity and specificity for nodule detection, probabilities for nodule type and size distributions, resource use and costs and utilities. Where possible, parameterisation was driven by findings from the test accuracy review. This was supported by clinical expert opinion and simulations (also informed by the test accuracy review) specifically carried out to generate parameters otherwise not available. Given the paucity of information, many assumptions and simplification were required to link the initial impact of AI software assistance to longer term costs and health outcomes in the full model.

Resource use and costs for both models were obtained from review of the cost-effectiveness literature, National Health Service (NHS) reference cost schedule, companies via NICE and Personal Social Service Research Unit Costs. All costs were reported in 2020/21 prices and discounted at a rate of 3.5% per annum.

The model estimated the mean costs incurred and the benefits accrued associated with each strategy and assumed that the patients entering the model were aged 60 years. The results of the economic analysis are presented in the form of an incremental cost-effectiveness ratio (ICER). Cost per correct detection of a person with an actionable nodule was estimated in the preliminary model. The economic analysis for the full model was carried out based on a primary outcome measure of cost per QALY, and the perspective adopted was that of the NHS and Personal Social Services (PSS). Secondary outcome measures (cost per correct detection of a person with an actionable nodule, cost per cancer detected and treated) were also analysed in the full model.

A deterministic analysis was undertaken for the base-case results. Additionally, we undertook scenario analyses, univariate and probabilistic sensitivity analyses.

Results

KQ1

Twenty-seven studies evaluating eight of the 13 NICE-specified technologies were identified that reported outcomes on nodule detection or measurement accuracy/concordance, practical implications and/or impact on patient management. All studies were at high risk of bias and had multiple applicability concerns. Twenty-four studies used retrospective datasets, 17 of which compared the performance of readers with and without concurrent AI use (primary comparison of interest). Nine of these studies also allowed comparison with stand-alone software (outside the scope of this assessment, only as supplementary information). One study evaluated readers with concurrent AI software use only (versus a reference standard); five studies evaluated stand-alone AI software only; one further study compared stand-alone AI to unaided readers. Only three studies reported on prospective screening experiences based on the same screening pilot trial conducted in Korea: two studies reported on software-assisted reading only, whereas one study used a before-after design that evaluated the performance of radiologists before and after AI software implementation as well as stand-alone AI.

Accuracy and reliability

Detection of any nodules – All four identified studies directly comparing readers with and without concurrent software use found that AI assistance significantly increased sensitivities for detecting any nodules. Three studies reported specificity based on per-person analysis. Specificity decreased slightly in two studies while improved slightly in one study. The remaining study reported no difference in false positive (FP) rates with and without AI assistance.

Detection of actionable nodules - All three studies evaluating software with nodule detection function and directly comparing readers with and without concurrent software use found that AI assistance significantly increased the sensitivity for detecting actionable nodules (at least 5 mm diameter). Specificity was significantly lower and number of FP detections per image were significantly increased with AI use in one study. The other two studies also reported increased number of FP detections per scan, but no statistical test was performed.

Detection of malignant nodules – Three studies directly compared the sensitivity for detecting malignant nodules in readers with and without concurrent AI use. AI use significantly increased the sensitivity in two studies, with one of them also reporting lower specificity and higher number of FP detections per image. The remaining study only included one cancer case that was detected by both, readers with and without software use.

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 speciality.

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

The detection performance of radiologists (with and without concurrent software use, respectively) was not significantly different between standard dose and low dose CT scans (1 study).

Three studies which 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 evidence was insufficient to draw a firm conclusion.

Nodule type determination - Inter-reader agreement in nodule type determination was similar in readers with and without software use (2 studies).

Nodule size measurement – Nodule diameters were similar (2 studies) or significantly larger (2 studies) with semi-automatic measurements compared to manual measurements. A significant correlation between software-aided and manual measurements was observed (2 studies). Inter-reader variability (3 studies) and intra-reader variability (1 study) in nodule size measurement was significantly reduced in readers with software use compared to manual measurement.

Classification into risk categories based on nodule type and size - Regarding all four possible nodule management recommendation categories based on the BTS guidelines, the AI-assisted readings of each radiologist showed a higher agreement with the consensus session (reference standard) than when readings were done unaided (1 study). Inter-reader agreement in risk category classification based on BTS (1 study), Lung-RADS (2 studies) and Fleischner (1 study) was consistently improved with concurrent software use. One study also reported reduced intra-reader variability in Fleischner risk categorisation with software use.

Whole read (detection and Lung-RADS categorisation) - One before-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 sensitivity, specificity, positive (PPV) and negative predictive values was observed between periods before and after software implementation. PPVs differed significantly according to measurement planes (transverse, maximum orthogonal, any maximum).

Nodule growth – No study provided data for primary comparison of interest. The sensitivity of stand-alone software to detect nodule pairs in subsequent scans of the same patient was 100.0% (23/23) with no FP-pairs (1 study). The mean growth percentage discrepancy was similar between unaided chest radiologists and stand-alone software (1 study). However, a single incorrect segmentation of the stand-alone software was observed, resulting in a twice as high upper end of its confidence interval compared to that of radiologists. Therefore, the study advises visual verification of the nodule segmentation by human readers.

Practical implications

Segmentation failure ranged from 0% to 57% of nodules (8 studies). However, the observed nodule segmentation failure might be mostly due to rejections of segmentation results by radiologists, rather than the inability of the system to segment the nodule. Failure rates seem to be higher in ground glass nodules (34%) and part-solid nodules (20%) compared to solid nodules (7%) (1 study). Manual modifications of the segmentation were required in 29% to 59% of nodules (2 studies).

Radiologist reading time was reduced with concurrent software use by 11.3%-78% compared to unaided reading (9 studies) but increased with 2nd-read software assistance (+26%, 1 study). When using a software with only vessel suppression function, the reading time was similar with and without software (1 study).

Impact on patient management

Characteristics of detected nodules: Regarding all detected nodules (true and false positives), the proportion of solid nodules was lower with concurrent software use compared to unaided reading (87.1% vs 90.6%) (1 study). Additional true positive nodules detected with AI use were 56-57% solid, due to larger improvements in the detection of sub-solid nodules (2 studies). Twenty-two percent of additional true positive nodules were 6 mm or larger (1 study).

Proportion of detected nodules that are malignant: The proportion of detected actionable nodules that are malignant was lower with software use (2 studies).

Impact of test result on decision making: With software use, readers tended to upstage Lung-RADS (3 studies) or Fleischner risk categories (1 study) rather than downstage.

Number of people having CT surveillance: The proportion of people classed as Lung-RADS category 3 or 4A increased with AI use (2 studies).

Number of people having biopsy or excision: Similar (1 study) or slightly higher (1 study) proportions of people were classed as Lung-RADS category 4B/4X.

Time to diagnosis: One retrospective study showed that substantial management discrepancies for lung cancer cases (Lung-RADS category 1/2 vs. 4A/B) between readers would be reduced by half and sensitivity would be improved with AI software use, which might translate into earlier diagnosis if confirmed in clinical practice.

KQ2

No studies were identified that reported on patient benefits and harms of AI-based software use compared to current CT reading practice without AI-based software use.

KQ3

Of the 1,988 records identified, 15 were considered potentially relevant and were reviewed at full text. All studies were excluded at full text. Two potentially relevant model-based economic analyses did not meet our inclusion requirements but were summarised, as these studies provided some evidence relevant for cost-effectiveness analysis.

De-novo cost effectiveness analysis (CEA)

Due to the complete absence of evidence related to clinical effectiveness, and substantial challenges in linking test accuracy evidence to clinical and economic outcomes, methods and findings presented here are highly uncertain and should be regarded as early indications and frameworks for future analyses when new evidence becomes available. Our preliminary model suggested 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 for the symptomatic and incidental populations, AI-assisted CT image analysis dominates the 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 the unaided reader. This is driven by the costs and disutilities associated with false positive results and CT surveillance. AI was deemed cost-effective for both symptomatic and

incidental populations in the scenario analyses where disutility associated with false positive results and CT surveillance were removed. 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 estimates of improved test specificity for AI. Although there was more data available to populate the screening population model there was very great uncertainty across all models.

Conclusions

Evidence from studies included in this diagnostic assessment shows that AI-assisted detection and analysis of lung nodules increase the consistency in nodule measurement and in risk classification according to clinical guidelines compared with unaided reading. The studies also suggest that AI assistance improves the sensitivity for lung nodule and cancer detection, and is often (but not always) accompanied by a decrease in specificity and an increase in false positive findings per scan, as well as pushing up risk categorisation of nodules based on clinical guidelines. The reported performance of AI-assisted reading varies substantially among published studies, possibly attributed to heterogeneous study populations, reader experience, speciality and reading conditions, other study design features and risk of bias in addition to potential differences in the performance of individual technologies.

No studies were identified that directly compared the performance of different AI software (and analysis of CT scan image assisted by them). Given the paucity of evidence, it is currently not possible to reliably establish the cost-effectiveness of AI-assisted reading compared with unaided reading, and the relative effectiveness and cost-effectiveness of strategies adopting different AI software to assist nodule detection and analysis.

Published studies have largely been conducted retrospectively in a research environment. The vast majority of studies identified in this diagnostic assessment report were judged to be at high risk of bias and have multiple applicability concerns for the UK settings. No studies evaluating intermediate clinical process and downstream clinical outcomes were identified. Further prospective studies of AI-based software that incorporate clinical process and outcome measures and that are undertaken in clinical practice settings are required.

PLAIN ENGLISH SUMMARY

Lung cancer is one of the most common types of cancer in the UK. People in the early stages of the disease may not have symptoms; therefore, lung cancer is often diagnosed late. Lung nodules are small (≤3 cm) abnormal areas of tissue in the lung. Most of the nodules are harmless, but some of them could be lung cancer.

Computed tomography (CT) is an imaging technology that doctors use to find lung nodules in the chest. At present, most healthcare professionals detect lung nodules on CT scan images without assistance from any computer software. After a nodule is found, its size needs to be measured (sometimes repeatedly over time to check if it grows) as this information helps doctors assess its cancer risk and decide what to do. Nevertheless, detecting and measuring nodules in CT images can be difficult for various reasons.

Computer software developed using methods that enable it to 'learn' from data and carry out tasks that is often done by humans (this is called artificial intelligence, or AI) could help health professionals detect nodules that might have been overlooked otherwise and may measure their size more consistently and quickly. However, the AI software may also detect more nodules that are harmless and cause unnecessary anxiety and investigations.

This report looked at the evidence on how good AI software is at helping healthcare professionals to find and measure lung nodules. The report also investigated benefits and harms of using such software and whether it offers value for money. This report covered people who had a CT scan that includes the chest due to symptoms suggestive of lung cancer, unrelated to lung cancer suspicion (e.g. after an accident), for lung cancer screening or as follow-up of a previously detected nodule.

We did not find any studies that compared radiologists (doctors who specialise in interpreting scans) with and without software use in clinical practice in the UK. Most identified studies were of low quality, and CT image assessment was performed retrospectively and was not affecting patient management. Findings from these low-quality studies suggest that:

- Software use could improve nodule detection, with bigger improvements seen for small and medium-size nodules.
- Software use might increase the number of false positive detections.
- Detection performance seems to be similar in standard and low dose CT scans.
- With software use, nodule size measurement as well as resulting cancer risk classifications could be more consistent between radiologists.

- Automatic nodule size measurement might fail or be deemed as unreliable by the radiologist in up to half of nodules.
- Radiologists reading time could be reduced with software use.

Our analysis shows that the use of AI software allows radiologists to identify more lung nodules that they should keep an eye on and detect more lung cancers. For lung cancer screening, we estimate that it is cost-effective, because it may be more accurate than humans alone and may cost less overall. For other groups of people including people with symptoms of lung cancer and people having a CT scan for other reasons, we estimate that it is not cost-effective. This is because it may harm more people with incorrect test results and unnecessary regular surveillance testing, which can be worrying and costs the NHS money.

1 BACKGROUND AND DEFINITION OF THE DECISION PROBLEM(S)

1.1 Lung nodules and lung cancer

Lung nodules are small rounded or irregular shaped growths with a diameter of 3 cm or less that are found inside the lung. They vary in size, which is strongly associated with risk of malignancy but in a nonlinear fashion.[1] A nodule with a diameter of less than 3 mm is referred to as a micronodule, the measurement of which is not recommended due to accuracy limitations.[2] Lung nodules with a diameter smaller than 5 mm have low probability of being lung cancer[3] and do not usually require further actions if they are detected incidentally. Unless otherwise stated, in this report we refer to nodules with a diameter of at least 5 mm as 'actionable nodules'.

Most lung nodules on a computed tomography (CT) scan appear as solid structures, but some are sub-solid. Sub-solid nodules have either 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 when people are (1) referred for a CT scan that includes the chest because of signs and symptoms suggestive of lung cancer, (2) 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 in the lung.

Lung cancer is one of the most common types of cancer in the UK. Its incidence rises steeply from around age 45-49.[4] 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, more than 65% of all 39,267 lung cancers in England were diagnosed at stage III (n=7,886) or 4 (n=18,104).[5] The NHS Long Term Plan sets out an ambitious target of diagnosing 75% of all cancers at an earlier stage, stages I or II, by 2028.[6]

While 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, identification and monitoring lung nodules using CT scans of the chest remain the primary means of detecting lung cancer at earlier stages.

1.2 Diagnostic and care pathway

1.2.1 Pathway to CT scan due to signs and symptoms suggestive of lung cancer

The identification of people with signs and symptoms suggestive of lung cancer often happens in primary care. The NICE guideline on recognition and referral for suspected cancer[7] recommends that people aged 40 and over 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 or
- appetite loss.

An urgent chest X-ray should also be considered for people aged 40 or over 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, referral to secondary care should be made using a suspected cancer pathway referral for an appointment within two weeks. During scoping, clinical experts noted if the X-ray findings do not show abnormalities but an ongoing suspicion of lung cancer remains, referral to secondary care for a CT scan may also be made. People aged 40 or over 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 direct access to CT where this is available for primary care.

In secondary care, people with known or suspected lung cancer should be offered a contrast-enhanced chest CT scan to further the diagnosis and stage the disease (NICE guideline on diagnosis and management of lung cancer).[8]

1.2.2 Lung cancer screening

In September 2022, the UK National Screening Committee (UK NSC) recommended targeted lung cancer screening for people aged 55 to 74 years identified as being at high risk of lung cancer.[9] NHS England are evaluating the Targeted Lung Health Check programme (TLHC) in some areas of

England,[10] which provides a feasible and effective starting point for the implementation of a targeted screening programme in England. In this programme, people aged over 55 years but less than 75 years who have ever smoked are invited to a lung health check. The lung health check involves collecting information about lung health, lifestyle and family and medical history, and measuring height and weight. Following the lung health check, people assessed as being at high risk of lung cancer are offered a low-dose CT 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.

1.2.3 Initial assessment and CT 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. solid or sub-solid).[11] The guideline recommends the same diagnostic approach for nodules detected incidentally, symptomatically, or through screening. The guideline recommendations are for lung nodules in adults. During scoping, clinical experts explained that lung nodules in children are very rarely malignant; therefore, lung nodules in children are not currently routinely investigated to avoid unnecessary CT scans.

Figure 1 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.[12] The nodule size (in diameter) and the number of nodules detected are among the inputs to this multivariable prediction model.[13]

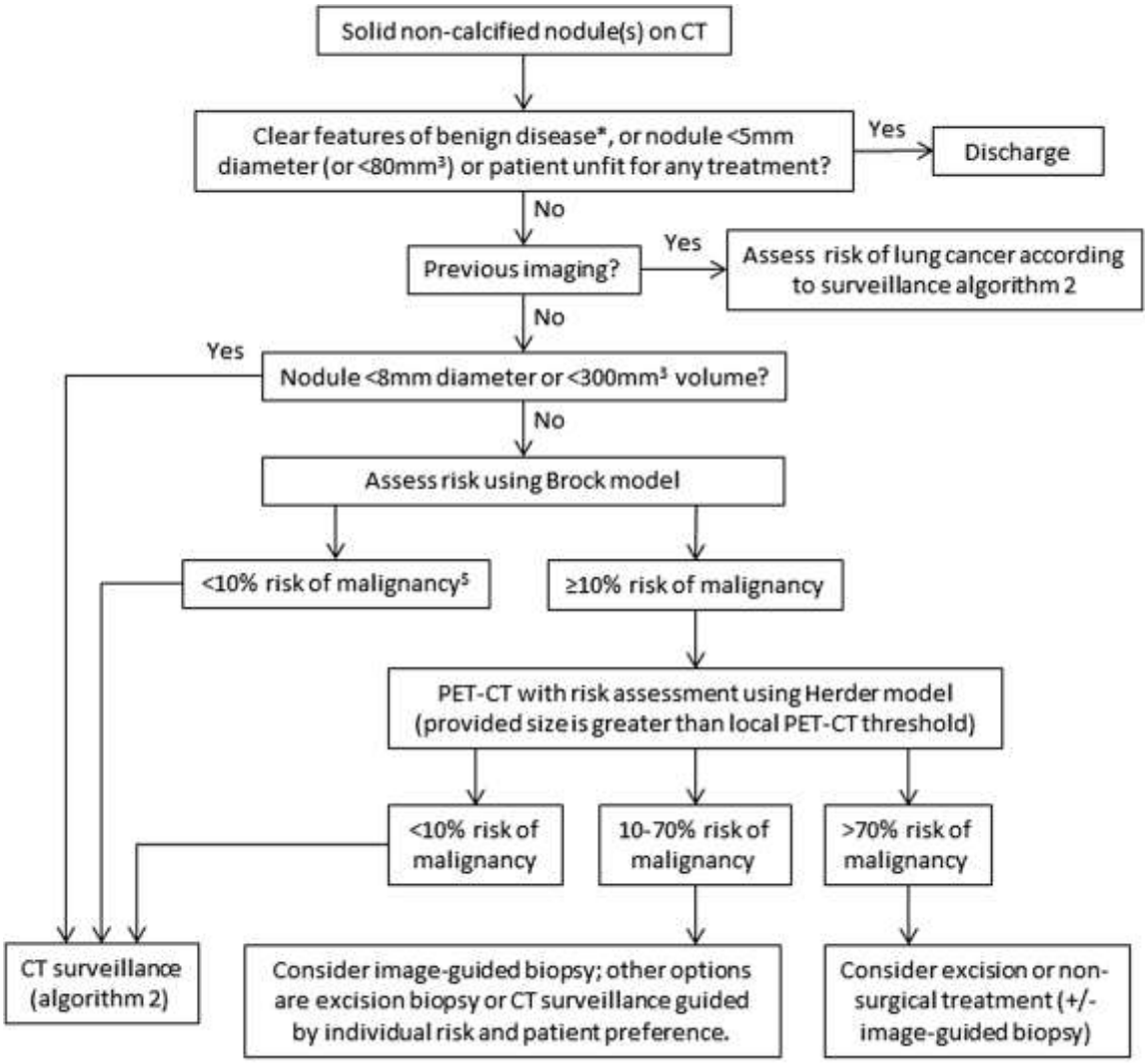


Figure 1. Initial assessment of solid lung nodules (reproduced with permission from Callister et al. 2015)[11]

* Some nodules seen may be attached to or very near the lining of the lungs (perifissural nodules), these are often pulmonary lymph nodes.

The initial assessment of sub-solid nodules (part-solid and ground glass) follows a similar pathway (see **Figure 2**). But because these nodules can sometimes disappear on their own, the pathway involves a repeat CT scan at 3 months before the use of the Brock malignancy risk model. Herder model[14] is not used for sub-solid nodules.

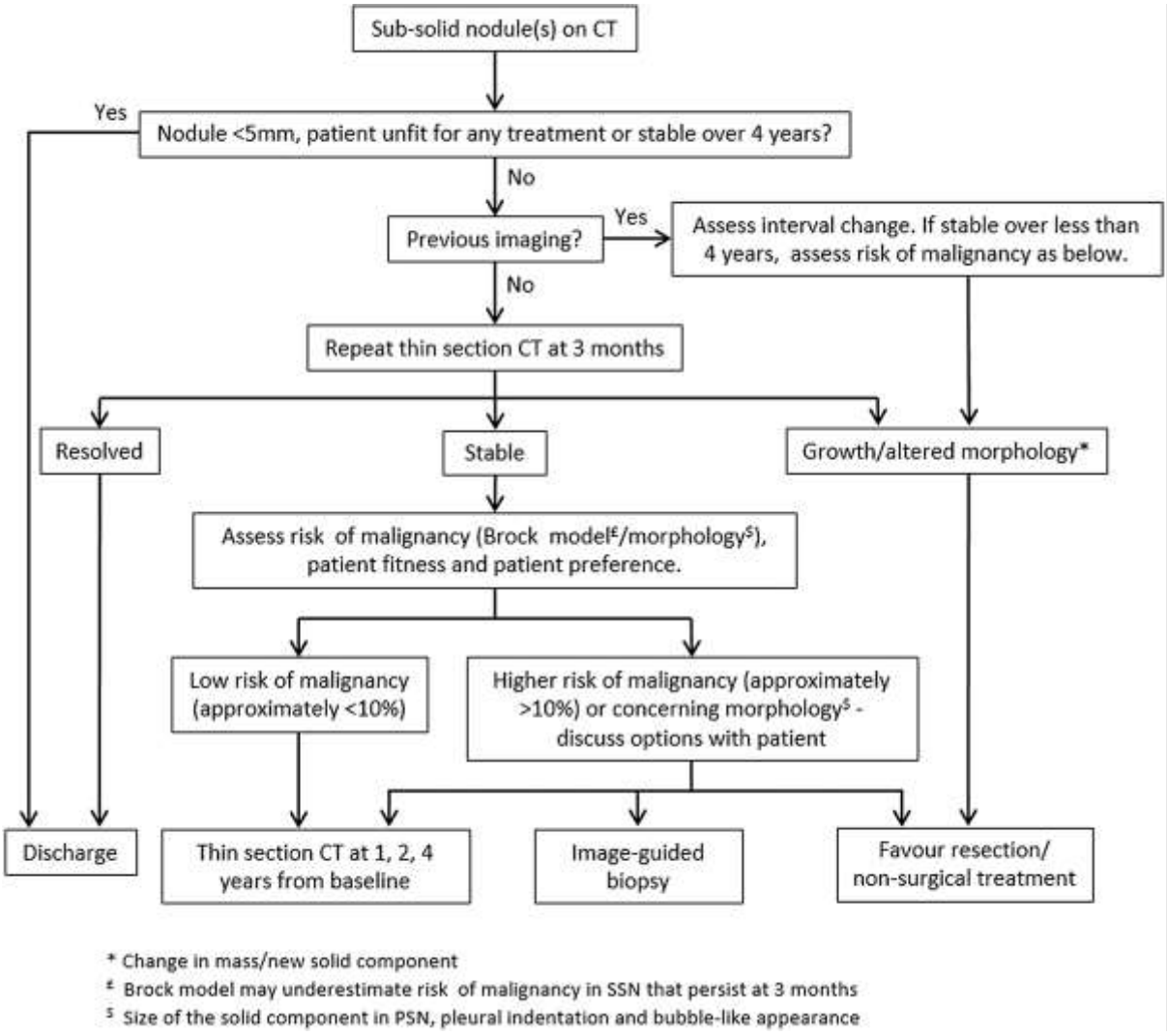


Figure 2. Sub-solid pulmonary nodules algorithm (reproduced with permission from Callister et al. 2015)[11]

PSNs, part-solid nodules; SSN, sub-solid nodules.

Figure 3 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 discriminate between benign and malignant nodules. The nodule growth should be assessed by estimating its volume doubling time (VDT). The surveillance period for sub-solid nodules is longer (4 years) than for solid nodules (one year with volume and two years with diameter measurements).

The BTS guidelines are currently being updated.[15]

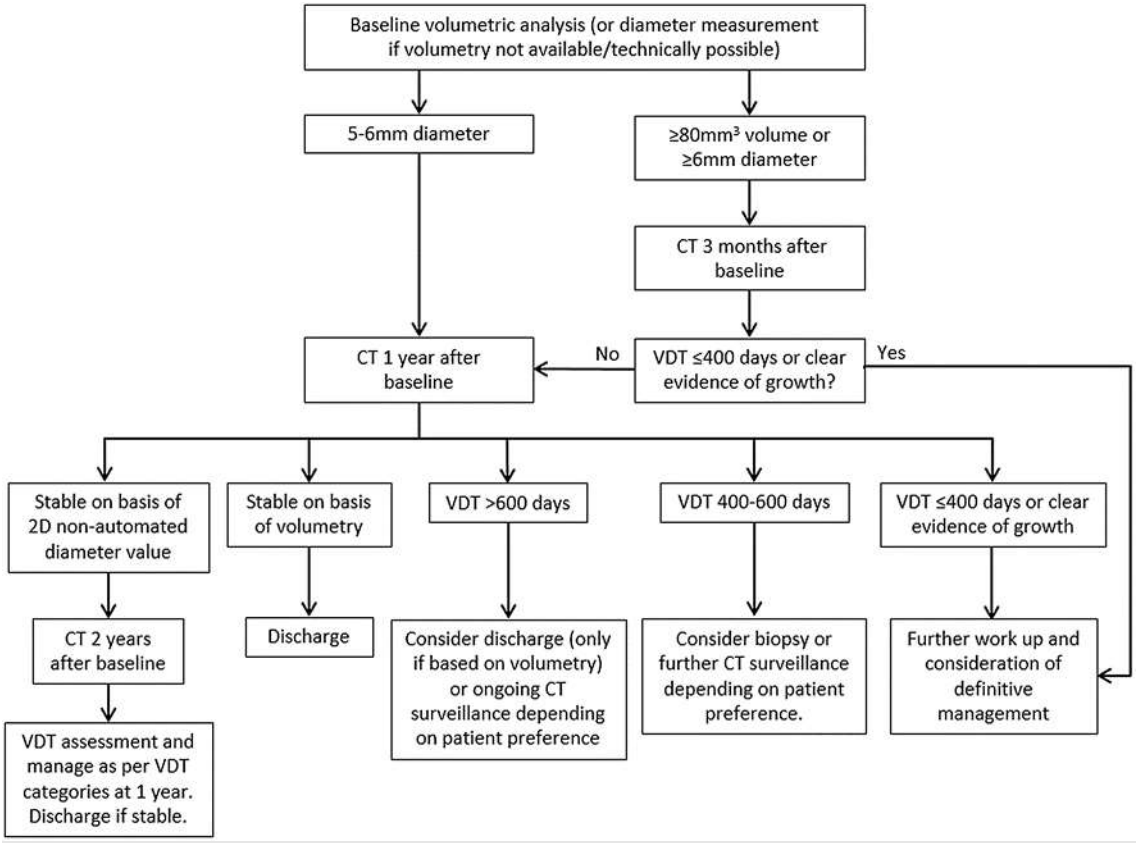


Figure 3. CT surveillance of solid lung nodules (reproduced with permission from Callister et al. 2015)[11]

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 clinical management of lung nodules identified by screening programmes,[16] and it was adopted in some of the studies assessed in this report. Lung-RADS allows categorisation of nodules according to their sizes and features into various categories 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 & 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 compared with the BTS guideline;[11] 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).

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

Currently, in routine clinical practice in the UK, radiologists or other healthcare professionals such as radiographers detect lung nodules on chest CT scan images without assistance from any software. 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 & Emergency [A&E] 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 or radiologists who yearly, as part of their normal clinical practice, report more than 500 thoracic CT scans of which a significant proportion are lung cancer CT scans.[17] 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.[18]

The 2015 BTS guidelines for the investigation and management of pulmonary nodules recommend that the size of an identified nodule is quantified as the volume of the nodule.[11] To do this, volumetry software needs to be used. In current practice, software is often part of the picture archiving and communication system (PACS), or a module on 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 of comparing 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 dates of the two scans into the BTS Pulmonary Nodule Risk Calculator.[13] 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.[13] 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 upon detection and analysis of pulmonary nodule in a number of ways as shown in **Figure 4** below.

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 **Section 3.3.1**; incorporated as a parameter for the economic model (**Section 7.4.3**).

(2) Accuracy in classification of nodule type: Evidence presented in **Section 3.3.2**; not included in the economic model as no clear evidence of an impact by AI.

(3) Accuracy and precision in measuring nodule size/volume: Evidence presented in **Sections 3.3.3** and **3.3.4**; incorporated into the model through simulation output (**Section 7.4** and **Appendix 7**).

(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 **Section 3.4**; incorporated into the economic model through simulation output (**Section 7.4** and **Appendix 7**)

(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 **Section 3.5.2**; incorporated as a parameter for the economic model.

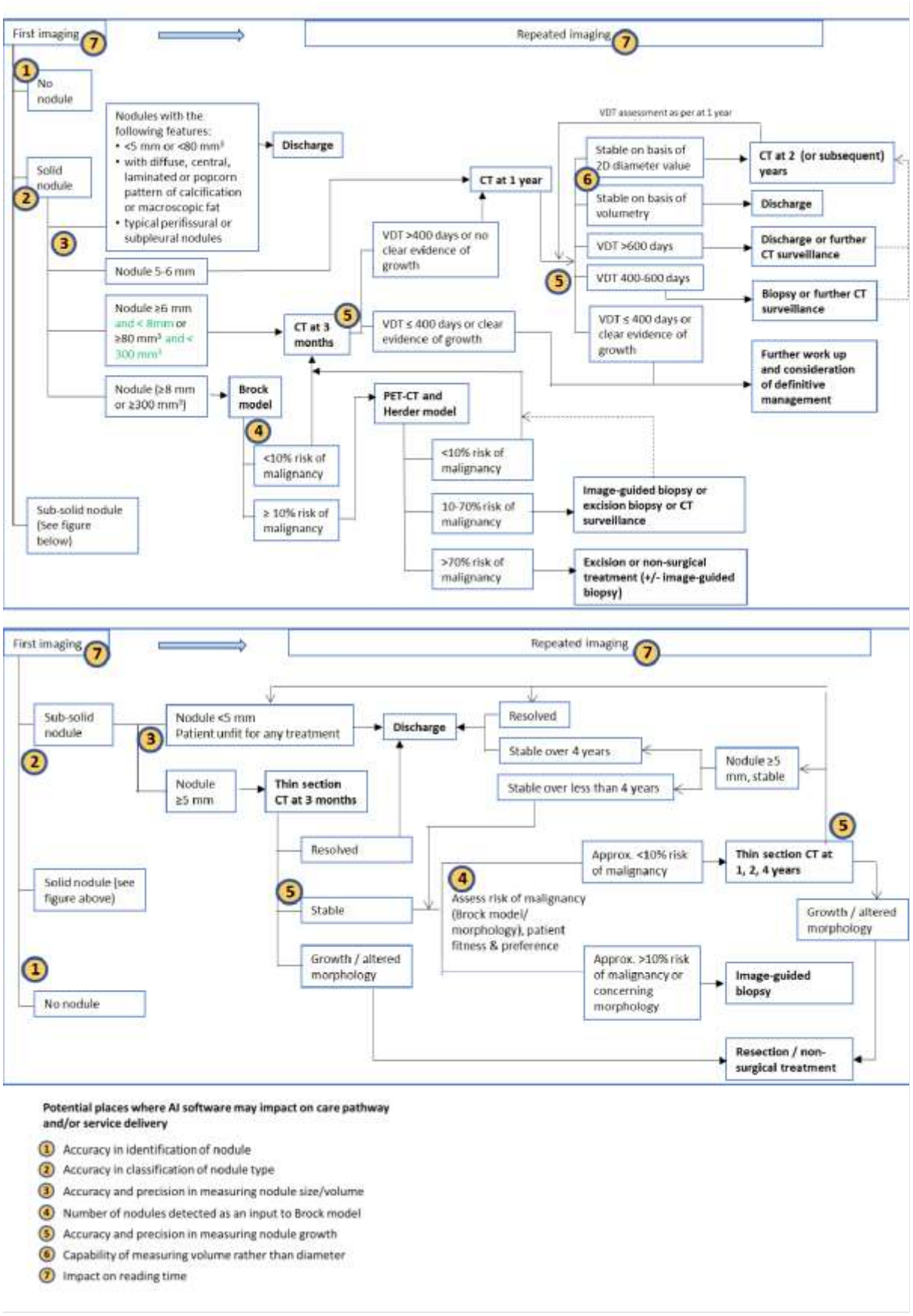


Figure 4. Points at which AI derived software may have an impact in the process of nodule detection and analysis and relevant evidence in this report

1.2.5 Diagnosis and staging of lung cancer

To guide the treatment of lung cancer, information about type and spread of the lung cancer (stage) are needed. The NICE guideline on diagnosis and management of lung cancer[8] recommends choosing investigations that give the most information about diagnosis and staging with the least 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-CT (PET-CT) scan and magnetic resonance imaging (MRI). Tissue diagnosis is often obtained by image-guided biopsy, endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) and endoscopic ultrasound-guided fine-needle aspiration (EUS-FNA), respectively.

1.2.6 Treatment for lung cancer

After diagnosis, treatment for lung cancer is based on several factors, such as 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 (NICE guideline on diagnosis and management of lung cancer[8]).

1.3 Population and relevant subgroups

This diagnostic assessment included people who have 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 who 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).

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

Use of the technologies for cancer staging and cancer 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 sub-solid nodules;
- Characteristics of the reader: General radiologist (or other healthcare professional) versus radiologist (or other healthcare professional) with thoracic speciality;
- Within the incidental population: different reasons for the CT scan.

1.4 Description of technology(ies) under assessment

This diagnostics assessment focuses on the use of computer software with artificial intelligence (AI)-derived algorithms for automated detection and analysis of lung nodules from CT scan images that include 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."[19] The technologies included in this diagnostic assessment were defined by the NICE final scope and comprise computer software that has been developed in a process that involves learning from data to detect and analyse lung nodules in 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, 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 (for example scans without contrast or low-dose CT) or in specified populations (for example people without symptoms suggestive of lung cancer or people aged 18 or older).

Thirteen relevant technologies have been identified by the NICE. The section below describes the specific technologies included in this assessment. The descriptions as well as **Table 1** are reproduced from the final scope issued by NICE.

1.4.1 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 and over. The software integrates with the PACS.

1.4.2 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 sub-solid nodules in low-dose chest CT scans. AVIEW LCS+ is indicated for use in adults. Other indications for use include 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.

1.4.3 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 sub-solid 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 and over 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 training of radiologists on how to use ClearRead CT is usually done within a day. The software is commercially available to the NHS directly from the manufacturer and through partner organisations.

1.4.4 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 sub-solid 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.

1.4.5 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 sub-solid 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 the use is recommended in people aged 18 and over. 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.[11] 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.

1.4.6 JLD-01K (JLK Inc.)

JLD-01K is a CE-marked (class I medical device) software. It can detect and measure solid and sub-solid 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.

1.4.7 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 sub-solid 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.

1.4.8 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.

1.4.9 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 and over. The software is compatible with slice thickness of up to 6 mm. qCT-Lung integrates with PACS.

1.4.10 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 sub-solid 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.

1.4.11 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 use in asymptomatic populations. 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 sub-solid nodules. Usually, 2 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[11] and integrates with the PACS. The company states that usually 4 to 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.

Superseded-
see erratum

1.4.12 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 sub-solid 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 and over. 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.[11] 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.

1.4.13 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 sub-solid lung nodules in low-dose chest CT scans. It is intended for use in lung cancer screening populations. The software integrates with PACS.

Table 1. Summary of the included technologies (reproduced from final NICE scope)

Product name (manufacturer)	CE mark	Available to the NHS	CT scan types	Detection	Volumetry
AI-Rad Companion Chest CT (Siemens)	Class IIa *	To be confirmed	Low dose, regular dose with and without contrast *	Yes *	Yes *
AVIEW LCS+ (Coreline Soft)	Class IIa *	Yes	Low dose *	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
JLD-01K (JLK Inc.)	Class I	To be confirmed	Without contrast	Yes	Yes
Lung AI (Arterys)	Class IIa *	To be confirmed	Low dose, regular dose with and without contrast *	Yes *	Yes *
Lung Nodule AI (Fujifilm)	To be confirmed	To be confirmed	To be confirmed	Yes	Yes
qCT-Lung (Qure.ai)	Class I *	To be confirmed	Without contrast *	Yes *	Research only *
SenseCare-Lung Pro (SenseTime)	Class IIb *	To be confirmed	Without contrast *	Yes *	Yes *
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 *	To be confirmed	Low dose *	Yes *	Yes *

* Information only from public domain.

1.5 Proposed position of the intervention in the diagnostic pathway

Figure 5 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 chest CT where pulmonary nodules are identified (a.). After nodules are detected, the nodule management pathway according to the 2015 BTS guidelines[11] depends on two main criteria: nodule type (solid or sub-solid;

c.) and nodule size (diameter or volume; d.). Depending on the predicted malignancy risk (e.), the guidelines recommend discharge, CT surveillance or further work-up and treatment.

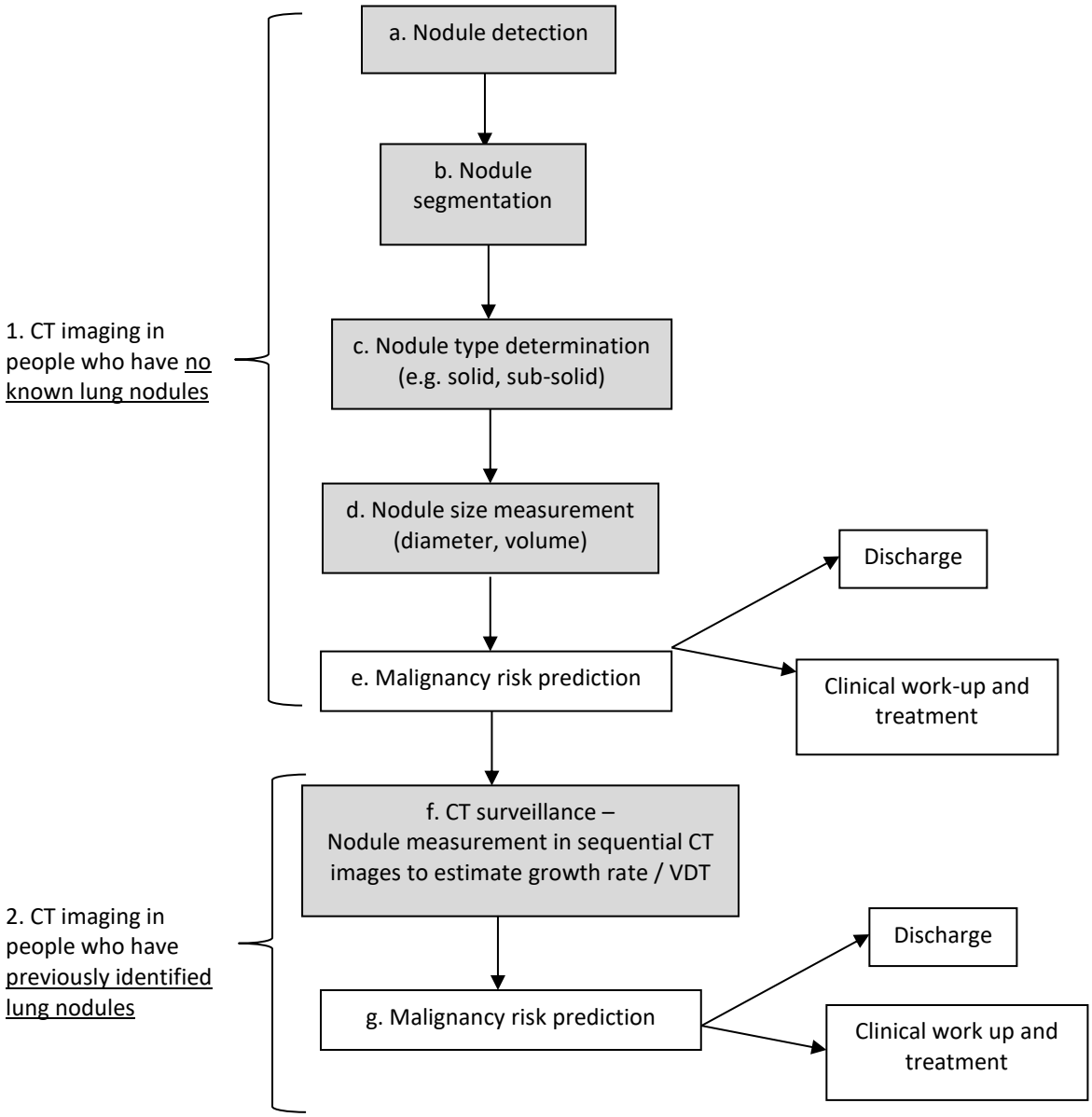


Figure 5. Proposed roles of the intervention in the process of diagnosing lung cancer

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 and make decision on further patient management (i.e. discharge, further CT surveillance or 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 reviewing 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 which 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 in 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 differ in confidence to overrule software depending on their experience and speciality (e.g. thoracic radiologists vs general radiologists).

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 (highlighted in grey in **Figure 5**):

1. In CT images from people without previously identified lung nodules (use case 1)
    a. Nodule detection;
    b. Nodule segmentation;
    c. Nodule type determination (solid or sub-solid);
    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.

1.6 Comparators

The comparator for this diagnostic assessment is review of chest CT scan images by a radiologist or another healthcare professional (such as a radiographer) without software for 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 **section 1.3.4**), 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 from general, trauma or thoracic radiologists. They further commented that the level of 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 **section 1.3.4**, the standard protocol for the TLHC programme in England stipulates specific requirements for specialised readers reviewing the CT scans in the programme.[17]

1.7 Outcomes

Key outcomes judged to be relevant to the assessment of the clinical 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 sections **2.1.2, 4.1.1.2** and **5.1.1.2**). In short, clinical effectiveness outcomes included: 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 and quality-adjusted life years. 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.

1.8 Objectives

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

The key questions for this diagnostic assessment report (DAR) are provided in the box below.

Key question 1

What is the accuracy of CT image analysis assisted by software for automated detection and analysis of lung nodules in people undergoing CT scans that include 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 to CT surveillance or having biopsy/excision)?

Sub-questions

1. Does the accuracy of CT image analysis assisted by 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; (3) of solid nodules and sub-solid nodules?
2. Does the accuracy of CT image analysis assisted by software for automated detection and analysis of lung nodules, its practical implications and impact on patient management differ by patients' ethnicity?
3. Does the accuracy of CT image analysis assisted by 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 CT image analysis assisted by software for 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 software support to detect and/or measure lung nodules from CT images?

   b) What is the concordance between readers using different software to detect and/or measure lung nodules from CT images?

   c) Does the use of software-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 software for automated detection and analysis of lung nodules from CT images compared with unassisted CT image analysis in people undergoing CT scans that include 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?

Sub-questions

1. Do the benefits and harms of CT image analysis assisted by software for 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; (3) of solid nodules and sub-solid nodules?
2. Do the benefits and harms of CT image analysis assisted by software for automated detection and analysis of lung nodules differ by patients' ethnicity?
3. Do the benefits and harms of CT image analysis assisted by software for 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 software for automated detection and analysis of lung nodules differ by reason for chest CT scan?

Key question 3

What is the cost-effectiveness of using software for automated detection and analysis of lung nodules from CT images compared with unassisted CT image analysis in people undergoing CT scans that include 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?

Sub-questions

1. Does the cost-effectiveness of CT image analysis assisted by software for 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; (3) of solid nodules and sub-solid nodules?
2. Does the cost-effectiveness of CT image analysis assisted by software for automated detection and analysis of lung nodules differ by patients' ethnicity?
3. Does the cost-effectiveness of CT image analysis assisted by software for 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 software for automated detection and analysis of lung nodules differ by reason for CT scan?

Ideally, priority of the assessment 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.[20] However, as no 'end-to-end studies' were found, we included and evaluated studies on test accuracy and practical implications, clinical effectiveness, costs and cost-effectiveness separately, and then synthesised the evidence using a linked evidence approach.[20]

2 SYSTEMATIC REVIEW OF ASSESSING 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 methods described below. The review followed the principles outlined in the Cochrane Handbook of Diagnostic Test Accuracy[21] and the NICE Diagnostic Assessment Programme manual.[20]

2.1 Identification and selection of studies

2.1.1 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 are likely to be difficult to assess. No date limits were applied. The draft MEDLINE search strategy was checked by an Information Specialist not otherwise involved in the project for any omissions or errors and adapted for the other databases. The final search strategies for all sources are provided in **Appendix 1: Literature search strategies.**

Systematic searches were conducted in January 2022 in the following databases:

MEDLINE All (via Ovid), Embase (Ovid), Cochrane Database of Systematic Reviews (Wiley), Cochrane CENTRAL (Wiley), Health Technology Assessment (HTA) database (CRD), International HTA database (INAHTA), Science Citation Index Expanded (Web of Science), Conference Proceedings - Science (Web of Science).

Records were exported to EndNote X9.3, where duplicates were systematically identified and removed.

In order to capture unpublished or ongoing studies, searches of MedRxiv preprint server (via the medrxivr app) and clinical trials registries (via clinicaltrials.gov and the WHO ICTRP portal) were undertaken. The trials registry searches were highly focussed, including search terms for the specific technologies of interest listed in the project scope, and their manufacturing companies. Websites of the technologies and their manufacturers were also checked for further information, as were websites of selected organisations and conferences of interest (see **Appendix 1: Literature search**

strategies for details). 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.

2.1.2 Study eligibility criteria

Studies that satisfied the following criteria were included:

Population	<u>All questions</u> 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 permits, the following subgroups may be considered: • Patient's ethnicity; • People who have a CT scan: (1) with or without contrast; (2) using a low-dose or a standard dose; (3) of solid nodules or sub-solid nodules; • For the incidental population, by reason for CT scan.
Target condition	<u>All questions</u> Lung nodules or lung cancer
Intervention	<u>All questions</u> CT scan review by a radiologist or another healthcare professional using any of the following software for automated detection and analysis of lung nodules: • AI-Rad Companion Chest CT (Siemens Healthineers) • AVIEW LCS+ (Coreline Soft) • ClearRead CT (Riverain Technologies)* • contextflow SEARCH Lung CT (contextflow)** • InferRead CT Lung (Infervision)* • JLD-01K (JLK Inc.) • Lung AI (Arterys) • Lung Nodule AI (Fujifilm) • qCT-Lung (Qure.ai) • SenseCare-Lung Pro (SenseTime) • Veolity (MeVis)* • Veye Lung Nodules (Aidence) • VUNO Med-LungCT AI (VUNO) * Indication for use specifies use in asymptomatic population, therefore the software cannot be assessed in symptomatic population. ** Indication for use specifies use in symptomatic population, therefore the software cannot be assessed in incidental or screening populations. Please note: specific indications for use for some of the technologies are unclear because only information in the public domain was available.

	Evidence on the performance of software alone (without review by a radiologist or other trained reader) will be included with applicability concerns highlighted. Where data permits, the following subgroups may be considered: - General radiologist/other healthcare professional with software support versus radiologist/other healthcare professional with thoracic speciality with software support.
Comparator	<u>All questions</u> CT scan review by a radiologist or another healthcare professional without software for automated detection and analysis of lung nodules (using diameter or volume to measure nodule size) or no comparator. Where data permits, the following subgroups may be considered: - General radiologist/other healthcare professional without software support versus radiologist/other healthcare professional with thoracic speciality without software support.
Reference standard	<u>Key question 1 and sub-questions 1-4</u> • 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).
Outcomes	<u>Key question 1 and sub-questions 1-4.</u> • 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 by judgement of experienced radiologist(s) and the accuracy to detect malignant nodules, respectively); • Accuracy to assess 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 number of unnecessary CT surveillance scans due to false positive nodules); • Number of people having a biopsy or excision (this may also include for example the number of people having a negative biopsy due to false positive nodules); • Number of cancers detected; • Stage of cancer at detection; • Time to diagnosis; • Acceptability and experience of using the software. <u>Sub-question 5.</u> • Concordance between readers with and without software; • Concordance between readers using different software; • Concordance between different software without human involvement; • Inter-observer variability (e.g. positive and negative agreement, Cohen's kappa);

	• Repeatability/reproducibility.
Study design	<u>All questions</u> • Prospective test accuracy studies; • Retrospective test accuracy studies; • Randomised controlled trials; • Cohort studies; • Historically controlled trials; • Before-after studies; • Retrospective multi-reader multi-case studies; • Qualitative studies for user experience/acceptability.
Publication type	<u>All questions</u> • 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	<u>All questions</u> English

Papers that fulfil the following criteria were excluded:

- Studies using PET-CT scan images, lung phantom images or where more than 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 will be excluded unless they report outcome data that have not been reported elsewhere, in which case they will 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 to confirm that no evidence relevant to their technology under investigation has been published before 2012.

2.1.3 Study screening and selection

Two reviewers (JG/AA) independently screened the titles and abstracts of records identified by 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 through consensus, with the inclusion of a third reviewer (CS, YFC) when required. Records that

were excluded at full text stage were documented, including the reasons for their exclusion (see **Appendix 2, Table 64** and **Table 65).**

2.2 Data extraction and risk of bias assessment

2.2.1 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 (**Appendix 3**). Any disagreements were resolved through consensus, with the inclusion of a third reviewer (CS, YFC) when required.

2.2.2 Assessment of study risk of bias

The risk of bias of test accuracy studies was assessed using a modified QUADAS-2 tool[22] combined with the QUADAS-C tool for comparative studies.[23] The 'COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument' was used to assess the risk of bias for 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,[24] and for studies of agreements/concordance between readers where a reference standard could not be defined. Quality appraisal tools used in this DAR are tailored for the specific topic and are provided in **Appendix 4.** Two reviewers (JG/AA) independently undertook risks of bias assessment and critical appraisal. Disagreements were resolved through consensus, with the inclusion of a third reviewer (CS, YFC) if required.

2.3 Methods of analysis/synthesis

Test accuracy results are firstly grouped by different software functionality, e.g. nodule detection, classification of nodule type (solid vs subsolid nodule), nodule size and volume measurements. Comparative evidence between different testing strategies (e.g. AI-assisted readers, stand-alone AI, unaided readers) are then presented in preference over 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/sub-solid nodule). Original data extracted from the studies were used to construct 2x2 tables. Pairs of sensitivities and specificities are also displayed in a paired forest plot to demonstrate scatter and uncertainty. Studies are grouped by software and its role in the workflow (e.g. stand-alone software, software-assisted reader).

Given the substantial heterogeneity in 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 being presented in structured tables and plotted in figures where feasible.

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

- Patients' ethnicity;
- Reason for CT scan (within the incidental population);
- CT scans with vs without contrast;
- CT scans using different radiation doses (e.g. ultra-low-dose, low-dose, standard dose);
- Solid nodules vs sub-solid nodules;
- General radiologist (or other healthcare professional) vs 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 technology being evaluated. If applicable, comparative evidence between different reading modes (e.g. AI-assisted readers versus unaided readers) are presented in preference over non-comparative evidence (e.g. individual testing strategy).

3 SYSTEMATIC REVIEW OF ASSESSING TEST ACCURACY, PRACTICAL IMPLICATIONS AND IMPACT ON PATIENT MANAGEMENT (KEY QUESTION 1) - RESULTS

Findings of systematic reviews and company submissions answering key question 1 are presented in the following sections.

3.1 Description of the evidence

3.1.1 Results of literature search

Electronic database searches yielded 6,330 results, of which 4,886 were published since 2012. Twenty-two records were judged to be relevant for key question 1 (**Figure 6**). An additional eight relevant records were identified through author contact of potentially relevant articles (n=1[25]), searching company websites (n=2[26, 27]), company submissions (n=3[28-30]), reviewers' Google search for published version of unpublished manuscript (n=1[31]) and tracking of registered clinical trials (n=1[32]), so 30 articles reporting 27 studies were included for key question 1.

The study by Murchison et al. is reported in two conference articles[26, 27] and a journal article.[31] As the two conference articles from 2019 only report minimal additional information, in-text citations from hereon only refer to Murchison et al. (2022).[31] The study by Hall et al. is reported in a conference abstract[33] and a full journal article.[25] As the conference abstract from 2019 only reports minimal additional information, in-text citations from hereon only refer to Hall et al. (2022).[25]

Eleven articles evaluated relevant technologies but were excluded because the population comprised more than 10% patients with extra-thoracic cancer or with previously diagnosed lung cancer.[34-44] These studies were not formally assessed, but the main study characteristics and outcome measures are summarised in **Appendix 2 Table 66.**

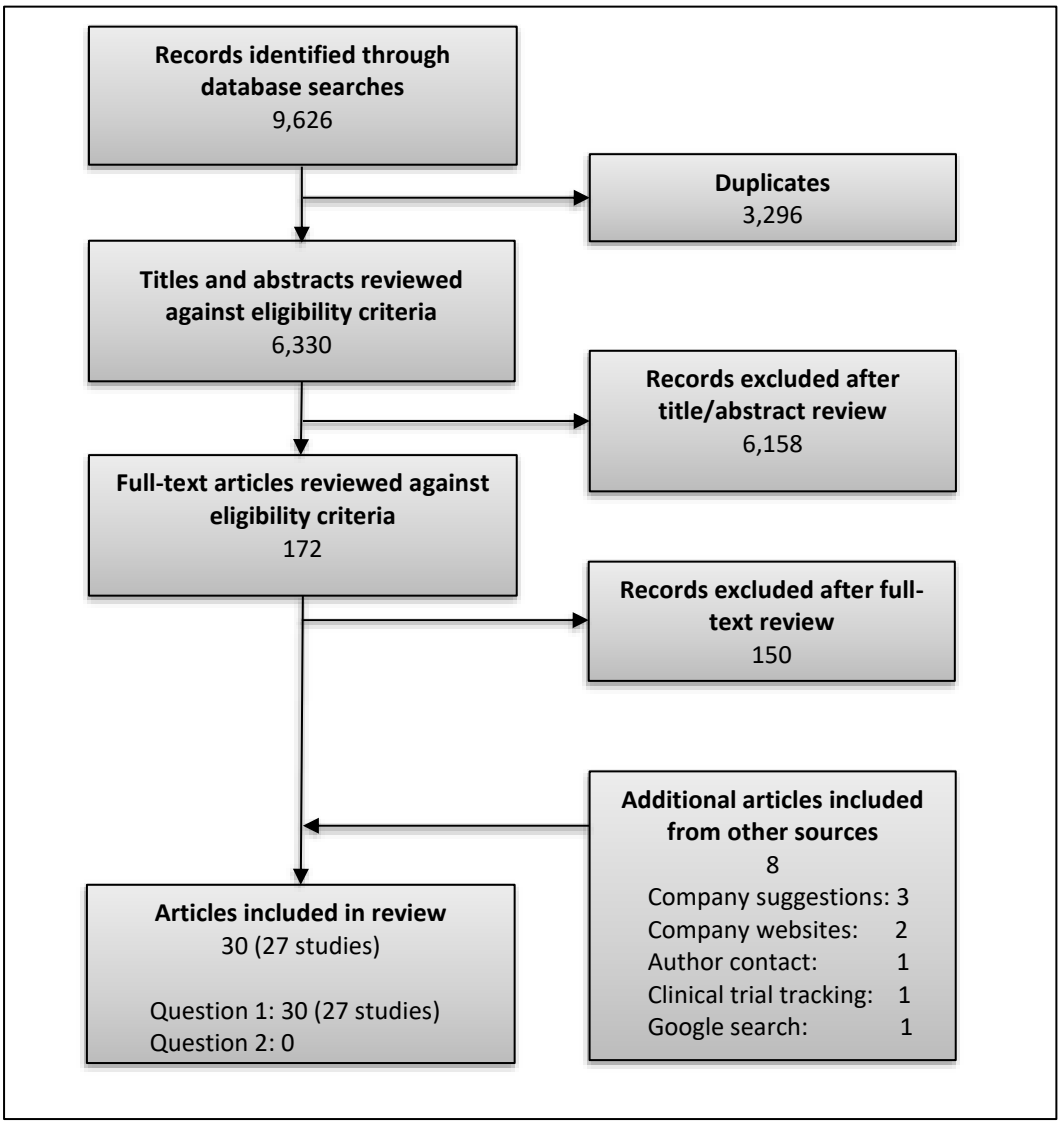


Figure 6. PRISMA diagram. Summary of publications included and excluded at each stage of the review

3.1.2 Characteristics of included studies

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

- AI-RAD Companion (Siemens Healthineers): 3 studies (USA 2, Germany 1);[45-47]
- AVIEW LCS+ (Coreline Soft): 4 studies (South Korea 3, Russia 1);[30, 48-50]
- ClearRead CT (Riverain Technologies): 6 studies (USA 2, Taiwan 2, Japan 1, Switzerland 1);[51-56]
- Contextflow SEARCH Lung CT (contextflow): 1 study (Austria 1);[29]
- InferRead CT Lung (Infervision): 3 studies (China 2, Japan 1);[57-59]
- Veolity (MeVis): 4 studies (**UK 1**, South Korea 2, USA [data]/Netherland/Denmark [readers] 1);[25, 60-62]
- Veye Lung Nodules (Aidence): 5 studies (**UK 1**, Netherlands 3, USA 1);[28, 31, 32, 63, 64]
- VUNO Med-LungCT AI (VUNO): 1 study (USA[data]/South Korea [readers] 1).[65]

Sixteen studies were multi-reader multi-case (MRMC) studies: Eight studies compared stand-alone AI software to human readers with and without concurrent AI software use under laboratory conditions.[30, 31, 51, 52, 54, 57, 58, 65] In "concurrent" AI software use, the software results are simultaneously displayed to readers during the reading. For brevity, we describe human reading with concurrent use of AI software as "concurrent AI" in this report. The study by Hsu et al. also assessed "assisted second-read" AI software use, where the human reader assessed the CT images without AI software first, then opened the software results, revised their assessment and made the final decision.[51] One MRMC study compared stand-alone software performance to unaided readers,[56] and six studies compared the performance of readers with and without concurrent software use, with both reading sessions performed under laboratory conditions.[29, 32, 53, 55, 61, 62] The remaining MRMC study compared software-assisted nodule measurement in CT images reconstructed with both filtered back projection (FBP) and model-based iterative reconstruction (MBIR) algorithms without comparison to unaided readers.[60]

Five studies were retrospective test accuracy studies evaluating the performance of stand-alone software only,[28, 46, 63, 64] or in comparison to original unaided reading (clinical practice).[47]

Three studies were classed as retrospective test accuracy studies as well as MRMC studies. One study performed a MRMC study comparing stand-alone AI and readers with concurrent AI to unaided reading, and additionally used the original radiologist reports as comparator.[45] The other two studies compared readers with concurrent AI use with reading performed under laboratory conditions to unaided radiologists in clinical practice.[25, 59]

Three studies reported prospective screening experiences: two studies only included software-assisted reading,[48, 50] whereas the remaining study was a before-after study that evaluated the performance of stand-alone software as well as that of the original readers before and after software implementation.[49]

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

- Symptomatic population (n=1):

  One study was performed in a randomly selected symptomatic population.[57]

- Incidental population (n=1):

  One study included a consecutive incidental population.[47]

- Screening population (n=11):

  Eleven studies included screening populations, of which six used consecutive or random sampling,[25, 46, 48-50, 59] and five were nodule-enriched (selection by nodule presence / absence, resulting in a higher nodule prevalence than expected for this population).[30, 52, 54, 62, 65]

- 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 sub-solid nodules who underwent preoperative CT scans and subsequent surgical resection.[60, 61]

- 'Mixed population' (n=11):

  In eleven studies, there were various indications for the chest CT scan: three studies included consecutive or random sampling[28, 51, 64] one study used convenience sampling,[58] five studies included enriched populations[29, 31, 32, 45, 56] and in the remaining two studies, the sampling method was unclear.[55, 63] The reasons for the CT scan are reported in **Appendix 3 Table 68**, 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, the indication for the chest CT scan was not reported.[53]

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) (3 studies)					
Abadia 2021,[45] 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 [C] Concurrent AI (MRMC study)	[D] Unaided reader (MRMC study) [E] Original radiologist reading (clinical practice)	Nodule detection accuracy Nodule size measurement Characteristics of nodules (FN, FP) Reading times Confidence in lung nodule detection
Chamberlin 2021,[46] USA	Retrospective test accuracy study	Screening (random)	[A] Stand-alone AI	None	Nodule detection accuracy Characteristics of detected nodules
Rueckel 2021,[47] Germany	Retrospective test accuracy study	Incidental (consecutive)	[A] Stand-alone AI	[E] Original radiologist reading (clinical practice)	Nodule detection accuracy Characteristics of detected nodules
AVIEW LCS+ (Coreline Soft) (4 studies)					
Hwang 2021a,[49] South Korea	Before-and-after study	Screening (consecutive)	[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	[E] Original radiologist reading (clinical practice)	Characteristics of detected nodules % detected nodules being malignant Nodule detection accuracy ([A]) Accuracy to detect lung cancer (whole read [C] with Lung-RADS) Number of people with positive screening result Technical failure rate
Hwang 2021b,[48] South Korea	Retrospective analysis of prospective cohort study	Screening (consecutive)	[B] 2nd-read AI for nodule detection [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) Characteristics of detected nodules % nodules being malignant Number of people with positive screening result Technical failure rate
Hwang 2021c,[50] South Korea	Prospective screening cohort	Screening (consecutive)	[B] Assisted 2nd-read AI for nodule detection	None	Characteristics of detected nodules Number of people having CT surveillance Number of people having excision/biopsy

Study, country	Study design	Target population	Index test	Comparator	Relevant outcomes reported
			[C] Concurrent AI for nodule measurement and whole read including Lung-RADS categorisation		Technical failure rate
Lancaster 2022,[30] Russia	MRMC study	Screening (nodule-enriched)	[A] Stand-alone AI [C] Concurrent AI	[D] Unaided reader	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
ClearRead CT (Riverain Technologies) (6 studies)					
Singh 2021,[54] USA	MRMC study	Screening (nodule-enriched)	[A] Stand-alone AI-AD (with vessel suppression and autodetection of pulmonary nodules) [C.1] Concurrent AI (with vessel suppression, without automatic nodule detection) [C.2] Concurrent AI (with vessel suppression and autodetection of pulmonary nodules)	[D] Unaided reader	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)
Lo 2018,[52] USA	MRMC study	Screening (nodule-enriched)	[A] Stand-alone AI [C] Concurrent AI	[D] Unaided reader	Nodule detection accuracy Radiologist reading time
Milanese 2018,[53] Switzerland	MRMC study	Unclear (consecutive)	[C] Concurrent AI (vessel-suppressed CT images) using semi-automatic segmentation software (MM Oncology, Siemens Healthcare)	[D] Unaided reader (standard CT images) using semi-automatic segmentation software (MM Oncology, Siemens Healthcare)	Measurement accuracy Inter-reader variability in nodule measurement Impact on clinical decision-making (categorisation according to Fleischner guidelines).[66]
Hsu 2021,[51] Taiwan	MRMC study	Mixed:	[A] Stand-alone AI [B] Assisted 2nd-read AI	[D] Unaided reader	Nodule detection accuracy Radiologist reading time

Study, country	Study design	Target population	Index test	Comparator	Relevant outcomes reported
		clinical routine; screening (consecutive)	[C] Concurrent AI		
Takaishi 2021,[55] Japan	MRMC study	Mixed (unclear selection)	[C] Concurrent AI	[D] Unaided reader	Nodule detection accuracy Reading time
Wan 2020,[56] Taiwan	MRMC study	Mixed (selected only patients with subsequent nodule excision)	[A] Stand-alone AI	[D] Consensus of 2 radiologists measuring diameter manually	Nodule detection accuracy Lung cancer detection accuracy Characteristics of missed nodules Measurement concordance between stand-alone AI and unaided reader consensus
Contextflow SEARCH Lung CT (contextflow) (1 study)					
Röhrich 2022,[29] Austria	MRMC study	Mixed (selected by presence/absence of diffuse parenchymal lung disease)	[C] Concurrent AI	[D] Unaided reader	Radiologist reading time Technical failure rate
InferRead CT Lung (Infervision) (3 studies)					
Kozuka 2020,[57] Japan	MRMC study	Symptomatic (random)	[A] Stand-alone AI [C] Concurrent AI	[D] Unaided reader	Nodule detection accuracy Reading time Characteristics of detected nodules
Liu 2019,[58] China	MRMC study	Mixed (convenience sample)	Evaluation 1: [A] Stand-alone AI Evaluation 4: [C] Concurrent AI	Evaluation 1 [D.1] Unaided reader Evaluation 4 [D.2] Unaided reader	Nodule detection accuracy Comparison of AI performance by radiation dose Radiologist reading time
Zhang 2021,[59] China	Retrospective test accuracy study and MRMC study	Screening (consecutive)	[C] Concurrent AI (MRMC study)	[E] Original radiologist reading (clinical practice)	Nodule detection accuracy 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.					

Study, country	Study design	Target population	Index test	Comparator	Relevant outcomes reported
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) (4 studies)					
Cohen 2017,[60] South Korea	MRMC study	Surveillance (preoperative CT scan for subsolid nodules and subsequent surgical resection) (consecutive)	[C] Concurrent AI (FBP versus MBIR reconstruction algorithms)	None	Diameter and volume measurement: Technical failure rate Inter-observer variability Repeatability / reproducibility Concordance between readers with software: FBP versus MBIR.
Kim 2018,[61] South Korea	MRMC study	Surveillance (preoperative CT scan for subsolid nodules and subsequent surgical resection) (consecutive)	[C] Concurrent AI	[D] Unaided reader	Diameter measurement: Concordance between readers with and without software Inter-observer variability Repeatability / reproducibility Technical failure rate Nodule classification by size of solid portion: Inter-observer variability Repeatability / reproducibility
Hall 2022,[25] UK	Retrospective test accuracy study and MRMC study	Screening (consecutive)	[C] Concurrent AI (MRMC study)	[E] Original radiologist reading (clinical practice)	Nodule detection accuracy Lung cancer detection accuracy Impact on decision making Radiologist reading time Software acceptability & experience Proportion of scans referred for CT surveillance Proportion of scans referred to MDT Characteristics of missed nodules % detected nodules being malignant
Jacobs 2021,[62] USA, Denmark, Netherlands	MRMC study	Screening (selected by Lung-RADS category)	[C] Concurrent AI	[D] Unaided reader	Lung-RADS categorisation: Inter-observer variability Repeatability / reproducibility

Study, country	Study design	Target population	Index test	Comparator	Relevant outcomes reported
					Radiologist reading time Technical failure rate Impact on decision-making
Veye Lung Nodules (Aidence) (5 studies)					
Blazis 2021,[63] Netherlands	Retrospective test accuracy study	Mixed (unclear selection)	[A] Stand-alone AI	None	Nodule detection accuracy
Hempel 2022,[32] Netherlands	MRMC study	Mixed (incidentally detected nodules or no nodules, with or without prior CT)	[C] Concurrent AI	[D] Unaided reader	Accuracy of BTS grade categorisation Characteristics of detected nodules Radiologist reading time Technical failure rate Inter-observer variability
Martins Jarnalo 2021,[64] Netherlands	Retrospective test accuracy study	Mixed (random)	[A] Stand-alone AI	None	Nodule detection accuracy Nodule type accuracy (solid, sub-solid) Size measurement accuracy Characteristics of detected (TP, FP) and missed (FN) nodules Technical failure rate Software acceptability and experience
Murchison 2022,[31] UK	MRMC study	Mixed - clinical routine mimicking a screening population in age and smoking history (selected)	[A] Stand-alone AI [C] Concurrent AI	[D] Unaided reader	Nodule detection accuracy Nodule type determination accuracy Measurement (volume, diameter): Inter-observer variability Concordance between stand-alone software and readers without software Technical failure rate Growth rate: Nodule registration accuracy Inter-observer variability Concordance between stand-alone software and readers without software
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Study, country	Study design	Target population	Index test	Comparator	Relevant outcomes reported
VUNO Med-LungCT AI (VUNO) (1 study)					
Park 2022,[65] USA, Korea	MRMC study	Screening (nodule-enriched)	[A] Stand-alone AI [C] Concurrent AI	[D] Unaided reader	Nodule detection and Lung-RADS categorisation: Lung cancer detection accuracy Concordance between stand-alone software and readers Inter-observer variability Impact on decision making

AI, artificial intelligence software; BTS, British Thoracic Society; CT, Computed tomography; FBP, Filtered back projection; FN, False negative; FP, False positive; Lung-RADS, Lung imaging reporting and data system; MDT, Multi-disciplinary team; MBIR, Model-based iterative reconstruction; MRMC, multi-case multi-reader study; TP, True positive.

To help navigating the results section, the tables below present for each pre-specified outcome the number of studies identified, study details and a link to the corresponding section in the report.

Table 3. Outcomes – Nodule detection and analysis: Accuracy, concordance and variability

Outcome	Section in report	Comparison	# studies	Target population, references
Use case 1: Nodule detection and analysis in people with no known lung nodules				
Nodule detection: Accuracy – Any nodule	3.3.1.1	[C] vs [D]	N=4	Screening[51, 59] Symptomatic[57] Mixed[51, 55]
		[B] vs [D]	N=1	Screening[51] Mixed[51]
	13.5.1	[A] vs [D]	N=4	Symptomatic[57] Incidental[47] Mixed[45, 58]
		None: [A]	N=6	Screening[49] Mixed[28, 45, 56, 63, 64]
Nodule detection: Accuracy – Actionable nodules	3.3.1.2	[C] vs [D]	N=5	Screening[25, 52, 54] Symptomatic[57] Mixed[31]
	13.5.2	[A] vs [D]	N=2	Symptomatic[57] Mixed[58]
		None: [A]	N=2	Screening[46] Mixed[28]
Nodule detection: Accuracy – Malignant nodules	3.3.1.3	[C] vs [D]	N=3	Screening[52, 65] Mixed[55]
	13.5.3	None	N=3	Screening[25, 49] Mixed[56]
Nodule detection: Effect modifiers	3.3.1.4 b)	Radiation dose	N=2	Mixed[51, 58]
	3.3.1.4 c)	Nodule type	N=7	Screening[49, 52, 54, 59] Symptomatic[57] Mixed[58, 64]
	3.3.1.4 e)	Radiologist experience	N=1	Screening[51] Mixed[51]
Nodule detection: Concordance	3.3.1.5	[A] [C] vs [D]	N=1	Mixed[45]
		Inter-observer	N=1	Screening[54]
Nodule type: Accuracy	3.3.2.1	None: [A]	N=2	Mixed[31, 64]
Nodule type: Concordance	3.3.2.3	Inter-observer	N=2	Screening[62, 65]
Diameter measurement: Accuracy	3.3.3.1	None: [C]	N=1	Unclear[53]
		None: [A]	N=2	Screening[54] Mixed[64]
Diameter measurement: Concordance	3.3.3.3	[A] [C] vs [D]	N=4	Surveillance with applicability concerns[61] Mixed[31, 45, 56]
		Inter-observer	N=5	Screening[62, 65] Surveillance with applicability concerns[60, 61] Mixed[31]

Outcome	Section in report	Comparison	# studies	Target population, references
		Intra-observer	N=2	Surveillance with applicability concerns[60, 61]
Volume measurement: Accuracy	3.3.4.1	None: [C]	N=1	Unclear[53]
Volume measurement: Concordance	3.3.4.3	[A] vs [D]	N=1	Mixed[31]
		Inter-observer	N=3	Surveillance with applicability concern[60] Mixed[31] Unclear[53]
		Intra-observer	N=1	Surveillance with applicability concerns[60]
Risk categorisation: Accuracy	3.3.5.1	[A] [C] vs [D]	N=3	Screening[30] Mixed[32] Unclear[53]
Risk categorisation: Concordance	3.3.5.4	[A] [C] vs [D]	N=2	Screening[62, 65]
		Inter-observer	N=5	Screening[62, 65] Surveillance with applicability concerns[60, 61] Mixed[32]
		Intra-observer	N=2	Surveillance with applicability concerns[60, 61]
Whole read: Accuracy for lung cancer	3.3.6.1	[C] vs [D]	N=1	Screening[49]
		None [C]	N=1	Screening[48]
Use case 2: Nodule growth monitoring in people with previously identified lung nodules				
Nodule registration: Accuracy	3.4.2.1 13.5.6.1	None [A]	N=1	Mixed[31]
Nodule growth rate: Concordance	3.4.2.3 13.5.6.2	[A] vs [D]	N=1	Mixed[31]
		Inter-observer	N=1	Mixed[31]

[A] Stand-alone AI; [B] Assisted 2nd-read AI; [C] Concurrent AI; [D] Unaided reading.

Table 4. Outcomes – Practical implications

Outcome	Section in report	# studies	Target population, references
Technical failure rate	3.5.1	N=12	Screening[25, 48-50, 54, 62]
			Surveillance with applicability concerns[60, 61]
			Mixed[29, 31, 32, 64]
Radiologist reading time	3.5.2	N=10	Screening[25, 52, 62]
			Symptomatic[57]
			Mixed[29, 32, 45, 51, 55, 58]
Acceptability and experience of using the software	3.5.4	N=3	Screening[25]
			Mixed[45, 64]

[A] Stand-alone AI; [B] Assisted 2nd-read AI; [C] Concurrent AI; [D] Unaided reading; NA, not applicable.

Table 5. Outcomes - Impact on patient management

Outcome	Section in report	Comparison	# studies	Target population, references
Characteristics of detected nodules				
All detected nodules (TP and FP)	3.6.1.1	[C] vs [D]	N=2	Screening[49] Mixed[32]
		[A] vs [D]	N=1	Mixed[45]
	13.5.8.1	None	N=3	Screening[48, 50] Mixed[64]
TP nodules	3.6.1.2	[C] vs [D]	N=2	Screening[59] Symptomatic[57]
		[A] vs [D]	N=1	Mixed[58]
	13.5.8.2	None	N=4	Screening[30, 49, 54] Mixed[64]
Additional TP nodules detected by software	3.6.1.3	[A] vs [D]	N=1	Incidental[47]
FP nodules	3.6.1.4	None ([A] only)	N=4	Screening[46] Incidental[47] Mixed[45, 64]
FN nodules	3.6.1.5	[C] vs [D]	N=2	Screening[59] Symptomatic[57]
	13.5.8.3	None	N=5	Screening[25, 49, 54] Mixed[56, 64]
Proportion of detected nodules that are malignant	3.6.2	[C] vs [D]	N=2	Screening[25, 49]
		None	N=1	Screening[48]
Impact of test result on clinical decision-making	3.6.3	[C] vs [D]	N=6	Screening[25, 54, 62, 65] Surveillance with applicability concerns[61] Unclear[53]
# of people having CT surveillance	3.6.4	[C] vs [D]	N=2	Screening[49, 62]
	13.5.8.4	None	N=3	Screening[25, 50] Symptomatic[57]
# of people having biopsy or excision	3.6.6	C] vs [D]	N=2	Screening[49, 62]
	13.5.8.5	None	N=3	Screening[25, 50] Symptomatic[57]
Time to diagnosis	3.6.8	[C] vs [D]	N=1	Screening[62]

[A] Stand-alone AI; [B] Assisted 2nd-read AI; [C] Concurrent AI; [D] Unaided reading.

3.2 Methodological quality of the evidence

The methodological quality of 22 studies[25, 28-32, 45-49, 51-59, 63, 64] that reported test accuracy outcomes was assessed using QUADAS-2[22] and, if applicable, QUADAS-C.[23]

Four studies[60-62, 65] reported concordance or agreement outcomes, and their quality was assessed using the COSMIN Risk of Bias tool (see Section **2.2.2**).[24] For the remaining study,[50] no quality appraisal was performed as the relevant outcomes for the DAR were neither related to accuracy nor reliability/measurement error.

3.2.1 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 **Table 6**, **Figure 7** and **Figure 8.**

Table 6. Quality assessment results based on QUADAS-2 and QUADAS-C tools (22 studies)

	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 2021a	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									

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 & timing; I, Index test; INCID, Incidental population; P, Population; R, Reference standard; SCREEN, Screening population; SURV, Surveillance population; SYMP, Symptomatic population.

3.2.1.1 Risk of bias

Sixteen of the 22 studies were comparative test accuracy studies. The risk of bias according to QUADAS-C was considered 'high' in three or more domains in 12 (75%) studies. In the remaining six non-comparative test accuracy studies, risk of bias (QUADAS-2) was considered high in three or more domains in one (17%) study. No comparative or non-comparative test accuracy study was rated as 'low' or 'unclear' risk of bias in all four domains. The number and proportion of studies with 'low', 'high' and 'unclear', respectively, risk of bias are presented in **Figure 7** for all 22 studies as well as separately for the 16 comparative studies (QUADAS-C) and the six non-comparative studies (QUADAS-2).The risk of bias in the four QUADAS-2 domains is discussed in more detail below.

<u>Patient selection domain</u>

The risk of bias was classed as 'high' in the patient selection domain in 15 (68%) out of 22 studies. The main reasons have been listed below:

- No consecutive or random sample: 8 studies;[29-32, 45, 52, 54, 58]
- Case-control design not avoided: 8 studies;[29-32, 45, 52, 54, 56]
- Systematic exclusion of 'easy to read' CT images (e.g. exclusion of patients without other, non-nodule related lung conditions): 2 studies;[29, 45]
- Exclusions by nodule number per image or unjustified (not based on management guidelines or minimal software manufacturer threshold, exclusion of certain nodule sizes): 6 studies;[28, 30, 32, 51, 56, 64]
- 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): 5 studies;[31, 32, 51, 55, 57]
- No fully paired or randomized design was used: 1 study.[49]

Four studies (18%) were classified as 'unclear' risk of bias,[25, 48, 53, 63] and the remaining three studies (14%) were classed as 'low' risk of bias in the patient selection domain.[46, 47, 59]

Index test domain

For the index test domain, three studies (14%) were classed as low risk of bias.[46, 48, 49] In 16 studies (73%), the risk of bias was considered as 'high' for the following reasons:

- Readers assessed the chest CT images outside clinical practice (MRMC studies): 14 studies;[25, 29-32, 45, 51-55, 57-59]
- AI software threshold clearly not pre-set by company or not pre-specified in methods: 4 studies.[28, 31, 58, 63]

For 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: 1 study;[64]
- Unclear if the threshold was pre-specified: 2 studies.[47, 56]

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 21 studies (29%) were classified as low risk of bias.[28, 52, 54, 56-58] 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;[25, 29, 45-47, 49, 51, 53, 55, 59, 64]
- Reference standard reader(s) were part of the index test(s) or not blinded to index test markings / decisions: 13 studies.[25, 30, 31, 45-47, 49, 51, 53, 55, 59, 63, 64]

For lung cancer detection, two out of six studies were rated as low risk of bias.[52, 56] Two studies were classed as high risk of bias as medical records were used as reference standard,[48, 49] and the clinicians undertaking the diagnostic follow up tests were not blinded to the results of the index test.[48, 49] In the remaining two studies, the risk of bias was rated as unclear as it was not stated how benign nodules were followed up[55] and no details about the reference standard were reported,[25] respectively.

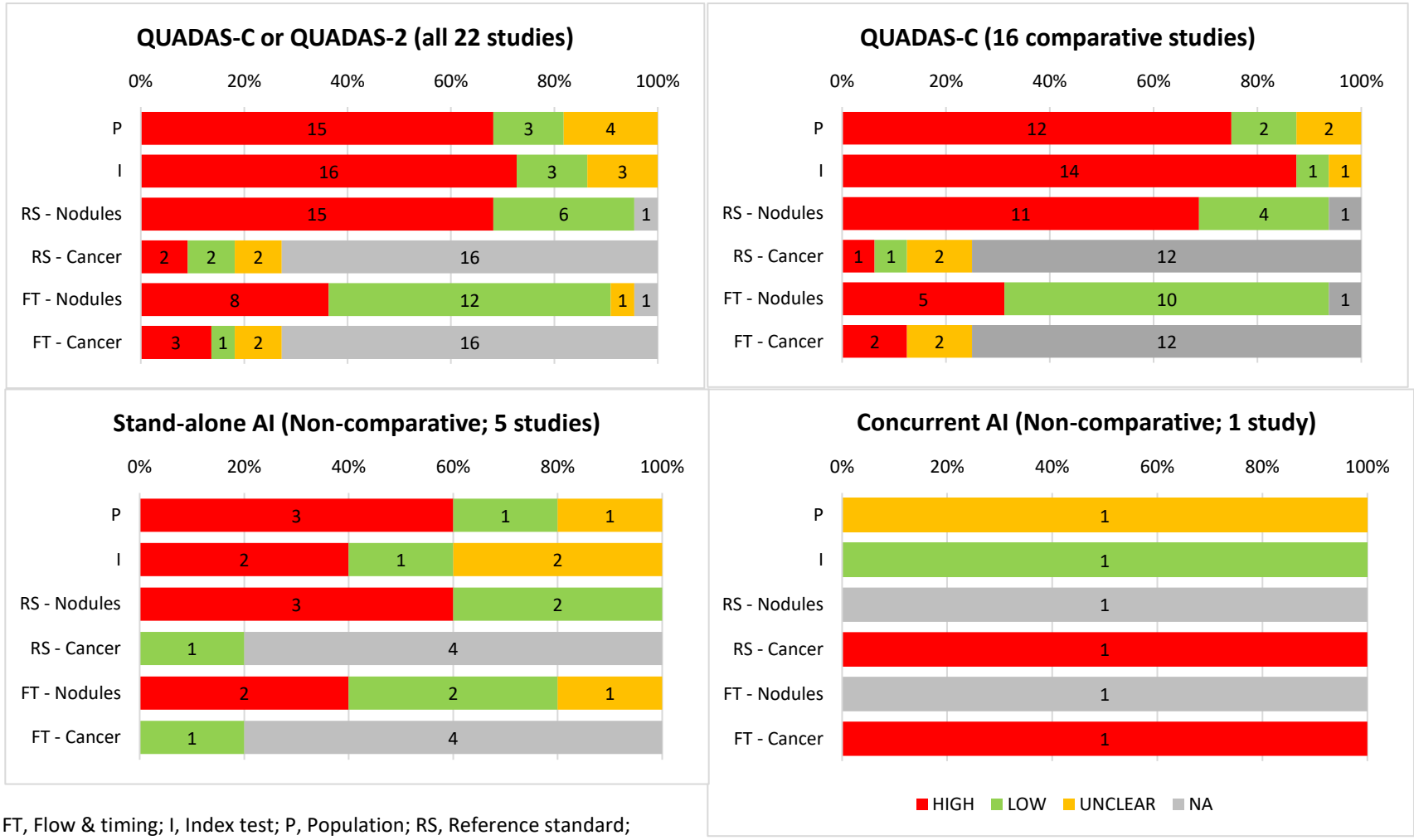
Flow and timing domain

In the 21 studies evaluating lung nodule detection accuracy, the risk of bias was rated as 'low' in 12 studies[29, 30, 32, 45, 47, 51, 52, 55-57, 59, 64] (57%) and as 'unclear' in one study[28] (5%). A high risk of bias was present in the remaining eight studies (38%) for the following reasons:

- There were significant exclusions (>10%; cut-off determined pragmatically) after the point of selecting the cohort: 6 studies;[25, 31, 53, 54, 58, 63]
- The number of CT images excluded due to software processing failures (e.g. segmentation failures) has not been reported: 3 studies;[31, 46, 49]

In the six studies reporting on lung cancer detection accuracy, the risk of bias was rated as 'low' in one study,[56] as 'unclear' in two studies,[25, 49] and as 'high' in three studies[48, 52, 55] for the following reasons:

- Not all patients received a reference standard: 1 study;[48]
- Not all patients received the same reference standard: 2 studies.[52, 55]



FT, Flow & timing; I, Index test; P, Population; RS, Reference standard; NA, Not applicable.

Figure 7. Findings of risk of bias assessment for all 22 studies as well as separately for comparative (QUADAS-C) and non-comparative (QUADAS-2) studies

3.2.1.2 Applicability concerns

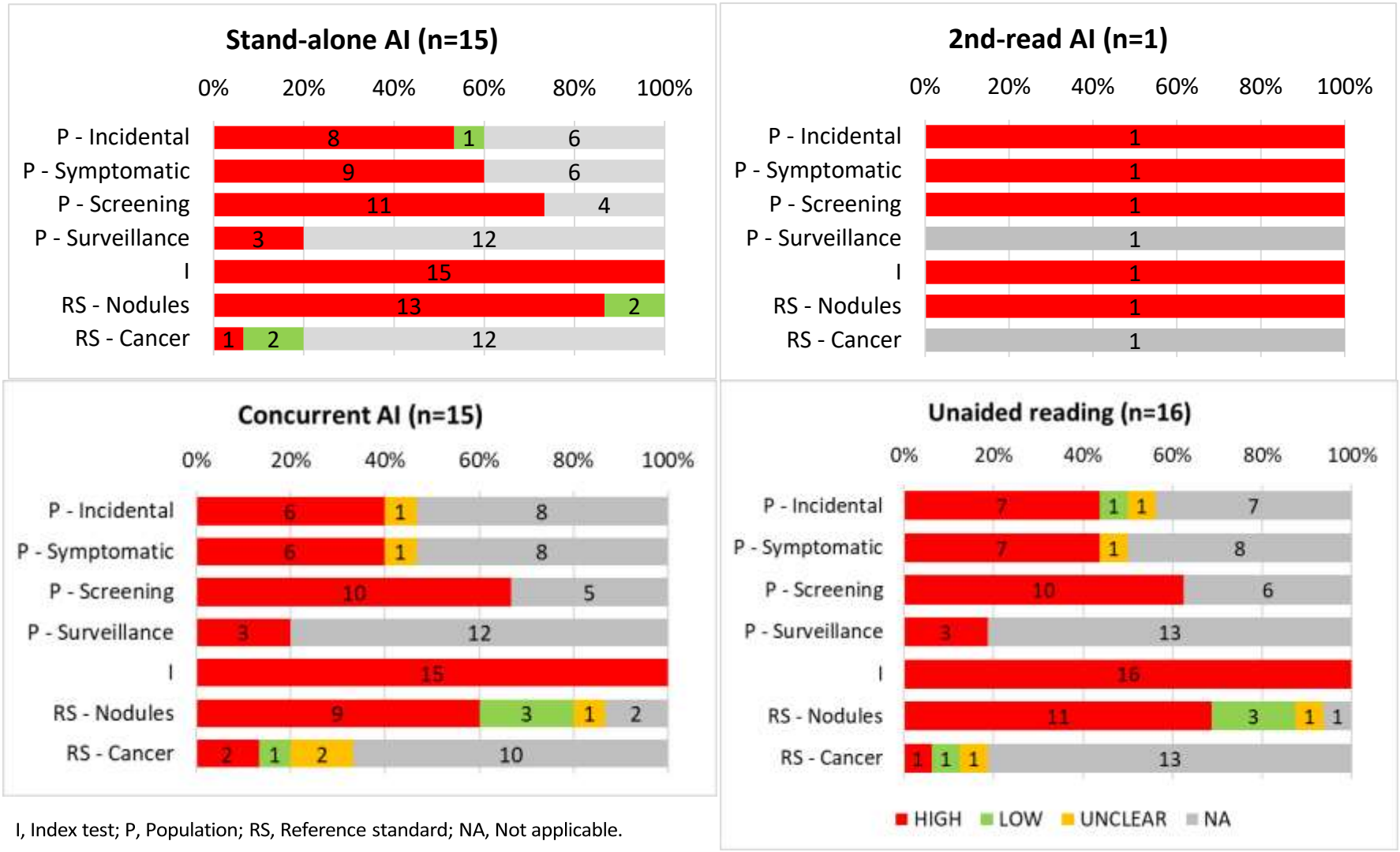
Overall, all 22 studies had 'high' applicability concerns in at least two of the three domains (i.e. population, index test, reference standard). The number and proportion of studies with 'low', 'high' and 'unclear', respectively, applicability concerns are presented in **Figure 8**, separately for each 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 listed below:

- Not a consecutive or random sample of patients / CT images: 9 studies;[29-31, 45, 52, 55, 58, 63, 67]
- Enriched sample (e.g. in-/exclusion by nodule number, nodule type and nodule size, respectively): 8 studies;[29, 30, 45, 51, 52, 56, 64, 67]
- Inclusion/Exclusion by age: 1 study;[31]
- Study not performed in the UK or another North-Western European country: 14 studies;[28, 30, 45, 46, 48, 49, 51, 52, 55-59, 67]
- >10% of included people have a different indication for the CT scan than the target population: 11 studies;[28, 29, 31, 45, 51, 53, 55, 56, 58, 63, 64]
- CT image acquisition details (dose, contrast use, slice thickness) different to UK practice for target population: 8 studies;[28, 30, 31, 51, 55, 56, 63, 64]
- Age not between 55-75 years in screening populations: 6 studies;[25, 30, 46, 53, 56, 59]
- Nodule size <5mm or >30mm maximal diameter; <80mm^3 in a surveillance population: 1 study.[53]

Only one study was classified as having 'low' applicability concerns for the 'incidental' population.[47] In another study,[53] the applicability to the 'Incidental' and 'Symptomatic' populations was 'unclear' as it was not reported if more than 10% of included people had a different indication for the CT scan than the target population.



I, Index test; P, Population; RS, Reference standard; NA, Not applicable.

Figure 8. Findings of applicability concern assessment (QUADAS-2) by index test

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): 2 studies;[45, 47]
- Integration of software into pathway not applicable to UK (e.g. stand-alone AI performance instead of concurrent or second-read software use: 12 studies;[28-31, 45, 49, 56-58, 63, 64, 67]
- Reader had no access to maximum intensity projections (MIP) and/or multiplanar reformations (MPR): 6 studies;[29, 31, 52, 55, 58, 67]
- Study did not use a pre-specified nodule size threshold similar to the UK 2015 BTS guidelines (i.e. ≥5mm maximum axial diameter or ≥80mm^3)[11]: 14 studies;[29, 30, 46, 48, 49, 51, 52, 55, 56, 58, 59, 63, 64, 67]
- Other nodule types used than in the 2015 BTS guidelines (nodule type should be classified as solid, part-solid or pure ground glass nodules)[11]: 1 study;[64]
- For stand-alone AI - false positive rate set to more than 2 FP per image: 3 studies;[28, 31, 63]
- For concurrent and assisted 2nd-read software use - more than 1 human reader involved per read: 1 study;[59]
- For the unaided reader (comparator) - human double reading instead of single human reader: 2 studies;[25, 59]
- Human reader's experience and/or specialty not representative of UK clinical practice (5 years training for radiologists. After that time, they are considered "fully trained".) for target population: 8 studies;[25, 29, 51-53, 55, 57, 59]
- Software only had vessel suppression function, not nodule detection and measurement functions: 1 study.[53]

Reference standard domain

Applicability concerns regarding the reference standard *for lung nodules* (21 studies) were rated as 'low' in three studies[25, 31, 53] and as 'unclear' in one study.[29] The remaining 17 studies (81%) were rated as having high applicability concerns for the following reasons:

- For "Actionable" nodule present/absent: different nodule size to BTS 2015 guideline definition ("actionable nodule" is ≥5 mm maximum axial diameter or ≥80 mm^3)[11]: 17 studies;[28-30, 45-47, 49, 51, 52, 55-59, 63, 64, 67]
- Other types used than in the BTS 2015 guidelines (nodule type should be classified as solid, part-solid or pure ground glass nodules)[11]: 1 study;[64]
- For nodule size measurement (volume/diameter) - nodule size not measured as volume or, if volumetry segmentation is not possible, as maximum axial diameter: 2 studies.[53, 56]

Applicability concerns regarding the reference standard for *lung cancer* (6 studies) were rated as 'low' in two studies.[52, 56] Two studies[25, 55] were rated as having 'unclear' applicability concerns as no details on the reference standard were reported in one study,[25] whereas in the other study,[55] it was unclear, if benign nodules were followed up for at least two years without lung cancer diagnosis. The remaining two studies[48, 49] 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).

3.2.2 Risk of bias for reliability and measurement error (COSMIN tool)

The methodological quality of four studies[60-62, 65] was assessed using the COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument.[24] Assessment results are summarised in **Table 7.**

Table 7. Quality of studies assessed by COSMIN Risk of Bias tool[24] (4 studies)

COSMIN	Cohen 2017[60]	Jacobs 2021[62]	Kim 2018[61]	Park 2022[65]
Final risk of bias rating – Reliability studies	Doubtful	Doubtful	Doubtful	Doubtful
Final risk of bias rating – Studies on measurement error	Doubtful	Doubtful	Doubtful	Doubtful

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? 1 study;[60]
- Were there any other important flaws in the design or statistical methods of the study? 4 studies;[60-62, 65]

- For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated? 3 studies.[61, 62, 65]

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

3.3.1 Nodule detection

In this section we summarise findings related to accuracy for nodule detection. Three main outcomes (targets of detection) are presented in each of the sub-sections: detection of any nodules; detection of actionable nodules; and detection of malignant nodules (see **Figure 9**). In each sub-section, we focus on comparative evidence between AI-assisted detection and unaided detection by human readers (the main comparison of interest for this DAR). Detailed description of comparisons between stand-alone AI and unaided readers, and additional non-comparative evidence, such as the accuracy of AI-assisted detection or detection by stand-alone AI compared with a reference standard, are presented in **Appendix 5**.

Key characteristics, reported outcome measures and quality ratings for studies reporting comparative and non-comparative results are shown in **Table 8** and **Table 9,** respectively**.**

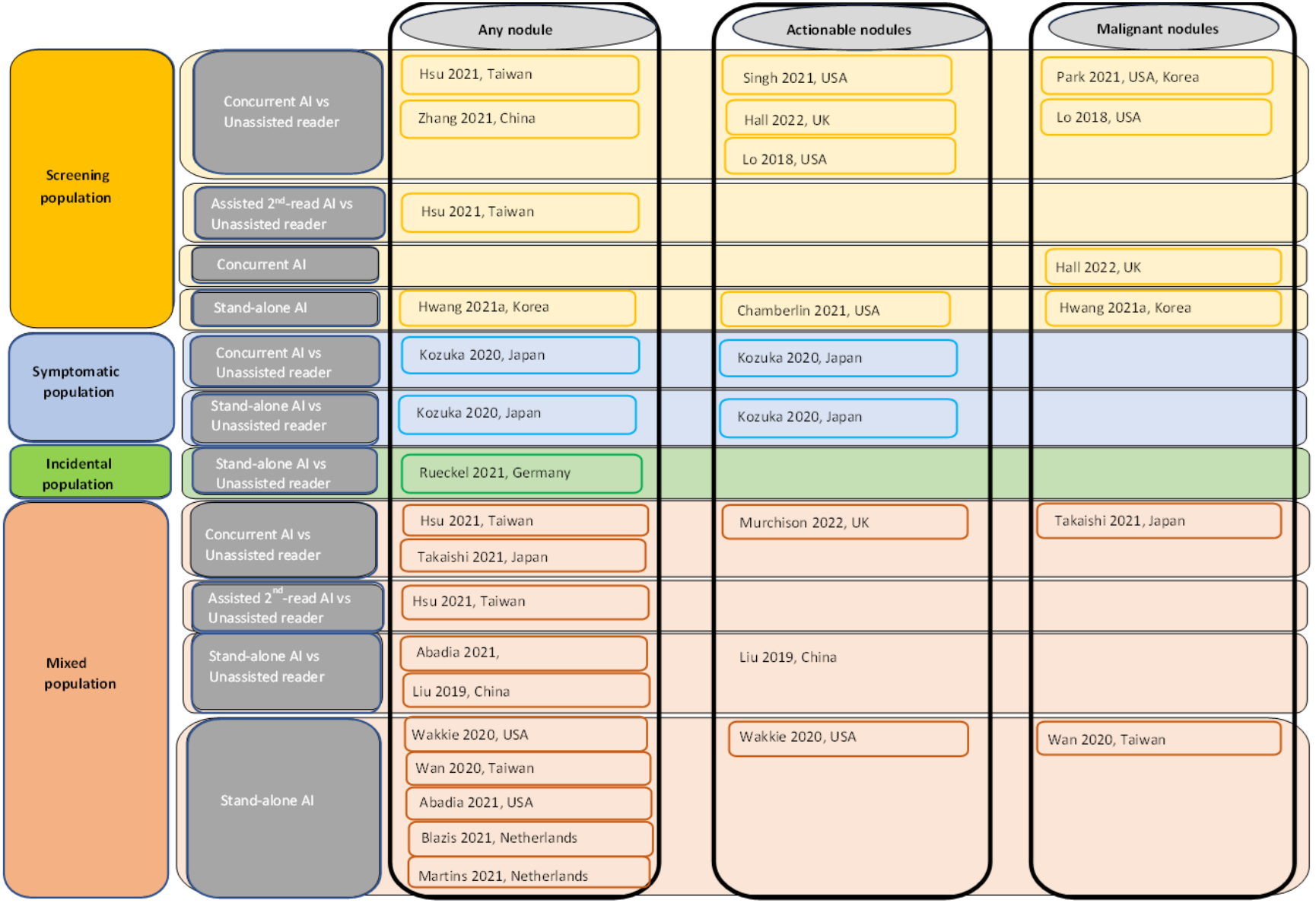


Figure 9. Visual map of included studies for detection accuracy based on population, comparison and reported outcomes (targets of detection)

Table 8. Characteristics of included studies with comparative results for nodule detection accuracy, and their quality ratings (12 studies)

Comparative results											
Study, country	**Population**	**Reading mode**	**Study design**	**Reader details**	**Nodule type**	**Nodule size**	**Sensitivity / Specificity / FP per scan**			**Quality of study**	
							Any nodule	**Actionable nodules**	**Malignant nodules**	**Risk of bias (QUADAS-C)**	**Applicability concerns**
Zhang 2021, China[59] **InferRead CT Lung (Infervision)**	Screening population	Concurrent AI vs unassisted reader	Retrospective test accuracy study and MRMC study	1 radiology resident with supervision of 1 experienced radiologist**	Solid, Part-solid, GGN	Solid: ≤5 mm, 6-7 mm, 8-14 mm, ≥15 mm; GGN & part-solid: all sizes	Sensitivity (per patient) / Specificity (per patient)			P: Low I: High RS (N): High F&T (N): Low	P: High I: High RS (N): High
Kozuka 2020,[57] Japan **InferRead CT Lung (Infervision)**	Symptomatic population	Concurrent AI vs unassisted reader; Stand-alone AI vs unassisted reader	MRMC study	2 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 I: High RS (N): Low F&T (N): Low	P: High I: High RS (N): High
Takaishi 2021,[55] Japan **ClearRead CT (Riverain Technologies)**	Mixed population	Concurrent AI vs unassisted reader	MRMC study	3 radiologists with < 10 years of experience	Any	≥4 mm	Sensitivity (per nodule) / FP per scan		Sensitivity (per nodule) / FP per scan	P: High I: High RS (N): High RS (C): Unclear F&T (N): Low F&T (C): High	P: High I: High RS (N): High RS (C): Unclear
Liu 2019,[58] Evaluation 4, China **InferRead CT Lung (Infervision)**	Mixed population	Concurrent AI vs unassisted reader	MRMC study	2 radiologists with approximately 10 years of experience	Any	NR	AUC			P: High I: High RS (N): Low F&T (N): High	P: High I: High RS (N): High
Liu 2019,[58] Evaluations 1-3, China **InferRead CT Lung (Infervision)**	Mixed population	Stand-alone AI vs unassisted reader	MRMC study	2 radiologists with 5 and 10 years of experience, respectively	Any, Solid, Sub-solid	Solid: ≤6 mm, >6 mm; Sub-solid: ≤5 mm, >5 mm	Sensitivity (per nodule) / FP/scan	Sensitivity (per nodule)		P: High I: High RS (N): Low F&T (N): High	P: High I: High RS (N): High

Comparative results											
Study, country	**Population**	**Reading mode**	**Study design**	**Reader details**	**Nodule type**	**Nodule size**	**Sensitivity / Specificity / FP per scan**			**Quality of study**	
							Any nodule	**Actionable nodules**	**Malignant nodules**	**Risk of bias (QUADAS-C)**	**Applicability concerns**
Hsu 2021,[51] Taiwan **InferRead CT Lung (Infervision)**	Mixed population (Screening population reported separately)	Concurrent AI vs unassisted reader; Assisted 2nd-read AI vs unassisted reader	MRMC study	6 readers: Junior group: 3 residents in radiology (1-2 years of CT experience and ≥6 months of chest CT experience); Senior group: 3 experienced chest radiologists (5, 10 and 25 years of experience, respectively).	Any	3-10mm	Sensitivity (per nodule) / Specificity (per patient)			P: High I: High RS (N): High F&T (N): Low	P: High I: High RS (N): High
Abadia 2021,[45] USA **AI-Rad Companion (Siemens Healthineers)**	Mixed population	Stand-alone AI vs unassisted reader	Retrospective test accuracy and MRMC study	Clinical practice: 1 of 5 single expert chest radiologist MRMC study: 1 expert chest radiologist (15 years of experience)	Any	≥4 mm	Sensitivity (per nodule) / FP per scan (for stand-alone AI only); Sensitivity (up to 3 largest nodules) / PPV			P: High I: High RS (N): High F&T (N): Low	P: High I: High RS (N): High
Rueckel 2021,[47] 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 I: Unclear RS (N): High F&T (N): Low	P: Low I: High RS (N): High
Singh 2021,[54] USA **ClearRead CT (Riverain Technologies)**	Screening population	Concurrent AI vs unassisted reader	MRMC study	2 radiologists (5 years and 10 years of thoracic CT experience)	GGN, Part-solid, Sub-solid	≥6 mm		Sensitivity (per nodule) / Specificity (per patient)		P: High I: High RS (N): Low F&T (N): High	P: High I: High RS (N): High
Lo 2018,[52] USA	Screening population	Concurrent AI vs unassisted reader	MRMC study	12 general radiologists certified by the American Board of	Any	5-44 mm		Sensitivity (per patient) / Specificity (per patient)	Sensitivity (per patient) / Specificity (per patient)	P: High I: High RS (N): Low RS (C): Low	P: High I: High RS (N): High RS (C): Low

Comparative results											
Study, country	**Population**	**Reading mode**	**Study design**	**Reader details**	**Nodule type**	**Nodule size**	**Sensitivity / Specificity / FP per scan**			**Quality of study**	
							Any nodule	**Actionable nodules**	**Malignant nodules**	**Risk of bias (QUADAS-C)**	**Applicability concerns**
ClearRead CT (Riverain Technologies)				Radiology (6–26 years of experience)						F&T (N): Low F&T (C): High	
Hall 2022,[25] UK **Veolity (MeVis)**	Screening population	Concurrent AI vs unassisted reader	Retrospective test accuracy study and MRMC study	[C] 2 radiographers without prior experience in chest CT reporting (MRMC study); [E] 5 radiologists (5-28 years of experience; 5% double reading) (clinical practice)	Any	≥5mm, ≥6mm		Sensitivity (per patient) / Specificity (per patient)		P: Unclear I: High RS (N): High F&T (N): High	P: High I: High RS (N): Low
Murchison 2022,[31] UK **Veye Lung Nodules (Aidence)**	Mixed population	Concurrent AI vs unassisted reader	MRMC study	2 thoracic radiologists (≥ 9 years' experience)	Any	5-30 mm		Sensitivity (per nodule) / FP per scan		P: High I: High RS (N): High F&T (N): High	P: High I: High RS (N): Low
Park 2022,[65] USA, Korea **VUNO Med-LungCT AI (VUNO)**	Screening population	Concurrent AI vs unassisted reader	MRMC study	5 readers: one 4-th year resident and 4 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

P, Patient selection domain; **I,** Index test domain; **RF (C),** Reference standard domain (lung cancer detection); **RF (N),** Reference standard domain (lung nodule detection); **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; NR, Not reported.

**[C] MRMC study: 1 radiology resident (5 years of experience) and 1 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 9. Characteristics of included studies with non-comparative results for nodule detection accuracy and quality ratings (8 studies)

Non-comparative results										
Study, country	**Population**	**Reading mode**	**Study design**	**Nodule type**	**Nodule size**	**Sensitivity / Specificity / FP rate**			**Quality of study**	
						Any nodule	**Actionable nodules**	**Malignant nodules**	**Risk of bias (QUADAS-2)**	**Applicability concerns**
Abadia 2021,[45] 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 I: Low RS (N): High F&T (N): Low	P: High I: High RS (N): High
Chamberlin 2021,[46] 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 I: Low RS (N): High F&T (N): High	P: High I: High RS (N): High
Hwang 2021a,[49] South 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 I: Low RS (N): High RS (C): High F&T (N): High F&T (C): Unclear	P: High I: High RS (N): High RS (C): High
Wan 2020,[56] Taiwan **ClearRead CT (Riverain Technologies)**	Mixed population	Stand-alone AI	MRMC study	Any	≤2 cm	Sensitivity		Sensitivity / Specificity	P: High I: Unclear RS (N): Low RS (C): Low F&T (N): Low F&T (C): Low	P: High I: High RS (N): High RS (C): Low
Blazis 2021,[63] 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 I: High RS (N): High F&T (N): High	P: High I: High RS (N): High

Non-comparative results										
Study, country	**Population**	**Reading mode**	**Study design**	**Nodule type**	**Nodule size**	**Sensitivity / Specificity / FP rate**			**Quality of study**	
						Any nodule	**Actionable nodules**	**Malignant nodules**	**Risk of bias (QUADAS-2)**	**Applicability concerns**
Martins Jarnalo 2021,[64] Netherlands **Veye Lung Nodules (Aidence)**	Mixed population	Stand-alone AI	Retrospective test accuracy study	Any, Solid, Sub-solid	4-30 mm	Sensitivity (per nodule) / FP per scan			P: High I: Unclear RS (N): High F&T (N): Low	P: High I: High RS (N): High
[redacted] **Veye Lung Nodules (Aidence)**	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]		P: High I: High RS (N): Low F&T (N): Unclear	P: High I: High RS (N): High
Hall 2022,[25] UK **Veolity (MeVis)**	Screening population	Concurrent AI	MRMC study: 2 radiographers without prior experience in chest CT reporting	Any	≥5 mm			Sensitivity	P: Unclear I: High RS (C): Unclear F&T (C): High	P: High I: High RS (C): Unclear

P, Patient selection domain; **I,** Index test domain; **RF (N),** Reference standard domain (lung nodule detection); **F&T (N),** Flow and timing domain (lung nodule detection); FP, False positive; GGN, Ground glass nodules; NR, Not reported.

3.3.1.1 Accuracy for identifying any nodules

a) Comparative results (7 studies)

Seven comparative studies[45, 47, 51, 55, 57-59] evaluated the **accuracy for detecting any nodules**. Of these, one included a screening population,[59] one included a symptomatic population,[57] one included an incidental population,[47] and four included mixed populations.[45, 51, 55, 58] The study by Hsu et al.[51] also reported accuracy data separately for the screening population subset.

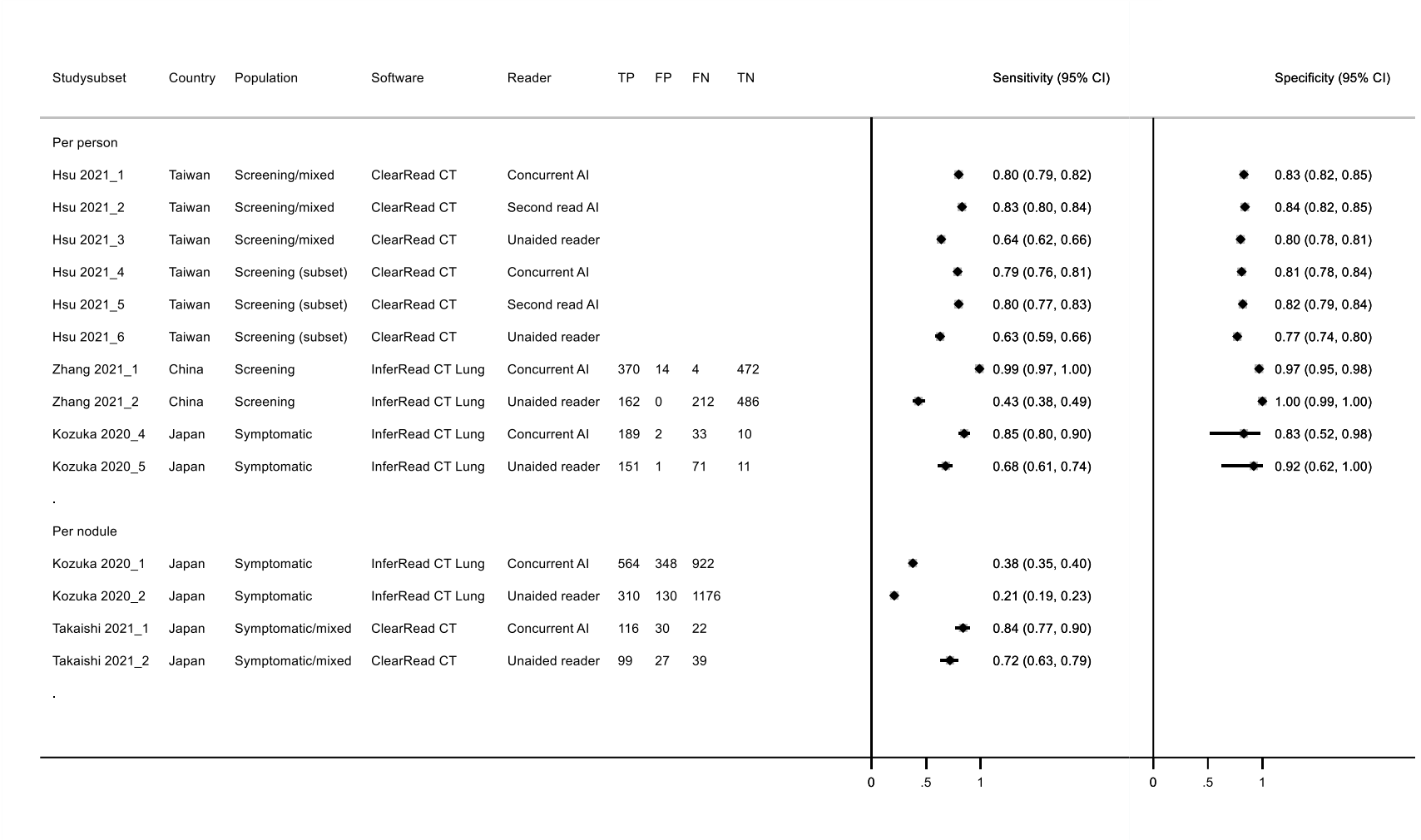
Four of the comparative studies provided evidence on the comparison between AI-assisted reading and unaided reading, and the findings are presented in **Figure 10. Evidence on AI-assisted reading compared with unaided reading for accuracy of detecting any nodules (7 studies)** Reported sensitivity varies widely among AI-assisted reading (range 0.38 to 0.99) and unaided reading (range 0.21 to 0.72) 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[57, 59] and slightly improved in one study[51] compared with unaided readers. Findings from Kozuka et al.[57] show that the per-person sensitivity tends to be higher than the per-nodule sensitivity, but the differences between reading with and without AI support remain similar (**Figure 10**). Further details from individual studies are provided below.

<u>Concurrent AI vs unassisted reader (4 studies)</u>

Screening population (2 studies)

Hsu 2021,[59] Taiwan - ClearRead CT (Riverain Technologies)

Hsu's study[59] included 150 consecutive cases with lung nodules ≤1 cm or no nodules (93 standard dose CT images from clinical routine and 57 low dose CT images [LDCT] 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 obtained for screening purposes. The mean per-nodule sensitivity of all six readers increased significantly from 63% (95% CI 59-66%) without software use to 79% (95% CI 76-81%) with software use ($p < 0.001$). The mean per-person specificity did not change significantly: 81% (95% CI 78-84%) with software use and 77% (95% CI 74-80%) for unaided readers ($p = 0.449$).



TP: true positive; FP: false positive; FN: false negative; TN: true negative. 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 excluding non-screening mixed populations

Figure 10. Evidence on AI-assisted reading compared with unaided reading for accuracy of detecting any nodules (7 studies)

Zhang 2021,[59] China - InferRead CT Lung (Infervision)

Zhang et al.[59] 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 (GGNs), 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 (1 study)**

Kozuka 2020,[57] Japan - InferRead CT Lung (Infervision)

One study[57] 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 lung cancer suspicion. Two less experienced radiologists (one and five years of diagnostic experience) assessed the CT images with and without software use. The per-nodule sensitivities for the pooled readers increased significantly from 20.9% (95% 18.8–23.0%) for the unaided reader to 38.0% (95% CI 35.5–40.5%) with concurrent AI ($p < 0.01$). The pooled PPV was 61.8% (95% CI 58.6-65.0%) with and 70.5% (95% CI 66.0-74.7%) without software. The pooled per-patient sensitivity increased significantly with software use from 68.0% (95% CI 61.4-74.1%) to 85.1% (95% CI 79.8-89.5%) ($p < 0.001$). The pooled specificity decreased from and 91.7% (11/12; 95% CI 61.5-99.8%) to 83.3% (10/12; 95% CI 51.6-97.9%) with concurrent software use.

- **Mixed population (2 studies)**

Hsu 2021,[51] Taiwan - ClearRead CT (Riverain Technologies)

Hsu's study[51] included 150 consecutive cases with lung nodules ≤1 cm or no nodules (93 standard dose CT images from clinical routine and 57 LDCT 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 79-82%) versus 64% (95% CI 62-66%) without software use

(p < 0.001). The mean specificity was 83% (95% CI 82-85%) with software use and 80% (95% CI 78%-81%) without software use (p = 0.25).

In the junior group, the mean per-nodule sensitivity increased significantly from 52% (95% CI 49-55%) without software use to 77% (95% CI 74-79%) with software use (p < 0.001). The mean specificity was 78% (95% CI 76-81%) with and 71% (95% CI 69-74%) without software use (p = 0.152). In the senior group, the mean per-nodule sensitivity was significantly higher with software use: 84% (95% CI 82-86%) compared to 74% (95% CI 72-77%) without software use (p < 0.001). The mean specificity was 88% (95% CI 87-90%) with and 87% (95% CI 85-89%) without software use (p = 0.729).

Takaishi 2021,[55] Japan - ClearRead CT (Riverain Technologies)

Takaishi et al.[55] 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 to 71.7% (99/138) without software use (p = 0.02). The average false positive rate was 21% for both concurrent AI (0.49 FP per scan) and unassisted reading (0.44 FP per scan) (p = 0.98).

Assisted 2nd-read AI vs unassisted reader (1 study)

- **Screening population (1 study)**

Hsu 2021,[51] Taiwan - ClearRead CT (Riverain Technologies)

Hsu's study[51] included 150 consecutive cases with lung nodules ≤1 cm or no nodules (93 standard dose CT images from clinical routine and 57 LDCT 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, 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 LDCT obtained for screening purposes. For all readers, the mean per-nodule sensitivity was significantly higher with software use: 80% (95% CI 77-83%) compared to 63% (95% CI 59-66%) without software use (p < 0.001). The mean specificity was 82% (95% CI 79-84%) with assisted 2nd-read AI and 77% (95% CI 74-80%) without software (p = 0.360).

In the junior group, the mean per-nodule sensitivity increased significantly from 52% (95% CI 47-57%) without software support to 76% (95% CI 72-80) with assisted 2nd-read AI use (p < 0.001). The

mean specificity was 76% (95% CI 72-80%) with and 68% (95% CI 64-73%) without software support (p = 0.333). For the senior group, the mean per-nodule sensitivity improved from 73% (95% CI 69-77%) without software support to 84% (95% CI 80-87%) with assisted 2nd-read software use (p = 0.001). The mean specificity was 88% (95% CI 85-91%) with vs 86% (95% CI 83-90%) without software support (p = 0.795).

- **Mixed population (1 study)**

Hsu 2021,[51] Taiwan - ClearRead CT (Riverain Technologies)

Hsu's reader study[51] retrospectively analysed data from consecutive cases with lung nodules ≤1 cm or no nodules (93 standard dose CT images from clinical routine and 57 LDCT from lung cancer screening) from a hospital in Taiwan. In assisted 2nd-read AI mode, the six readers read the CT images without AI first 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 62-66%) without software use to 82% (95% CI 80-84%) with assisted 2nd-read AI (p < 0.001). The mean specificity was 84% (95% CI 82-85%) using assisted 2nd-read AI compared to 80% (95% CI 78-81%) with unaided reading (p = 0.177).

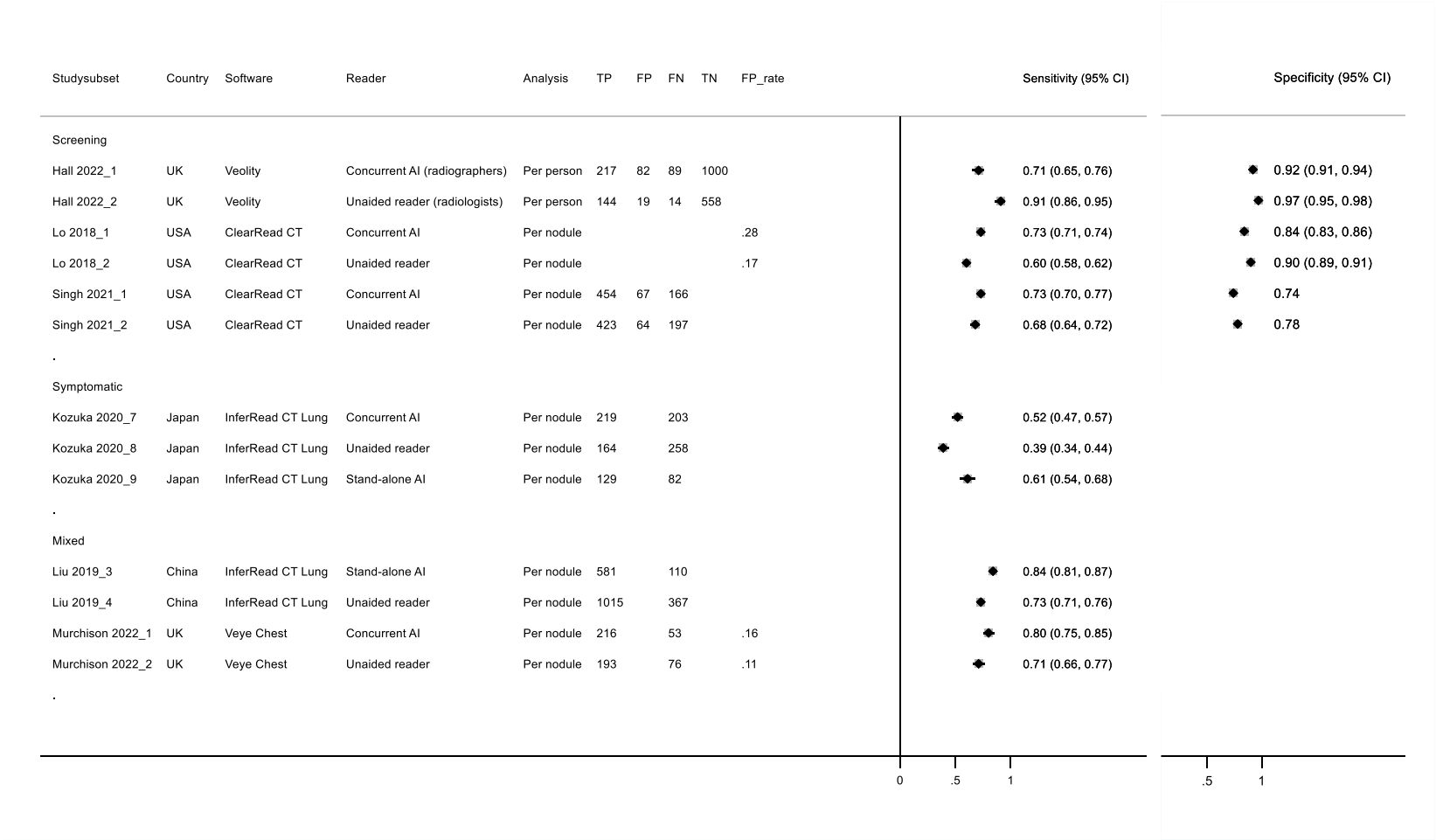
For the three junior readers, the mean per-nodule sensitivity was 79% (95% CI 76-81%) with and 52% (95% CI 49-55%) without software use (p < 0.001). Their mean specificity was 79% (95% CI 77-82%) with and 71% (95% CI 69-74%) without software use (p = 0.088). For the three senior readers, the mean per-nodule sensitivity was 85% (95% CI 83-87%) with and 74% (95% CI 72-77%) without software use (p < 0.001). Their mean specificity was 88% (95% CI 87-90%) with and 87% (95% CI 85-89%) without software use (p = 0.729).

3.3.1.2 Accuracy for detecting actionable nodules

a) Comparative results (6 studies)

Six comparative studies[25, 31, 52, 54, 57, 58] evaluated the **accuracy for detecting actionable nodules (⩾5 or 6 mm)**. Of these, three included a screening population,[25, 52, 54] one included a symptomatic population,[57] and two included mixed populations.[31, 58] Only one study (Hall 2022)[25] reported per person analysis. Key results reported in these studies are shown in **Figure 11**. 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 to 0.73). Only a small number of studies reported specificity or 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 compared with unaided readers (**Figure 11**).

One UK study (Hall 2022)[25] 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 compared with unassisted, experienced radiologists (see **Figure 11**).



TP: true positive; FP: false positive; FN: false negative; TN: true negative

Figure 11. Comparative evidence for accuracy of detecting actionable nodules (6 studies)

Concurrent AI vs unassisted reader (5 studies)

- **Screening population (3 studies)**

Singh 2021,[54] USA - ClearRead CT (Riverain Technologies)

Singh et al.[54] selected 150 LDCT from the US-based National Lung Screening Trial (NLST): the first 125 patients with mixed attenuation or GGNs 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 GGNs, 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) vs 70% (216/308), and the average specificity was 85% vs 76% in vessel-suppressed vs standard CT images, respectively. For all sub-solid nodules, the pooled per-nodule sensitivity was 73% (453/620) vs 68% (423/620), and the mean specificity was 74% vs 78% on vessel-suppressed vs standard CT images.

Lo 2018,[52] USA - ClearRead CT (Riverain Technologies)

Lo's study[52] included 324 LDCT 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 6.2–18.6%) 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 FP nodules/scan to 0.28 FP nodules/scan ($p < 0.01$) with software use.

Hall 2022,[25] UK - Veolity (MeVis)

Hall's study[25] included all 770 LDCT 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 BTS guideline size threshold for warranting surveillance. Therefore, the specificity of the unaided radiologists to identify people without actionable nodules was 96.7% (558/577).

- **Symptomatic population (1 study)**

Kozuka 2020,[57] Japan - InferRead CT Lung (Infervision)

Kozuka et al.[57] 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 (one and five years of experience). The pooled per-nodule sensitivity for the detection of nodules ≥6 mm was 51.9% (219/422) with vs 38.9% (164/422) without software support (calculated by reviewers; no level of significance reported).

- **Mixed population (1 study)**

Murchison 2022,[31] UK - Veye Lung Nodules (Aidence)

Murchison's study[31] 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 datasets 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 75.2-85.0%) with and 71.7% (95% CI 66.0-77.0%) without software use ($p < 0.01$), with an average number of FP detections per image of 0.16 and 0.11, respectively.

Assisted 2nd-read AI vs unassisted reader (No study)

No data available.

3.3.1.3 Accuracy for detecting malignant nodules

Evidence related to the accuracy for detecting malignant nodules is summarised in **Table 10**. It is worth highlighting that direct detection or classification of malignant nodules by AI assisted reading is outside the scope of this assessment. 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[52] 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) for AI-assisted reading compared with unassisted radiologist reading (**Table 10**). False positive detections per image nearly doubled for AI-assisted reading (increased from 0.22 to 0.39).

The other five studies generally reported sensitivity of above 0.70 for the detection of malignant nodules for AI-assisted reading but did not provide information on specificity or false positive detections per image. One study[56] reported high sensitivity (0.94) and low specificity (0.39) for stand-alone AI (**Table 10**). More detailed descriptions of the evidence from individual studies are provided below.

Table 10. Summary of evidence related to accuracy of AI-assisted reading and stand-alone AI for detecting malignant nodules (6 studies)

Study, country, image readers	Malignant nodules/ total scans	Measure of accuracy [a]	Index test [b]	Comparator [b]	Difference	P value for difference
Screening population						
Lo 2018,[52] USA 12 general radiologists (6-26 yrs)	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 2022,[65] USA/South Korea 5 chest radiologists (1-20 yrs)	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 2021a,[49] South Korea	27/4666	Sensitivity	[A] 0.704 (0.498 to 0.862)	NA	NA	NA
Hall 2022,[25] UK 2 radiographers	33/716	Sensitivity [c]	[C] 0.857 (0.746 to 0.933) [e]	NA	NA	NA
Mixed population						
Takaishi 2021,[55] Japan 3 radiologist (2-8 yrs)	1/61	Sensitivity	[C] 1.00 [d]	[D] 1.00 [d]	0	NR
		PPV	[C] 0.020 (1/49)	[D] 0.024 (1/42)	-0.004	NR
Wan 2020,[56] 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

Numbers shown in brackets are 95% confidence intervals unless otherwise stated. NA: not applicable; NR: not reported; PPV: positive predictive value; SD: standard deviation
Technologies evaluated in the studies: Hall 2022: Veolity; Hwang 2021a: AVIEW LCS+; Lo 2018 & Takaishi 2021: ClearRead CT; Park 2022: VUNO Med-LungCT AI;

[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] Only included one malignant nodule, which was detected by both concurrent AI and unaided reader.

[e] Calculated by review authors based on data provided in the original article.

a) Comparative results (3 studies)

Three comparative studies[52, 55, 65] evaluated the **accuracy for detecting malignant nodules**. Of these, two included a screening population,[52, 65] and one included a mixed population.[55]

Concurrent AI vs unassisted reader (3 studies)

- **Screening population (2 studies)**

Lo 2018,[52] USA - ClearRead CT (Riverain Technologies)

The study by Lo et al.[52] 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. Twelve general radiologists certified by the American Board of Radiology (with 6–26 years of experience) participated in a MRMC study. The study found a 15.4% (95% CI 8.2% to 22.5%; p = 2.50 x 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 to 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 2022,[65] USA, Korea - VUNO Med-LungCT AI (VUNO)

Park et al. included a nodule- and cancer-enriched screening population (200 baseline LDCT; 31 cancer cases) selected from the US-based NLST dataset.[65] Five readers participated in the MRMC study. They consisted of one 4th-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 (South Korea). The pooled sensitivity to detect malignant nodules was 91.6% (95% CI 81.7-96.4%) with and 85.2% (95% CI 74.2-92.0%) without software use (p = 0.004).

- **Mixed population (1 study)**

Takaishi 2021,[55] Japan - ClearRead CT (Riverain Technologies)

Takaishi et al.[55] 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, respectively. The PPV was 2.4%

(1/42) without and 2.0% (1/49) with software use (average of 3 readers) (no level of significance reported).

Assisted 2nd-read AI vs unassisted reader (No study)

No data available.

3.3.1.4 Potential effect modifiers of nodule detection accuracy (Sub-question 1).

a) Sub-question 1-1: Effect of contrast use

No subgroup analysis based on contrast use was performed.

b) Sub-question 1-2: Effect of radiation dose (2 studies)

Two studies performed in mixed populations from China[58] and Taiwan,[51] respectively, assessed the effect of radiation dose on nodule detection.

Mixed population - ClearRead CT (Riverain Technologies) (1 study)

The study by Hsu et al.[51] reported accuracy results for the detection of any nodules for both standard dose CT and low dose CT (**Table 11**). It included 150 consecutive cases with lung nodules ≤1 cm or no nodules (93 standard dose CT images from clinical routine and 57 LDCT 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 AUC for both junior and senior readers and all readers ($p > 0.05$).

Table 11. Accuracy for the detection of any nodules in standard dose and low dose CT scans according to Hsu et al.[51]

	Dose of CT	Total scans	Total nodules	Per-nodule sensitivity, % (95% CI)	Per-patient specificity, % (95% CI)
Assisted 2nd-read AI	Standard dose	93	222	83 (81-85)	87 (84-87)
	Low dose	57	118	80 (77-83)	82 (79-84)
Concurrent AI	Standard dose	93	222	81 (79-83)	83 (83-87)
	Low dose	57	118	79 (76-81)	82 (78-84)
Unaided reading	Standard dose	93	222	63 (61-66)	80 (79-83)
	Low dose	57	118	63 (59-66)	72 (74-80)

CI, Confidence interval; CT, Computed tomography.

Mixed population – InferRead CT Lung (Infervision) (1 study)

The study by Liu et al. evaluated 187 LDCT and 942 standard-dose CT images (SDCT).[58] The deep learning-based algorithm (InferRead CT Lung, Infervision) showed no dose-level dependence of nodule detection sensitivity (x2 = 1.1036, p = 0.9538). The same result was observed for the two unaided radiologists (radiologist 1: x2 = 1.6562, p = 0.8944; radiologist 2: x2 = 1.5293, p = 0.9097). The false-positive rate of the stand-alone software was also independent of the dose (x2 = 0.5640, p = 0.4527).

c) Sub-question 1-3: Effect of nodule type (7 studies)

Screening population - Concurrent AI vs unaided reader (2 studies)

Two studies[54, 59] reported detection accuracy for concurrent AI and unaided for different type of nodules (**Table 12**).

Zhang et al.[59] 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% CI 96.5-99.8%) for solid nodules, 100.0% (95% CI 75.3-100.0) for part-solid nodules and 99.1% (95% CI 95.1-99.9%) for GGNs. For the unaided readers in clinical practice, the per-subject sensitivity was 52.4% (95% CI 46.0-58.7%) for solid nodules, 23.1% (95% CI 5.0-53.8%) for part-solid nodules, and 25.2% (95% CI 17.5-34.4%) for GGNs.

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

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 (95% CIs overlap though) and GGNs (no overlap in 95% CIs). Concurrent software use seems to result in bigger sensitivity improvements for part-solid nodules (+76.9%) and GGNs (+73.9%) than for solid nodules (+46.4%).

Singh et al.[54] selected 150 LDCT from the NLST: the first 125 patients with mixed attenuation or GGNs 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. The evaluated software did not possess nodule detection function though. The study reported mean per-nodule sensitivities of 76% for part-solid and 67% for GGNs on vessel-suppressed CT images. On standard CT images, the mean per-nodule sensitivities were 70% for part-solid and 67% for GGNs. The mean specificities were 85% for part-solid nodules and 78.5% for GGNs and 74% for all sub-solid nodules on vessel-suppressed CT images (there might have been a mix up in the table of the article though!). On standard CT images, the mean specificities were 76% for part-solid nodules, 84% for GGNs and 77.5% for all sub-solid nodules.

Symptomatic population - Concurrent AI vs unaided reader (1 study)

Kozuka et al.[57] reported the per-nodule sensitivity of concurrent AI and unaided readers by nodule type (**Table 13**). This study was a retrospective analysis of 120 randomly selected chest CT images (117 cases included in analysis) from cases with lung cancer suspicion. Two less experienced radiologists (one and five 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 29.8-35.6%) for solid, 58.4% (95% CI 49.5-67.0%) for part-solid and 40.1% (95% CI 32.7-47.9%) for GGNs. In the unaided reading session, the pooled per-nodule sensitivity was 18.6% (95% CI 16.3-21.1%) for solid, 31.5% (95% CI 23.7-40.3%) for part-solid nodules and 18.0% (95% CI 12.6-24.6%) for GGNs.

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

Symptomatic population - Stand-alone AI vs unaided reader (1 study)

Kozuka et al.[57] reported per-nodule and per-patient accuracy for stand-alone AI and unaided readers by nodule type (**Table 13**). This study was a retrospective analysis of 120 randomly selected chest CT images (117 cases included in analysis) from cases with lung cancer suspicion. Two less experienced radiologists (one and five 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 63.9–72.1%) for solid, 70.8% (95% CI 58.2–81.4%) for part-solid and 72.1% (95% CI 61.4–81.2%) for GGNs. For the unaided readers, the pooled per-nodule sensitivity was 18.6% (95% CI 16.3-21.1%) for solid, 31.5% (95% CI 23.7-40.3%) for part-solid nodules and 18.0% (95% CI 12.6-24.6%) for GGNs.

Table 12. Effect of nodule type on nodule detection accuracy in screening populations - Concurrent AI vs unaided reader (2 studies)

Author / Year, Software	Nodule type	# scans	# nodules	Sensitivity, % (95% CI)		Specificity, % (95% CI)	
				Concurrent CAD	Unaided reader	Concurrent CAD	Unaided reader
Zhang 2021,[59]	Solid nodules	250	NR	98.8 (96.5-99.8)	52.4 (46.0-58.7)	99.2 (98.1-99.7)	100.0 (99.4-100)
InferRead CT Lung	Part-solid nodules	13	NR	100.0 (75.3-100)	23.1 (5.0-53.8)	100.0 (99.6-100)	100.0 (99.6-100)
(Infervision)	Ground glass nodules	111	NR	99.1 (95.1-99.9)	25.2 (17.5-34.4)	98.8 (97.7-99.5)	100.0 (99.5-100)
Singh 2021,[54]	Sub-solid nodules	NR	310	73	68	74	77.5
ClearRead Vessel	Part-solid nodules	NR	154	76	70	85	76
Suppression (Riverain	Ground glass nodules	NR	156	67	67	78.5	84
Technologies)							

CAD, Computer-aided detection; CT, Confidence interval; NR, Not reported.

Table 13. Effect of nodule type on nodule detection accuracy in a symptomatic population - Concurrent AI / stand-alone AI vs unaided reader (1 study)

Author / Year, Software	Nodule type	# scans	# nodules	Per-nodule sensitivity, % (95% CI)		
				Stand-alone AI	Concurrent AI (pooled 2 readers)	Unaided reader (pooled 2 readers)
Kozuka 2020,[57]	Solid nodules	NR	518	68.1 (63.9-72.1)	32.6 (29.8–35.6)*	18.6 (16.3–21.1)
InferRead CT Lung	Part-solid nodules	NR	65	70.8 (58.2–81.4)	58.5 (49.5–67.0)*	31.5 (23.7–40.3)
(Infervision)	Ground glass nodules	NR	86	72.1 (61.4–81.2)	40.1 (32.7–47.9)*	18.0 (12.6–24.6)

AI, Artificial intelligence; CI, Confidence interval; NR, Not reported.

* $p < 0.01$ versus unaided reader.

Screening population - Stand-alone AI (2 studies)

Hwang et al.[49] included 4,666 participants who had undergone lung cancer screening as part of the K-LUCAS project after the implementation of the software AVIEW Lungscreen (Coreline Soft). The per-nodule sensitivity of stand-alone AI was 51% (95% CI 50-53%) for solid nodules, 49% (95% CI 36-61%) for part-solid nodules and 21% (95% CI 16-29) for GGNs (**Table 14**).

The study by Lo et al.[52] 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, part-solid and GGNs, respectively (**Table 14**).

Table 14. Effect of nodule type on nodule detection accuracy in screening population - Stand-alone AI (2 studies)

Author / Year, Software	Nodule type	# scans	# nodules	Sensitivity, % (95% CI)
Hwang 2021,[49] AVIEW Lungscreen (Coreline Soft)	Solid nodules	NR	4,032	51 (50-53)
	Part-solid nodules	NR	70	49 (36-61)
	Ground glass nodules	NR	178	21 (16-29)
Lo 2018,[52] 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.

Mixed population - Stand-alone AI alone (1 study) or versus unaided reader (1 study)

Liu et al.[58] 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 and low dose CT scans **(Table 15).** On 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 sub-solid 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 sub-solid nodules ≤5 mm and 81.1% for sub-solid nodules >5 mm.

Martins et al.[64] 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, sub-solid and mixed (solid/sub-solid) nodules, respectively **(Table 15)**.

Table 15. Effect of nodule type on nodule detection accuracy in mixed populations - Stand-alone AI alone (1 study) or vs unaided readers (1 study)

Author / Year, Software	Nodule size and dose	Nodule type	# nodules	Per-nodule sensitivity, %		
				Stand-alone AI	Unaided Reader 1	Unaided Reader 2
Liu 2018,[58] InferRead CT Lung (Infervision)	>6 mm, Conventional dose	Solid nodules	215	87.9	77.2	69.3
	≤6 mm, Conventional dose	Solid nodules	2,680	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
	>5 mm, Conventional dose	Sub-solid nodules	371	81.1	58.2	85.2
	≤5 mm, Conventional dose	Sub-solid nodules	993	68.1	26.2	56.9
	>5 mm, Low dose	Sub-solid nodules	61	85.2	67.2	82.0
	≤5 mm, Low dose	Sub-solid nodules	333	61.3	22.5	56.2
Martins Jarnalo 2021,[64] Veye Chest (Aidence)	4-30 mm	Solid nodules	73	89.0	NA	NA
	4-30 mm	Sub-solid	16	81.3	NA	NA
	4-30 mm	Mixed (solid / sub-solid)	2	100.0	NA	NA

d) Sub-question 1-4. Effect of patient's ethnicity

No subgroup analysis based on ethnicity was performed.

e) Sub-question 1-5. Effect of radiologist speciality & experience (1 study)

Hsu 2021,[51] Taiwan - ClearRead CT (Riverain Technologies)

The study in a mixed population (with data for the screening subgroup reported separately) reported accuracy for detecting any nodules by 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 47-57%) without software use to 74% (95% CI 70-78%) with concurrent AI ($p < 0.001$). The mean specificity did not change significantly and was 74% (95% CI 70-78%) with vs 68% (95% CI 64-73%) 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 69-77%) to 83% (95% CI 79-86%) ($p < 0.01$). The mean specificity was 88% (95% CI 85-91%) with and 86% (95% CI 83-90%) without software use ($p = 0.795$).

f) Sub-question 1-6: For the incidental population, effect of reason for CT scan

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

3.3.1.5 Sub-question 2: Concordance and variability in nodule detection

a) Concordance between readers with and without software (1 study)

No study was identified that reported on the concordance in nodule detection between readers with and without software use. However, one study reported on the percentage agreement in nodule detection between stand-alone AI and the original unaided reading.[45]

Mixed population – AI-RAD Companion CT Chest (Siemens Healthineers) (1 study)

Abadia et al. 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).[45] The highest agreement in nodule detection between AI-RAD software and the original radiology reports was achieved in the sub-population 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 (1 study)

One study reported on the inter-reader variability between unaided readers in the detection of the risk-dominant nodule.[54]

Screening population – Unaided readers (1 Study)

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

3.3.2 Nodule type determination

3.3.2.1 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 13.5.4.

3.3.2.2 Sub-questions 1 to 6: Potential factors influencing nodule type determination

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

3.3.2.3 Sub-question 2: 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 (2 studies)

Two MRMC studies 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.[62, 65] Both studies found that software use did not affect the proportion of disagreements in nodule type between the readers.

Screening population – Veolity (MeVis) (1 study)

Jacobs et al.[62] 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 x 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) (1 study)

Park et al.[65] reported that for all 2,000 possible paired observations among the five readers (10 reader pairs x 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).

3.3.3 Nodule diameter measurement

3.3.3.1 Accuracy of measurement (3 studies)

Three studies compared diameter measurements of stand-alone software[54, 64] or readers with concurrent software use[53] to the measurements of a reference standard. The studies were performed in a screening population,[54] a mixed population[64] and a population with unclear indication for the chest CT scan,[53] respectively. Results on the diameter measurement accuracy of stand-alone software are inconsistent with one study reporting significantly smaller nodule diameters measured by the software,[53] while the other study reported that in 83% of size disagreements, the nodule size was overestimated by the software.[64] Substantial agreement with the reference standard was reported for semi-automated longest diameters measured on vessel-suppressed CT images in the third study.[53]

Table 16. Main findings, risk of bias, applicability concerns and input into modelling

Study, AI software, country	**Population, design & sample**	**Main findings**	**Risk of bias (RoB), applicability concerns (AppC) & input into modelling (Model)**
Singh 2021[54] ClearRead CT USA	Screening, MRMC, nodule-enriched sample, Risk-dominant sub-solid nodule (n=100)	Average diameter [a] **Stand-alone AI**: Mean 12, SD 3 mm **Radiologist consensus:** [b] Mean 14, SD 5 mm P=0.02	**RoB**: Research setting; excluded scans which could not be processed by the software (n=27) **AppC**: Research setting; sub-solid nodules only **Model**: no; Stand-alone AI rather than concurrent AI

Study, AI software, country	Population, design & sample	Main findings	Risk of bias (RoB), applicability concerns (AppC) & input into modelling (Model)
Martins Jarnalo 2021[64] Veye Chest Netherlands	Mixed, retrospective test accuracy study, randomly selected sample, 80 nodules (all nodule types, 4-30 mm).	Diameter measurements **Stand-alone AI vs unaided radiologist consensus**: [d] Agreement (same millimetre): 67.5% (54/80) +1 mm: 20.0 % (16/80) +2 mm: 2.5% (2/80) +4 mm: 1.25% (1/80) -1 mm: 2.5% (2/80) -2 mm: 2.5% (2/80) Failure: 3.75% (3/80)	**RoB**: research setting; scans with >5 nodules were excluded. **AppC**: single hospital; stand-alone AI rather than concurrent AI **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 **section 13.8.1.4**)
Milanese 2018[53] ClearRead CT for vessel suppression; MM Oncology for semi-automatic 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:** [c] 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 **AppC**: Research setting; population characteristics unclear; solid nodules only; radiologists <5 years of experience; AI software only used for vessel suppression, not for measurement **Model**: no; Lin's CCC does not allow the derivation of relative measurement accuracy or precision

AppC: applicability concerns; EAG: External assessment group; 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] Compared with reference standard, which was the average semi-automatic measurements by the two readers on standard CT images (without AI for vessel-suppression). Radiologists 1 and 2 had 3 and 1 year of experience in chest CT, respectively.

a) Non-comparative results (3 studies)

Screening population – ClearRead CT (Riverain Technologies) (1 study)

In a nodule-enriched screening population, Singh et al. found that for the same risk-dominant, sub-solid 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, SD 3 mm) compared to the reference standard measurement obtained by consensus reading of two

experienced chest radiologists, with a third experienced radiologist resolving discrepancies (mean 14, SD 5 mm) (p=0.02).[54]

Mixed population – Veye Chest (Aidence) (1 study)

Martins Jarnalo et al. compared the diameter measurements of stand-alone software (Veye Chest, Aidence) to a reference standard of consensus reading of one experienced radiologist and one resident radiologist, with discrepancies resolved by a third experienced chest radiologist.[64] 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 another 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; for one (1 mm discrepancy) there were surrounding spiculae, and another (2 mm discrepancy) was a cavitating nodule.

Unclear indication for CT scan – ClearRead CT (Riverain Technologies) (1 study)

Milanese et al. 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 semi-automatic 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 longest diameters measured on standard CT images by the two readers used as reference standard.[53] 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 semi-automated longest 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 values near 1 meaning perfect concordance).

3.3.3.2 Sub-questions 1 to 6: potential factors associated with nodule diameter measurement accuracy and precision

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

3.3.3.3 Sub-question 2: Concordance and variability in nodule diameter measurement

a) Concordance between readers with and without software (4 studies)

One study was identified that evaluated the concordance of nodule diameter measurements between readers with and without software in patients with previously detected sub-solid nodules (surveillance population with applicability concerns).[61] Another three studies reported on the concordance of stand-alone software measurements compared to manual diameter measurements in mixed populations.[31, 45, 56]

The studies found similar[56, 61] or significantly larger[45] nodule diameters with semi-automatic measurements compared to manual measurements. Two studies reported a significant correlation between the measurements.[45, 56] One study concluded that the segmentation of pulmonary nodules of stand-alone software and the resulting diameter measurements are comparable to manual measurement performed by experienced thoracic radiologists.[31]

Surveillance population with applicability concerns – Veolity (MeVis) (1 study)

Kim et al. included 89 patients with sub-solid nodules who underwent preoperative non-enhanced CT and subsequent surgical resection.[61] The diameter of the 102 sub-solid nodules was not statistically different between the semi-automated 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 semi-automated and manual measurements for Reader 1 (6.3±4.9 mm vs 6.3±4.9 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 semi-automated diameter measurements being larger than manual measurements.

Mixed population - AI-RAD Companion Chest CT (Siemens Healthineers) (1 study)

Abadia et al. 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 (IQR 6.5 to 11.5) by AI-RAD and 6.6 mm (IQR 5.0 to 9.5) by the expert radiologist ($p<0.0001$).[45] 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) (1 study)

The UK-based reader study by Murchison et al. 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.[31] 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 3-30 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, 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), which was similar to the geometric mean difference observed between the individual expert radiologists (1.15 mm [95% CI 1.00, 1.58]).

Mixed population - ClearRead CT (Riverain Technologies) (1 study)

Wan et al. included LDCT images from 50 Taiwanese patients with mixed indications who had subsequent excision of their nodule(s).[56] 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 (7.83 ± 3.06 mm versus 8.13 ± 3.49 mm, mean ± SD, $p = 0.624$) with a Pearson correlation coefficient of 0.926.

b) Concordance between readers using different software (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 (5 studies)

Inter-observer variability (5 studies)

Five MRMC studies were identified that reported on the inter-observer variability in nodule diameter measurements.[31, 60-62, 65] Three of them compared the inter-reader variability between manual diameter measurements and semi-automatic measurements and consistently found reduced disagreements in nodule sizes between readers with software use.[61, 62, 65] The variability between readers using semi-automatic software was similar in CT images reconstructed with filtered back projection (FBP) and images reconstructed with model-based iterative reconstruction (MBIR) algorithms.[60]

Screening population – Veolity (MeVis) (1 study)

The study by Jacobs et al. included a nodule-enriched screening population.[62] 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 that there were 67% (207 vs 68) fewer Lung-RADS category disagreement pairs that were 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) (1 study)

Park et al. included a nodule- and cancer-enriched screening population (200 baseline LDCT) selected from the US-based NLST dataset.[65] 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) (2 studies)

Two studies were performed at the same hospital in Korea and included (potentially overlapping) surveillance populations with applicability concerns: 89[61] and 73 patients,[60] respectively, with preoperative CT scans for sub-solid nodules. In both MRMC studies, two radiologists with concurrent use of the software Veolity (MeVis) independently performed nodule diameter measurements, but only one study[61] compared semi-automatic with manual diameter measurements.

Kim et al. found that in 102 sub-solid nodules measured by semi-automated segmentation software, the inter-reader variability of two experienced radiologists ranged from -1.9 mm (95% CI -2.3 to -1.6) to 2.1 mm (95% CI 1.7–2.4) for the whole nodule diameter and from -2.1 mm (95% CI -2.5 to -1.8) to 2.1 mm (95% CI 1.7–2.5) for the solid portion diameter.[61] With manual measurement, inter-reader variability ranged from -2.8 mm (95% CI -3.3 to -2.4) to 2.4 mm (95% CI 2.0 to 2.9) for the whole nodule diameter and from -5.1 mm (95% CI -5.7 to -4.4) to 2.8 mm (95% CI 2.1 to 3.5) for the solid portion diameter. The inter-reader variability of semi-automatic measurement was significantly lower than those of manual measurement for both the whole nodules as well as the solid portion diameters ($p<0.001$ for all).

Cohen et al. compared semi-automatic measurement using CT images reconstructed with FPB and MBIR reconstruction algorithm. This study did not include a 'manual measurement' comparator. Regarding the semi-automatic measurement of the longest diameter of the whole sub-solid 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 reconstruction algorithm.[60] For the diameter of the solid component of the sub-solid 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, for MBIR. There were no significant differences concerning inter-reader variability between FBP and MBIR reconstructed CT images ($p > 0.05$).

Mixed population – Manual measurement (1 study)

The UK-based reader study by Murchison et al. 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.[31] 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 3-30 mm, the average inter-reader Dice coefficient for nodule segmentation was 0.83 (95% CI: 0.39, 0.96), and the geometric mean diameter difference of the largest axial diameter was 1.15 mm (95% CI 1.00, 1.58).

Reproducibility/repeatability (2 studies)

Two studies reported on the intra-reader variability in nodule diameter measurements in patients with previously detected sub-solid nodules.[60, 61] The intra-reader variability with semi-automatic measurement was significantly lower compared to manual measurement for the whole nodule diameter and the solid portion diameter, respectively,[61] and was similar between FBP and MBIR reconstructed CT images.[60]

Surveillance population with applicability concerns – Veolity (MeVis) (2 studies)

The two MRMC studies were both performed at the same hospital in Korea and included (potentially overlapping) surveillance populations with applicability concerns: 89[61] and 73 patients,[60] respectively, with preoperative CT scans for sub-solid nodules.

In the study by Kim et al., one experienced radiologist performed the nodule diameter measurements twice with concurrent use of the software Veolity (MeVis), and twice without software use in 102 sub-solid nodules.[61] With semi-automatic 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 semi-automatic 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., two radiologists with four and five years of experience performed the semi-automatic measurements with concurrent use of the software Veolity (MeVis), twice on FBP

reconstructed CT images and twice on FBIR reconstructed CT images.[60] In 66 sub-solid 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) between FBP and MBIR reconstructed CT images.

3.3.4 Nodule volume measurement

3.3.4.1 Accuracy in nodule volume measurement (1 study)

One MRMC study reported on the accuracy of volume measurement in solid nodules and found substantial agreements of semi-automated volumetric measurements in vessel-suppressed CT images with the reference standard.[53] The percentages of error of semi-automated volumetric measurement were similar between standard CT images and vessel-suppressed CT images.

a) Comparative results – Reader with and without software (1 study)

Unclear indication for chest CT scan – ClearRead CT (Riverain Technologies) (1 study)

This MRMC study included 93 consecutive patients referred for clinical non-enhanced, LDCT (unclear indication for the chest CT scan).[53] One radiologist with three years of experience in chest CT and a radiology resident independently performed semi-automatic 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 reference standard. Agreement of semi-automatic 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 datasets). 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. concluded that vessel-suppressed CT datasets can be used for semi-automated measurements of solid pulmonary nodules.

3.3.4.2 Sub-questions 1 to 6: Potential factors associated with nodule volume measurement

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

3.3.4.3 Sub-question 2: Concordance and variability in nodule volume measurement

a) Concordance between readers with and without software (1 study)

No study was identified that reported on the concordance of volume measurements between readers with and without software. However, one study evaluated the concordance of volume measurements between stand-alone software and unaided readers.[31] 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) (1 study)

The UK-based reader study by Murchison et al. 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.[31] Nodules were manually segmented (Apple Pencil) by two or three experienced thoracic radiologists. Software segmentation was successful in 95% of 428 nodules of all types between 3-30 mm. The average Dice coefficient between Veye Chest's and each individual radiologist's segmentation was 0.86 (95% CI 0.51, 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, 3.38), which was similar to the volume difference observed between the expert radiologists (1.39 mm^3 [95% CI 1.01, 3.19]).

b) Concordance between readers using different software or 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 (3 studies)

Inter-observer variability (3 studies)

Three MRMC studies reporting on the inter-observer variability in nodule volume measurement were identified.[31, 53, 60] Between-readers agreement using semi-automatic software was almost perfect on both standard CT images as well as vessel-suppressed CT images.[53] The inter-reader variability of semi-automatic volumetric measurement was similar between FBP and MBIR reconstructed CT images.[60] The third study only reported inter-observer agreement between unaided readers.[31]

Surveillance population with applicability concerns – Veolity (MeVis) (1 study)

The study by Cohen et al. included a surveillance population with applicability concerns: 73 patients with preoperative CT scans for sub-solid nodules.[60] Two radiologists with four and five years of experience independently performed the semi-automatic measurements with concurrent use of the software Veolity (MeVis) on FBP reconstructed CT images as well as on MBIR reconstructed CT images. In 66 sub-solid 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%) for 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%) on FBP and -21.4 mm^3 (-12.7%) for MBIR (p = 0.11).

Mixed population – Unaided readers (1 study)

The UK-based reader study by Murchison et al. 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.[31] Nodules were manually segmented (Apple Pencil) by two or three experienced thoracic radiologists. In 428 nodules between 3-30 mm, the average Dice coefficient between each reader's segmentation and the segmentation from the other readers was

0.83 (95% CI 0.39, 0.96). The geometric mean volumetric discrepancy between radiologists was 1.39 mm^3 (95% CI 1.01, 3.19).

Unclear indication for CT scan – ClearRead CT (Riverain Technologies) (1 study)

This MRMC study by Milanese et al. included 93 consecutive patients referred for clinical non-enhanced, chest LDCT (unclear indication).[53] One radiologist with three years of experience in chest CT and a radiology resident independently performed semi-automatic 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 as well as 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 8 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 (1 study)

One study reported on the reproducibility of semi-automatic volume measurements and found similar intra-reader variability in FBP and MBIR reconstructed CT images, respectively.[60]

Surveillance population with applicability concerns – Veolity (MeVis) (1 study)

Cohen et al. included 73 patients with preoperative CT scans for sub-solid nodules from a single hospital in Korea.[60] Two radiologists performed the semi-automatic measurements with concurrent use of the software Veolity (MeVis), twice on FBP reconstructed CT images and twice on MBIR reconstructed CT images.[60] In 66 sub-solid 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% for FBP and -5.9% for MBIR ($p = 0.07$). Intra-observer variability was similar ($p > 0.05$) between FBP and MBIR reconstructed CT images.

3.3.5 Classification into risk categories based on nodule type and size

3.3.5.1 Accuracy for risk classification based on 2015 BTS guidelines (1 study)

One study reported on the performance of readers with and without concurrent software use for identifying patients classed as BTS grade A (discharge recommended) at consensus.[32] This study also reported on the agreement in nodule management recommendations (4 grades based on the 2015 BTS guidelines[11]) 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 in software-aided readers were higher compared to 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 done unaided, but no level of significancy or 95% CIs were reported.

a) Comparative results – Reader with and without software (1 study)

Mixed population – Veye Chest (Aidence) (1 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.[32] 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[11] (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 reference standard. With concurrent use of Veye Chest software, the sensitivities and specificities to identify patients with BTS grade A (no clinical follow-up required) were higher for both readers, but 95% CIs overlapped (see **Table 17**).

Table 17. Accuracy of readers with and without concurrent use of Veye Chest to identify patients with BTS grade A (no clinical follow-up recommended)[32]

	Sensitivity (95% CI)		Specificity (95% CI)	
	Unaided	**Aided**	**Unaided**	**Aided**
Reader 1	0.83 (0.61-0.95)	0.85 (0.66–0.96)	0.85 (0.66–0.96)	1.00 (0.85–1.00)
Reader 2	0.76 (0.55-0.91)	0.92 (0.73–0.99)	0.84 (0.64–0.95)	0.96 (0.80-1.00)

CI, Confidence interval.

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 of 95% CIs was reported.

3.3.5.2 Accuracy for risk classification based on other risk categories (2 studies)

Two studies were identified that evaluated the accuracy of stand-alone software[30] and software-assisted readers[30, 53] in classifying solid nodules into other risk categories based on volume.

One study was performed in a selected screening population,[30] and in the other study, the indication for the chest CT scan was not reported.[53]

"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 semi-automatic volumetric measurements in vessel-suppressed CT images in one study using three volume-based risk categories.[53] Using two volume-based risk categories, another study found misclassifications by stand-alone software in 22% and by software-assisted readers in 10% to 15% of cases.[30]

a) Comparative results – Reader with and without software (1 study)

Screening population – AVIEW LCS (Coreline Soft) (1 study)

Lancaster et al. included 283 participants who underwent a baseline ultra-LDCT thorax scan and had at least one solid nodule of any size.[30] In a MRMC study, five thoracic radiologists with more than seven years of experience independently interpreted the CT images with visual nodule detection and software use for semi-automated 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) to automatically detect, measure, and classify solid nodules was also evaluated. As reference standard, an independent consensus read was performed by a panel of three radiologists with more than 10 years' experience and an experienced IT technologist of the 283 largest nodules. Compared to the reference standard, the stand-alone software had 61 (21.6 %; 53 false positive, 8 false negative) misclassifications reported, compared to 43 discrepancies (15.1 %; 22 false positive, 21 false negative) for Reader 1, 36 (12.7 %; 25 false positive, 11 false negative) for Reader 2, 29 (10.2 %; 25 false positive, 4 false

negative) for Reader 3, 28 (9.9 %; 6 false positive, 22 false negative) for Reader 4 and 50 (17.7 %; 15 false positive, 35 false negative) discrepancies for Reader 5.

b) Non-comparative results (1 study)

Unclear indication for CT scan – ClearRead CT (Riverain Technologies) (1 study)

The MRMC study by Milanese et al. included 93 consecutive patients referred for clinical non-enhanced, low-dose chest CT (unclear indication).[53] One radiologist with three years of experience in chest CT and a radiology resident independently performed semi-automatic 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.[66] 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 reference standard. The agreement between the Fleischner management categories[66] based on semi-automated volumetric measurements performed on vessel-suppressed CT images and the reference standard was reported as "excellent" (**Table 18**).

Table 18. Risk categorisation using standard CT images and vessel-suppressed CT images for semi-automatic volume measurement (modified from Milanese et al.[53])

		Reader 1	Reader 2	Reference standard
Semi-automatic measurement on standard CT images	<100 mm^3 100-250 mm^3 >250 mm^3	48 (73.8%) 11 (16.9%) 6 (9.2%)	48 (73.8%) 11 (16.9%) 6 (9.2%)	49 (75.4%) 10 (15.4%) 6 (9.2%)
Semi-automatic measurement on vessel-suppressed CT images	<100 mm^3 100-250 mm^3 >250 mm^3	50 (76.9%) 9 (13.8%) 6 (9.2%)	49 (75.4%) 9 (13.8%) 7 (10.8%)	NA NA NA

3.3.5.3 Sub-questions 1 to 6: Potential factors associated with risk classification

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

3.3.5.4 Sub-question 2: Concordance and variability in risk classification

a) Concordance between readers with and without software use (2 studies)

One study was identified that reported on the concordance in Lung-RADS categorisation between readers with and without software use.[62] A second study reported on the concordance in Lung-RADS categorisation between stand-alone software and readers with and without software use.[65] Both studies were performed in nodule-enriched screening populations. Agreement in Lung-RADS categorisation between each reader with and without software as assessed by mean Cohen weighted k value was 0.67.[62] The agreement between stand-alone software and each reader increased with software use.[65]

Screening population – Veolity (MeVis) (1 study)

In the study by Jacobs et al., 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.[62] The intra-observer agreement in Lung-RADS categorisation for each reader with and without software use was assessed using mean Cohen weighted k value and constituted 0.67 (range: 0.59 to 0.76 for individual readers).

Screening population – VUNO Med-Lung CT AI (VUNO) (1 study)

Park et al. investigated the agreement in nodule Lung-RADS categorisation between stand-alone software and five readers with and without software use (VUNO Med-Lung CT AI from VUNO).[65] 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, 0.68) to 0.70 (95% CI 0.62, 0.78).

b) Concordance between readers using different software (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 (5 studies)

Inter-reader variability (5 studies)

Categorisation based on 2015 BTS guidelines (1 study)

One study[32] reported on the inter-reader agreement in nodule management recommendations based on the 2015 BTS guidelines.[11] 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) (1 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.[32] 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[11] (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) compared to unaided readers (linear weighted kappa 0.61), but no level of significance or 95% CIs were reported.

Categorisation based on Lung-RADS categories (2 studies)

Two studies were identified that reported on the inter-reader variability in nodule Lung-RADS categorisation.[62, 65] Both studies were performed in nodule-enriched screening populations and found marginally improved[65] and improved[62] inter-reader agreement with software use.

Screening population – Veolity (MeVis) (1 study)

In the study by Jacobs et al., 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.[62] When using the standard PACS-like viewer without software support, the inter-reader agreement in Lungs-RADS categorisation had a Fleiss k value of 0.58 (95% CI 0.55 to 0.60). When readers were using the dedicated CT lung screening viewer with Veolity software, the Fleiss k value increased to 0.66 (95% CI 0.64 to 0.68). The mean pairwise Cohen weighted k 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 with using the standard PACS-like viewer.

Screening population – VUNO Med-Lung CT AI (VUNO) (1 study)

In the study by Park et al., five readers assessed the LDCT images with and without software use (VUNO Med-Lung CT AI from VUNO).[65] Inter-reader agreement of five readers for Lung-RADS categorisation as assessed by Fleiss kappa was 0.60 (95% CI 0.57-0.63) without software use, and improved marginally to 0.65 (95% CI 0.63-0.68) 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 (2 studies)

Two studies were identified that reported on the inter-reader variability in categorising sub-solid nodules according to Fleischner's guidelines[68] into (1) pure ground glass, (2) part-solid with a solid component ≥5 mm or (3) part-solid with a solid component less than 5 mm.[60, 61] Semi-automatic segmentation was able to significantly improve inter-reader variability compared to manual measurement (p = 0.022), especially the subclassification of part-solid nodules according to the diameter of the solid portion.[61] The inter-observer agreement for semi-automated measurements performed on FBP and MBIR reconstructed CT images was not statistically significant (p = 0.22).[60]

Surveillance populations with applicability concerns – Veolity (MeVis) (2 studies)

Both studies were performed at the same hospital in Korea and included (potentially overlapping) surveillance populations with applicability concerns: 89[61] and 73 patients,[60] respectively, with preoperative CT scans for sub-solid 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., the two readers were also assessing CT images without software use performing manual diameter measurement.[61]

In the study by Kim et al., the inter-reader variability (kappa) regarding the classification of 102 sub-solid nodules was 0.861 (95% CI 0.769 to 0.953) for semi-automatic measurement and 0.683 (95% CI 0.561 to 0.805) for manual measurement (p=0.022).[61] Percentage inter-reader agreement was 92.2% (94/102) for semi-automatic measurement and 80.4% (82/102) for manual measurement.

Cohen et al. found that the inter-observer variability in categorising 66 sub-solid nodules as assessed by Kappa values was 0.66 and 0.77 for FBP and MBIR, respectively.[60] The inter-observer agreement for both image reconstruction algorithms was not statistically significant (p = 0.22).

Repeatability/reproducibility (2 studies)

Two studies were identified that reported on the intra-reader reproducibility in categorising sub-solid nodules according to Fleischner's guidelines[68] into (1) pure ground glass, (2) part-solid with a solid component ≥5 mm or (3) part-solid with a solid component less than 5 mm.[60, 61] One study reported a significantly higher intra-reader reproducibility with semi-automatic measurement compared to manual measurement.[61] In readers with semi-automatic measurement, the intra-reader agreement was significantly higher with MBIR compared to FPB reconstructed images.[60]

Surveillance populations with applicability concerns – Veolity (MeVis) (2 studies)

Both studies were performed at the same hospital in Korea and included (potentially overlapping) surveillance populations with applicability concerns: 89[61] and 73 patients,[60] respectively, with preoperative CT scans for sub-solid nodules.

In the reader study by Kim et al., one experienced radiologist performed the nodule diameter measurements twice with concurrent use of the software Veolity (MeVis), and twice without software use in 102 sub-solid nodules.[61] The intra-reader reproducibility (kappa) of nodule classification was 0.894 (95% CI 0.812 to 0.976) for semi-automatic 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., two radiologists with four and five years of experience performed the semi-automatic measurements with concurrent use of the software Veolity (MeVis), twice on FBP reconstructed CT images and twice on MBIR reconstructed CT images.[60] The intra-observer reproducibility (kappa) for the classification of the 66 sub-solid 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).

3.3.6 Whole read

3.3.6.1 Accuracy for lung cancer detection based on whole read (2 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[48, 49] or without[49] concurrent software use (AVIEW, Lungscreen, Coreline Soft) in a prospective screening population from Korea. Positivity was based on Lung-RADS category 3 or higher, and the reference standard was medical record review. The comparative study did not find a statistical difference in sensitivity, specificity, PPV and NPV before and after software implementation, when measurements were performed on transverse planes. After software implementation, PPVs differed significantly according to measurement planes used (transverse, maximum orthogonal, any maximum).

a) Comparative results – Reader with and without software (1 study)

Screening population – AVIEW Lungscreen (1 study)

In a before-after study, Hwang et al.[49] included 6,487 consecutive participants of the K-LUCAS project: 1,821 participants were screened before the implementation of the software (including 16 cases of lung cancer) and 4,666 participants received screening after the implementation of the software (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 implementation of the AVIEW Lungscreen software and 93.5% after the implementation (p = 0.979). The specificity was 90.9% before and 89.6% after the implementation of the software (p = 0.132). There were also no significant differences in PPV and NPV (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 PPVs

(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 (1 study) are reported in **Appendix 13.5.5.**

3.3.6.2 Sub-questions 1 to 4: Factors potentially associated with accuracy for lung cancer detection based on whole read

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

3.3.6.3 Sub-question 5: Concordance and variability

No evidence was identified for sub-questions 5a) – 5c).

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

3.4.1 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[11] or other thresholds.

3.4.2 Nodule registration and growth assessment

3.4.2.1 Accuracy of nodule registration (1 study)

No study was identified that compared the accuracy of nodule registration between readers with and without AI software use. However, Murchison et al. 2022 evaluated the accuracy of stand-alone AI (Veye Chest, Aidence) to detect nodule pairs in subsequent scans of the same patient.[31] The study found a sensitivity for detecting nodule pairs of 100.0% (23/23) with no false positive pairs (see **Appendix 13.5.6.1**).

3.4.2.2 Sub-questions 1 to 4: Factors potentially associated with accuracy of nodule *registration or growth rate estimation*

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

3.4.2.3 Sub-question 5

a) Concordance between readers with and without AI software use (1 study)

No study was identified that reported on the concordance of readers with and without AI software use. However, the same study mentioned above (Murchison et al. 2022) reported on the mean growth percentage difference between stand-alone AI (Veye Chest, Aidence) and unaided expert radiologists.[31] The geometric mean growth rate difference was similar between stand-alone AI and unaided readers. However, due to a single incorrect segmentation of the stand-alone AI, the upper end of its confidence interval is twice as high compared to that of readers, illustrating that visual verification of the nodule segmentation by human readers is still advised (see Appendix 13.5.6.2).

b) Concordance between readers using different software (No study)

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

c) Intra-reader and inter-reader variability (1 study)

One study was identified that reported on the inter-reader variability in nodule growth assessment between unaided readers.[31] The mean growth rate difference for 23 nodule pairs between two unaided expert radiologists was 1.30%.

Mixed population – Unaided readers (1 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). 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, 2.21).

3.5 Practical implications

3.5.1 Technical failure rate (12 studies)

Twelve records were identified that reported on the technical failure rate of AI-based software assessing chest CT images.[25, 29, 31, 32, 48-50, 54, 60-62, 64] Six studies were performed in a screening population,[25, 48-50, 54, 62] two studies were performed in a surveillance population with applicability concerns,[60, 61] and the remaining four studies included mixed populations.[29, 31, 32, 64] The identified studies used five different technologies: Veye Chest (Aidence) as stand-alone software[31, 64] or in concurrent mode,[32, 60, 61] Veolity (MeVis) in concurrent mode,[25, 60-62] ClearRead CT (Riverain Technologies) as stand-alone software,[54] AVIEW Lungscreen (Coreline Soft) in concurrent mode,[48-50] and contextflow SEARCH Lung CT (contextflow) in concurrent mode.[29] Segmentation failure ranged from 0% to 57% of nodules (8 studies; see **Table 19**). However, one study discussed that the observed nodule segmentation failure is mostly due to 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%) compared to solid nodules (7%) (1 study). Manual modifications of the segmentation were required in 29% to 59% of nodules (2 studies).

Screening population – Veolity (MeVis) (2 studies)

The MRMC study by Jacobs et al. included a nodule-enriched screening population.[62] 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; 1 observer) or 1.3% (2/160; 2 observers) nodules.

The study by Hall et al. was performed in London (UK) and is a sub-study of the LSUT trial.[25] In a MRMC study, two radiographers without prior experience in thoracic CT reporting independently

read all 770 LDCT 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) (1 study)

Singh et al. included a nodule-enriched screening population.[54] Using ClearRead CT from Riverain Technologies in stand-alone and concurrent mode, respectively, 27/150 (18%) chest CT exams could not be processed with the AI algorithm since they had artifacts, thicker sections and/or missing images in the downloaded datasets.

Screening population – AVIEW Lungscreen (Coreline Soft) (3 studies)

The three identified studies by Hwang et al.[48-50] are all 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 South 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 4,666 CT images taken between April 2017 and March 2018 containing 4,990 lung nodules. Semi-automated segmentation failed in 13.4% (669/4,990) of nodules.[49]

A second analysis[48] included 10,424 CT images taken between April 2017 and December 2018 with a total of 10,080 nodules identified. Ninety-one percent of nodules (9,206/10,080) were measured by semi-automated segmentation, while 9% (874/10,080) of nodules failed to be semi-automatically segmented and were measured manually. Segmentation failures occurred in 7.3% (688/9,465) of solid nodules, 19.7% (31/157) of part-solid nodules and 33.8% (155/458) of ground glass nodules.

A third analysis 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 in screening practice and also compared them to retrospective central review of the same CT images by one experienced chest radiologist.[50] 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%) compared to 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) (2 studies)

Kim et al. included 89 patients with sub-solid nodules who underwent preoperative non-enhanced CT and subsequent surgical resection.[61] Veolity version 1.2 (MeVis) was used in concurrent mode by two experienced radiologists. The segmentation success rate of the software in 109 sub-solid nodules was 93.6 % (102/109).

The study by Cohen et al. included 73 patients in whom preoperative CT scans for sub-solid nodules were reconstructed on a single CT system and compared the effects of MBIR and FBP algorithms on software (Veolity, MeVis) semi-automatic measurements.[60] Adequate nodule segmentation was obtained in 66/73 (90.4%) images with FBP and in 68/73 (93.2%) of image 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). Using MBIR, reader 1 performed manual modifications in 21 cases and reader 2 in 39 (median 30). The number of manual modifications was similar between FBP and MBIR ($p = 0.58$).

Mixed population – Veye Chest (Aidence) (3 studies)

The study by 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) (337 chest CT images from 314 subjects).[31] The Veye Chest software from Aidence was able to successfully segment 95% of the total 428 nodules between 3-30 mm.

Martins Jarnalo et al. randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital.[64] The study found that Veye Chest (Aidence) reported an unknown diameter for 3/80 (3.8%) nodules between 4-30 mm.

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.[32] 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 semi-automated volumetry tool (Vitrea Enterprise Solutions, Vital Images Inc.) and once using Veye Chest (Aidence) for automatic diameter and volume measurement. When using the semi-automated volumetry tool, reader 1 and reader 2 deemed 54.6% (35/64) and 44.4% (28/63) of volume measurements as not reliable (and chose to report longest axial diameter instead), whereas with use of Veye Chest, only 2.4% (1/41) and 4.5% (2/44) of volume measurements were deemed as not reliable.

Mixed population – Contextflow SEARCH Lung CT (contextflow) (1 study)

From all patients who had CT images performed on one scanner model at a single hospital in Austria in 2018, Röhrich et al. 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.[29] Two of 216 readings (0.9%) with concurrent software use (Contextflow SEARCH Lung CT) had to be excluded due to “technical difficulties” (no further details reported).

Table 19. 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 (6 studies)				
Hwang 2021a,[49] Korea	K-LUCAS (Korea) 4,666 LDCT taken between April 2017 and March 2018; 4,990 nodules 4,686 (93.9%) solid 78 (1.6%) part-solid 226 (4.5%) Pure ground glass. Non-enhanced CT, slice thickness < 1.5 mm.	**AVIEW Lungscreen (Coreline Soft)**	Failure of semi-automatic segmentation (clinical practice): All nodules	 669/4,990 (13.4%)
Hwang 2021b,[48] Korea	K-LUCAS (Korea) 10,424 LDCT taken between April 2017 and December 2018; 10,080 nodules: 9,465 (93.9%) solid 157 (1.6%) part-solid 458 (4.5%) Pure ground glass. 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%) 688/9,465 (7.3%) 31/157 (19.7%) 155/458 (33.8%)
Hwang 2021c,[50] Korea	K-LUCAS (Korea) 3,353 LDCT taken between April 2017 and December 2017. 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)	 497/3,452 (14.4%) Range 0 to 57.0% (CV 1.28) 1.1% (107/9,389)
Singh 2021,[54] USA	NLST dataset (USA): 150 LDCT first 125 patients with sub-solid nodules; first 25 patients with no nodules. 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 2021,[62] Denmark, Netherlands	NLST dataset (USA): 160 LDCT selected by Lung-RADS category; 40 Lung-RADS 1 or 2; 40 Lung-RADS 3; 40 Lung-RADS 4A;	**Veolity (MeVis)**	Need to manually tune segmentation parameters	28% of nodule segmentations 3/160 (1.9%) nodules (1 reader)

Reference and country	Population / Nodule characteristics / Slice thickness	Technology	Details of technical failure	Failure rate
	40 Lung-RADS 4B Non-enhanced CT, slice thickness: 1.0 to 3.2 mm		Manual diameter measurement deemed necessary: Retrospective MCMR study	2/160 (1.3%) nodules (2 readers) 0/160 nodules (4 readers)
Hall 2022,[25] UK	LSUT study (UK): All 770 LDCT with a lung health check appointment between November 2015 and July 2017; 158 with ≥1 nodule (≥5 mm or ≥80 mm3). 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 (2 studies)				
Cohen 2017,[60] Korea	1 hospital in Seoul (Korea): 73 patients with preoperative CT scans for sub-solid nodules taken between July 2014 to May 2015; 73 sub-solid nodules. Non-enhanced CT, slice thickness 0.625 mm. Reconstructed with FBP and MBIR, respectively.	**Veolity (MeVis)**	Failure of semi-automatic segmentation (MRMC study): Sub-solid nodules - FBP Sub-solid nodules - MBIR Manual modifications of nodule segmentation required (MRMC study): Sub-solid nodules - FBP Sub-solid nodules - MBIR	 7/73 (9.6%) 5/73 (6.8%) 27/73 (37.0%) for reader 1 43/73 (58.9%) for reader 2 (median 35/73, 47.9%). 21/73 (28.8%) for reader 1 39/73 (53.4%) for reader 2 (median 30/73, 41.1%). FBP versus MBIR (p = 0.58).
Kim 2018,[61] Korea	1 hospital in Seoul (Korea): 89 patients with preoperative CT scans for sub—solid nodules taken between November 2014 and July 2016; 109 sub-solid nodules. Non-enhanced CT, slice thickness 0.625 mm.	**Veolity (MeVis)**	Failure of semi-automatic segmentation (MRMC study): Sub-solid nodules	 7/109 (6.4%)

Reference and country	Population / Nodule characteristics / Slice thickness	Technology	Details of technical failure	Failure rate
Mixed population (4 studies)				
Röhrich 2022,[29] Austria	1 hospital in Austria in 2018; first 100 patients with lung pathologies (22 unique, verified diagnoses, but none with lung nodules), first 8 patients without pathological lung findings. Slice thickness: 1 mm.	**Contextflow SEARCH Lung CT (contextflow)**	"Technical difficulties" (not further specified), Retrospective MRMC study	2/216 (0.9%)
Hempel 2022,[32] Netherlands	1 hospital in the Netherlands: 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. 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 2021,[64] Netherlands	1 hospital in the Netherlands: Random 145 chest CT scans performed for various indications between December 2018 and May 2020; 91 nodules: 16 sub-solid nodules, 73 solid nodules, 2 mixture of solid/sub-solid. 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 2022,[31] UK	1 hospital in Edinburgh (UK): 337 scans of 314 current smokers, ex-smokers and/or those with radiological emphysema between 55-74 years taken between January 2008 and December 2009. [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 sub-solid nodules. Slice thickness 1.0-2.5mm.	**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.

3.5.2 Radiologist reading time (10 studies)

Ten studies were identified that reported on the reading time of radiologists with and without software support.[25, 29, 32, 45, 51, 52, 55, 57, 58, 62] Three studies included chest CT images from screening populations,[25, 52, 62] one study included a symptomatic population,[57] and the remaining six studies included mixed indications for the CT scans.[29, 32, 45, 51, 55, 58] The included studies compared the reading time between unaided readers and readers supported by six different technologies: AI-Rad Companion Chest CT (Siemens Healthineers) in stand-alone and concurrent mode, respectively,[45] ClearRead CT (Riverain Technologies) in concurrent[51, 52, 55] and assisted second read mode[51], respectively, contextflow SEARCH Lung CT (Contextflow) in concurrent mode,[29] InferRead CT Lung (Infervision) in concurrent mode,[57, 58] Veolity (MeVis) in concurrent mode[25] and Veye Chest (AIdence) in concurrent mode.[32] Nine of the 10 identified studies reported reduced radiologist reading times by 11.3% to 78% with concurrent software use,[25, 29, 32, 45, 51, 52, 57, 58, 62] whereas one study[55] found similar reading times when using software with vessel suppression function only (**Table 20**). Software assistance as second reader resulted in a significant increase in radiologist reading times by 26% in one study.[51]

Symptomatic population – InferRead CT Lung (Infervision) (1 study)

Kozuka et al. randomly selected 120 chest CT images from cases of suspected lung cancer in patients at a single hospital in Japan.[57] In a MRMC study, two less experienced radiologists independently read the CT images first without software and then (after at least 14 days interval) 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 min, reducing the mean reading time for one case from 3.1 min without software to 2.8 min with software (no level of significance reported).

Screening population – Veolity (MeVis) (2 studies)

The study by Jacobs et al. included 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 lower 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) without software support to 105 seconds (IQR 61–158) 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) in unaided readers to a median of 74 seconds with software support (IQR 46–128) ($p < 0.0001$).

The MRMC study by Hall et al. included all 770 patients who received LDCT for lung cancer screening as part of the LSUT study.[25] 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) 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%) 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) for the pooled radiologists versus 3 minutes (IQR 2-5) for Radiographer 1 and 5 minutes (IQR 4-8) for Radiographer 2 ($p < 0.001$ for both comparisons).

Screening population – ClearRead CT (Riverain Technologies) (1 study)

The MRMC study by Lo et al. included a nodule-enriched screening population.[52] 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 dataset.

Mixed population – AI-RAD Companion Chest CT (Siemens Healthineers) (1 study)

Abadia et al. 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.[45] 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) (1 study)

In the study by Liu et al., chest CT scans (screening and in-patient) performed at multiple hospitals in China were retrospectively collected with convenience sampling.[58] The total dataset comprised 12,574 CT scans, of which 1,129 CT scans 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, 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) (2 studies)

The study by Hsu et al. retrospectively included 150 consecutive cases with lung nodules ≤1cm or no nodule on chest CT performed at a single hospital in Taiwan.[51] 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 month of chest CT experience) and a 'Senior group' (three experienced chest radiologists with 5, 10 and 25 years of experience, respectively). In assisted 2nd-read mode, readers read the CT images without software first 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 assisted 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 for the concurrent-read mode compared to 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 assisted 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 assisted 2nd-read mode and 107 seconds for concurrent mode ($p < 0.001$ for all).

Takaishi et al. included 61 thoracic or thoracic-abdominal unenhanced CT images conducted at a single hospital in Japan for various reasons.[55] 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 versus 31.2 seconds) in Reader A. The average reading time of all three radiologists was slightly longer with software use (29.2 seconds versus 32.3 seconds, +9.5%, $p = 0.11$).

Mixed population – contextflow SEARCH Lung CT (contextflow) (1 study)

From all patients who had CT images performed on one scanner model at a single hospital in Austria in 2018, Röhrich et al. included the first 100 patients with lung pathologies as well as the first eight patients without pathological lung findings.[29] 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 (2 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 compared to where they did not (110 vs 39 s 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 (1 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.[32] 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[11] (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.

Table 20. 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 (1 study)						
Kozuka 2020,[57] Japan	120 chest CT images from cases of suspected lung cancer in patients, 1 hospital in Japan	**InferRead CT Lung (Infervision)**	MRMC, 2 less experienced radiologists, Concurrent mode	MRMC, Same as 'Index test', Unaided	To detect any nodules ≥3 mm	Concurrent mode: (↓) (-11.3%)
Screening population (3 studies)						
Lo 2018,[52] USA	324 LDCT from the NLST dataset and 2 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 2021,[62] Denmark, Netherlands	NLST dataset (USA): 160 CT images (40 per Lung-RADS category)	**Veolity (MeVis)**	MRMC, 3 experienced chest radiologists and 4 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 2022,[25] UK	All 770 LDCT from Lung Screen Uptake Trial (LSUT), London (UK)	**Veolity (MeVis)**	MRMC, 2 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%)

Reference and country	Population	Technology	Index test	Comparator test	Reader task	Effect of software use on reading time compared to unaided reading
Mixed population (6 studies)						
Abadia 2021,[45] 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, 1 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 5 largest nodules	Concurrent mode: ↓ (-78%)
Hsu 2021,[51] Taiwan	150 consecutive cases with lung nodules ≤1cm 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, 'Junior group': 6 radiology residents, >6 month of chest CT experience; 'Senior group': 6 experienced chest radiologists, 5, 10 and 25 years of experience, respectively; Concurrent mode; 2nd read mode	MRMC, Same as 'Index test', Unaided	To detect any nodule (3-10 mm)	Concurrent mode: ↓ for all readers (-21%) ↓ for radiology residents (-23%) ↓ for experienced chest radiologists (-16%). Assisted 2nd-read mode: ↑ for all readers (+26%) ↑ for radiology residents (+28%) ↑ for experienced chest radiologists (+24%).
Takaishi 2021,[55] 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, 6 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: = All readers (+9.5%) = Reader A ↑ Reader B ↓ Reader C
Röhrich 2022,[29] Austria	108 CT images from 1 hospital in Austria;	**Contextflow SEARCH Lung CT (contextflow)**	MRMC, 6 radiology residents (mean 2.1 ± 0.7 years of experience),	MRMC, Same readers as 'Index test' but each image only	To interpret the CT images (diagnosis of lung pathologies)	Concurrent use: ↓ (-31.3%) (↓) Radiology residents (↓) Attending general radiologists

Reference and country	Population	Technology	Index test	Comparator test	Reader task	Effect of software use on reading time compared to unaided reading
	first 100 patients with lung pathologies (no lung nodules), first 8 patients without pathological lung findings.		4 attending general radiologists (mean 12 ± 1.8 years of experience), Each image read by 1 radiology resident and 1 attending general radiologist, Concurrent mode	read once by each reader (with or without software), Unaided		
Liu 2019,[58] China	123 (Batch 1) and 148 (Batch 2) chest CT images (screening and in=-patient) from >10 hospitals in China.	**InferRead CT Lung (Infervision)**	MRMC, 2 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%)
Hempel 2022,[32] 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 1 Dutch hospital.	**Veye Chest (Aidence)**	MRMC, 1 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%). 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 study, Multi-reader multi-case study; NR, Not reported.

↑ 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.

3.5.3 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.

3.5.4 Acceptability and experience of using the software (3 studies)

Three studies were identified that assessed readers' acceptability and experience of using AI-based software for the detection and analysis of lung nodules.[25, 45, 64] One study was performed in a screening population[25] and the other two in mixed populations.[45, 64]

Screening population – Veolity (MeVis) (1 study)

This sub-study of the LSUT trial performed in London (UK) included all 770 patients who received LDCT for lung cancer screening.[25] 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) 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) (1 study)

Abadia et al. 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.[45] 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. 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 radiologist reported increased confidence for lung nodule detection for all 20 cases (100%).

Mixed population – Veye Chest (Aidence) (1 study)

Martins Jarnalo et al. randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital.[64] The authors reported in the discussion that the single system threshold setting for nodule detection of the Veye Chest software from

Aidence for various different uses (e.g., follow-up versus screening) has been found to be a limitation, and that it would be a useful improvement to be able to set different thresholds.

3.5.5 Other non-prespecified outcomes

One study 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.[30] This outcome was not pre-specified in the protocol; result are reported in **Appendix 5, section 13.5.7.1.**

3.5.6 Sub-questions 1 to 6.

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

3.6 Impact on patient management

3.6.1 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.

3.6.1.1 All detected nodules (true positive and false positive) (6 studies)

Six studies were identified that reported on the characteristics of all detected nodules (true positives and false positives).[32, 45, 48-50, 64] Three studies were performed in consecutive screening populations,[48-50] and the remaining three studies included mixed populations.[32, 45, 64] Only one MRMC study[32] 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 compared to unaided reading (87.1% vs 90.6%, no level of significance reported).[32] A second study[49] used an unpaired design and reported nodule characteristics before and after software implementation in prospective screening practice. In contrast, this study observed significantly higher ($p < 0.001$)

number of nodules detected per participants 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.

a) Comparative results – Reader with and without software (2 studies)

Mixed population – Veye Chest (Aidence) (1 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.[32] For this MRMC study, two experienced radiologists independently assessed the CT images to determine nodule management recommendation grade based on the 2015 BTS guidelines[11] twice, first unaided and then aided by Veye Chest software (Aidence). The readers were tasked to report relevant pulmonary nodules that contributed to their management decision. A summary of the nodule types and sizes is reported in **Table 21**. 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 being solid was lower (87.1%) than without software use (90.6%) (no level of significance reported).

Table 21. Nodule number, type and size in patients with incidentally detected nodules on CT, with and without concurrent use of Veye Chest[32]

	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.

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

In a before-after study, Hwang et al. included 6,487 consecutive participants of the K-LUCAS project: 1,821 participants were screened before the implementation of the AVIEW Lungscreen software and 4,666 participants received screening after the implementation of the software.[49] The study observed a significantly higher number of detected nodules per participant (0.76 vs. 1.07, $p < 0.001$) and higher proportion of solid nodules (90.2% vs. 93.9%, $p < 0.001$) in participants screened after software implementation (**Table 22**). No significant difference in nodule size was observed when nodules were measured on transverse planes after software implementation ($p = 0.441$), but sizes of nodules were significantly greater when nodules 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 though as the study did not use a fully paired design, but different CT images were analysed by different readers before and after software implementation.

Table 22. Characteristics of detected nodules (true and false positives) in consecutive screening populations from Korea (3 studies)

Reference and country	Technology / Reading details for detection	# nodules or participants	Nodule type	Nodule size	Lung-RADS category
Hwang 2021,[49] 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 2 – 84.8% 4B – 1.2% 3 – 9.1% 4X – 1.7% 4A – 3.2%
	AVIEW Lungscreen (Coreline Soft) Assisted 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% 4B – 1.1% 3 – 6.5% 4X – 0.7% 4A – 2.5%
	Unaided reader	1,821 participants	NR	NR	Per-participant: Transverse plane 1 – 58.6% 4A – 2.3% 2 – 31.5% 4B – 0.7% 3 – 5.3% 4X – 1.5%
	AVIEW Lungscreen (Coreline Soft) Assisted 2nd-read mode	4,660 participants	NR	NR	1 – 51.5% 4A – 2.9% 2 – 37.6% 4B – 1.2% 3 – 6.1% 4X – 0.8%
Hwang 2021,[48]Korea	AVIEW Lungscreen (Coreline Soft) Assisted 2nd-read mode	10,080 nodules	Solid 93.9% Part-solid 1.6% Pure GGN 4.5%	Average transverse diameter Solid: Median 3.6 mm <5 mm: 75.1% 5-6 mm: 8.1% 6-8 mm: 6.0% ≥8 mm: 4.6% Part-solid: Median 11.9 mm <5 mm: 0.008% ≥5 mm: 1.5% Pure GGN: Median 5.8 mm <5 mm: 1.7% ≥5 mm: 2.8%	NR

Reference and country	Technology / Reading details for detection	# nodules or participants	Nodule type	Nodule size	Lung-RADS category	
	AVIEW Lungscreen (Coreline Soft) Assisted 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% Solid ≥10 mm: 4.7% Part-solid: 2.7% Pure GGN: 3.4%		NR	
	AVIEW Lungscreen (Coreline Soft) Assisted 2nd-read mode	10,424 participants	NR	NR	1 – 53.0% 2 – 26.9% 3 – 11.7%	4A – 4.3% 4B/X – 4.1%
Hwang 2021,[50] Korea	AVIEW Lungscreen (Coreline Soft) Assisted 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) Assisted 2nd-read mode (original institutional reading)	3,353 participants	NR	NR	1 – 53.0% 2 – 26.9% 3 – 11.7%	4A – 4.3% 4B/X – 4.1%

GGN, ground glass nodule; NR, Not reported.

b) Comparative results – Stand-alone AI versus unaided reader (1 study)

One study reported the size of nodules detected by stand-alone AI as well as by an expert unaided reader in a mixed population.[45]

Mixed population – AI-RAD Companion Chest CT (Siemens Healthineers) (1 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.[45] The nodule 2-D 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 23**.

Table 23. Nodule 2-D axial diameter in all detected nodules in patients with complex lung disease[45]

Lung condition	Stand-alone software		Unaided expert chest radiologist	
	No. of nodules detected	Nodule size (mm) Median (IQR)	No. 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.

Three studies reported characteristics of nodules detected by software-assisted readers[48, 50] and stand-alone software,[64] respectively, without comparator. These non-comparative results are reported in **Appendix 5, section 13.5.8.1.**

3.6.1.2 True positive nodules (7 studies)

Seven studies reported characteristics of correctly detected nodules.[30, 49, 54, 57-59, 64] Four studies were performed in screening populations,[30, 49, 54, 59] in one study, the indication for the chest CT scan was lung cancer suspicion,[57] and in the remaining two studies, the indication for the chest CT scan was mixed.[58, 64] Of these, 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).[57, 59] Additional true positive nodules detected with software use were 56-57% solid, due to larger improvements in the detection of sub-solid 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 percent of additional true positive nodules were 6 mm or larger.[57]

a) Comparative results – Reader with and without software (2 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).[57, 59]

Symptomatic population – InferRead CT Lung (Infervision) (1 study)

Kozuka et al. randomly selected 120 chest CT images from cases of suspected lung cancer at a single hospital in Japan.[57] 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 24**. An additional 254 true positive nodules were identified by the two readers with software use. The additional nodules had the following composition: 57% solid, 14% part-solid, 15% ground glass and 14% calcified. Seventy-eight percent of additional nodules had a diameter of 3-6 mm, and 22% were 6 mm or larger.

Table 24. Nodule type and size in a random symptomatic population from Japan[57]

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) GGNs 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) GGNs 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) GGNs 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) GGNs 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) GGNs 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.

Screening population – InferRead CT Lung (Infervision) (1 study)

Zhang et al. included 860 consecutive patients who had undergone lung cancer screening at one Chinese hospital as part of the NELCIN-B3 project.[59] 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 25**. 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 GGNs. Of 126 additional participants with solid or part-solid nodules, 67% had a nodule diameter of 5 mm or

smaller, and 33% had nodules that were 6 mm or larger. The 82 additional participants with pure GGNs had nodules with a diameter less than 20 mm.

Table 25. Nodule characteristics of subjects with at least 1 nodule in a consecutive screening population from China, by mode of detection[59]

Nodule type	Nodule diameter category	Subjects 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)*	+75%	50.3% (188/374)
	6-7 mm	9.9% (16/162)	11.1% (41/370)*	+156%	11.2% (42/374)
	8-14 mm	4.9% (8/162)	5.1% (19/370)*	+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)***	**+89%**	**66.8% (250/374)**
Part-solid	≤5 mm	1.9% (3/162)	2.1% (8/370)*	+167%	2.1% (8/374)
	≥6 mm	0	1.4% (5/370)*	NA	1.3% (5/374)
	All	**1.9% (3/162)**	**3.5% (13/370)***	**+333%**	**3.5% (13/374)**
Ground glass	≤19 mm	17.3% (28/162)	29.7% (110/370)*	+293%	29.7% (111/374)
	≥20 mm	0	0	NA	0
	All	**17.3% (28/162)**	**29.7% (110/370)***	**+293%**	**29.7% (111/374)**

* Indicates significant difference (p<0.001) by the Chi-square test between unaided and AI-assisted reading.

b) Comparative results – Stand-alone software versus unaided reader (1 study)

One study[58] reported the proportions of detected nodules by size and type for unaided radiologists, stand-alone software as well as consensus expert reading.

Mixed population – InferRead Lung CT (Infervision) (1 study)

Liu et al. included a test set consisting of 1,129 CT images (screening and in-patients) from more than 10 hospitals in China using convenience sampling.[58] The chest CT images were retrospectively assessed by stand-alone software (InferRead Lung CT) as well by two experienced radiologists without software use. **Table 26** reports the proportions of detected nodules by size and type for unaided radiologists, stand-alone software as well as consensus expert reading (2 experienced radiologists, reference standard).

Table 26. Characteristics of correctly detected nodules in a mixed population from China obtained via convenience sampling.[58]

Nodule type	Nodule size	Reference standard	Correctly detected nodules		
			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%
Sub-solid	**≤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.

Superseded-

Four studies reported on characteristics of true positive nodules detected by stand-alone software,[49, 64] by software-assisted readers,[54] and/or by the reference standard[30, 54, 64] without a comparator. These non-comparative results are reported in **Appendix 5, section 13.5.8.2.**

see erratum

3.6.1.3 Additional true positive nodules detected by software compared to unaided reading (1 study)

Incidental population – AI-RAD Companion Chest (Coreline Soft) (1 study)

The study by Rueckel et al. included 105 consecutive patients who received a whole-body CT scan in the emergency department of a single German hospital.[47] Retrospective reading by stand-alone software (AI-RAD Companion Chest CT prototype, Siemens Healthineers) detected three additional true positive nodules compared to 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.

3.6.1.4 False positive nodules (4 studies)

Four studies reported on characteristics of false positive nodules detected by stand-alone software in a random screening population,[46] an incidental population[47] and mixed populations,[45, 64] respectively. No study compared characteristics of false positive nodules between readers with and without concurrent software use.

a) Non-comparative results (4 studies)

Incidental population – AI-RAD Companion Chest (Siemens Healthineers) (1 study)

The study by Rueckel et al. included 105 consecutive patients who received a whole-body CT scan in the emergency department of a single German hospital.[47] Nodules were detected retrospectively by stand-alone software (AI-RAD Companion Chest CT prototype, Siemens Healthineers) and compared to 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). Of 81 additional nodules detected by the stand-alone software, three were true positive nodules. The remaining 78 false positive nodules were classed as trauma-associated (27%), scarred/post-inflammatory (38%), perifissural lymph nodes (6%), granuloma (6%) or could not be confirmed visually (22%).

Screening population – AI-RAD Companion CT Chest (Siemens Healthineers) (1 study)

Chamberlin et al. 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.[46] 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) (1 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.[45] 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 to 17.2). If the nodule was near a blood vessel, overestimation of nodule size was occasionally observed. A few false positives were also caused by incorrect lung segmentation.

Mixed population – Veye Chest (Aidence) (1 study)

Martins Jarnalo et al. randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital.[64] 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 sub-solid (**Table 27**). 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, e.g., a gland, bronchiectasis, or a large consolidation.

Table 27. Characteristics of all detected nodules, true positive, false positive and false negative nodules – Stand-alone software in a random mixed population[64]

Reference and country	Detection details	Number of nodules	Nodule type	Nodule size (mm) (mean ± SD)
Martins Jarnalo 2020,[64] Netherlands	Reference standard	93 nodules	Solid 80% Sub-solid 18% Mixed solid/sub-solid 2%	7.0 ± 4.1
	Veye Chest (Aidence) Stand-alone	130 detected nodules (TP and FP)	Solid 85% Sub-solid 14% Mixed solid/sub-solid 1%	9.0 ± 7.1
		80 true positive nodules	Solid 81% Sub-solid 16% Mixed solid/sub-solid 3%	7.0 ± 3.8
		50 false positive nodules	Solid 90% Sub-solid 10% Mixed solid/sub-solid 0%	11.8 ± 10.0
		11 false negative nodules	Solid, 4 mm: n=5 Solid, calcified, 4 mm: n=3 Sub-solid, 4 mm: n=1 Sub-solid, 18 mm: n=1 Sub-solid, 20 mm: n=1	6.7 ± 6.1

FP, False positive; SD, Standard deviation; TP, True positive.

3.6.1.5 False negative (missed) nodules (9 studies)

Nine studies reported characteristics like nodule size and type of missed nodules: four studies were performed in screening populations,[25, 49, 54, 59] one study was performed in a symptomatic population,[57] and the other four studies were performed in populations with mixed indication for the chest CT scan.[45, 56, 58, 64] Of these, two studies compared the characteristics of missed nodules in readers assessing the same CT images with and without concurrent software use (InferRead CT Lung, Infervision).[57, 59] 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 sub-solid nodules than nodules missed by unaided readers.

a) Comparative results – Reader with and without software (2 studies)

Symptomatic population – InferRead CT Lung (Infervision) (1 study)

Kozuka et al. randomly selected 120 chest CT images from cases of suspected lung cancer at a single hospital in Japan.[57] 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 24**. With software use, the two readers missed less nodules (922 vs 1,176; -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 to unaided reading.

Screening population – InferRead CT Lung (Infervision) (1 study)

Zhang et al. included 860 consecutive patients who had undergone lung cancer screening at one Chinese hospital as part of the NELCIN-B3 project.[59] In the real-world radiologist observation, one of 14 residents drafted the diagnostic report, and one of 15 board-certified radiologist 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 GGNs (**Table 28**). Missed nodules were solid and larger than 5 mm in 17.5%, part-solid and larger than 5 mm in 2.4% and ground glass nodules smaller than 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 solid nodules ≤5 mm, one (25%) had solid nodules larger than 5 mm and the remaining participant had a ground glass nodules smaller than 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 GGNs (-98.8%) compared to solid nodules (-97.5%).

Table 28. Characteristics of missed nodules in a consecutive screening population from China[59]

Nodule type	Nodule diameter category	Missed subjects with ≥1 nodule		
		Unaided (clinical practice)	AI-assisted (MRMC study)	Difference in numbers missed, n (%)
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%)**
Ground glass	≤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%)**

b) Comparative results – Stand-alone AI versus unaided reader (2 studies)

Mixed population – AI-RAD Companion Chest CT (Siemens Healthineers) (1 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.[45] 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 to 14.4), whereas the unaided expert chest radiologist missed 8.4% (37/441) of nodules with a median size of 6.1 mm (IQR 5.1 to 9.2). 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) (1 study)

Liu et al. included a test set consisting of 1,129 CT images (screening and in-patients) from more than 10 hospitals in China using convenience sampling.[58] The chest CT images were retrospectively assessed by stand-alone software (InferRead Lung CT) as well by two experienced radiologists without software use. **Table 29** reports the proportions of missed nodules by size and type for stand-alone software as well as for the unaided radiologists.

Table 29. Characteristics of missed nodules in a mixed population from China obtained via convenience sampling.[58]

Nodule type	Nodule size	Missed nodules		
		Stand-alone software	Reader 1 – unaided	Reader 2 – unaided
Total	**All**	1,879	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%
Sub-solid	**≤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%

Non-comparative results (5 studies) are reported in **Appendix 5, section 13.5.8.3.**

3.6.2 Proportion of detected nodules that are malignant (3 studies)

Three studies performed in consecutive screening populations reported on the proportion of detected nodules that were diagnosed as lung cancer.[25, 48, 49] 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.[25, 49]

a) Comparative results – Reader with and without software (2 studies)

Screening population – Veolity (MeVis) (1 study)

The study by Hall et al. was performed in London (UK) and is a sub-study of the LSUT trial.[25] It included 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 to the number 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 multi-disciplinary team (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; for Radiographer 2, the proportion of detected nodules being malignant was 19.4% (30/155).

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

In a before-after study, Hwang et al. included 6,487 consecutive participants of the K-LUCAS project: 1,821 participants were screened before the implementation of the AVIEW Lungscreen software and 4,666 participants received screening after the implementation of the software.[48, 49] 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 or higher, and cases of lung cancer were identified by medical record review. The proportion of all detected nodules (Lung-RADS category 2 or higher) that were later diagnosed as lung cancer was 1.2% (16/1,391) before the implementation of the software and 0.6% (31/4,990) after the implementation of the AVIEW Lungscreen software. Of the screen-positive (actionable) nodules (Lung-RADS category 3 or

higher), 6.6% (14/212) and 5.2% (28/538) were malignant before and after implementation of the software, respectively.

b) Non-comparative results (1 study)

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

The other study by Hwang et al. included 10,424 concurrent baseline LDCT scans obtained after the implementation of the AVIEW Lungscreen software as part of the Korean K-LUCAS project.[48] 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 **Table 30**. In all 4,642 risk-dominant nodules, 1.1% (52/4,642) were diagnosed as lung cancer within one year after LDCT, and 1.2% (58/4,642) were diagnosed with any lung cancer after LDCT. The highest proportion of malignant nodules was found in solid nodules ≥10 mm (14%) and in part-solid nodules (13%).

Table 30. Proportion of detected risk-dominant nodules that are malignant, by nodule type and size, in a consecutive screening population from Korea[48]

	Solid nodules						**Part-solid nodules**	**Non-solid nodules**	**Total**
	<6 mm	6-7 mm	7-8 mm	8-9 mm	9-10 mm	≥10 mm			
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.

3.6.3 Impact of test result on clinical decision-making (6 studies)

Six comparative studies were identified that reported the impact of software use on clinical decision making.[25, 53, 54, 61, 62, 65] Four studies were performed in screening populations,[25, 54, 62, 65] one study was performed in a surveillance population with applicability concerns,[61] and in the remaining study, the indication for the chest CT scan was not reported.[53] Four studies consistently reported that with software use, readers tended to upstage Lungs-RADS[62, 65] or Fleischner risk categories[32, 61] rather than downstage.

a) Comparative results – Reader with and without software (6 studies)

Screening population – MeVis (2 studies)

The study by Jacobs et al. included a nodule-enriched screening population.[62] One-hundred and sixty LDCT images were selected from the US-based NLST dataset stratified by Lung-RADS category (40 Lung-RADS 1 or 2; 40 Lung-RADS 3; 40 Lung-RADS 4A; 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 nodule type and size with and without concurrent use of the software Veolity (MeVis) (**Table 31**).

Table 31. Lung-RADS category with and without concurrent software use in a nodule-enriched screening population[62])

Lung-RADS category	7 readers with concurrent software use (160 LDCT scans each)	7 readers without concurrent software use (160 LDCT 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 using the dedicated CT lung screening viewer with software support, 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. was performed in London (UK) and is a sub-study of the LSUT trial.[25] It included 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) and common incidental findings and had to make patient management recommendations. The study reports on the concordance of management decisions against BTS guidelines[11] 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 (1 cancer) and for 40.5% (53/131) a less active follow-up was recommended (3 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 (4 cancers) and for 16.0% (24/150) a less active follow-up was recommended (1 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 (3 cancers) and for 12.9% (20/155) a less active follow-up was recommended (1 cancer).

Screening population - VUNO Med-Lung CT AI (VUNO) (1 study)

Park et al. included a nodule- and cancer-enriched screening population (200 baseline LDCT), selected from the US-based NLST dataset.[65] 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 32**).

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%) comp 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 or higher) occurred in 6/200 (3%) to 26/200 (13%) of CT images that were assessed with software

use. Between 0/200 to 18/200 (9%) of CT images were down-staged by the five readers with software use compared to unaided reading.

Table 32. Lung-RADS category based on stand-alone software and readers with and without concurrent software use in a nodule-enriched screening population[65]

Lung-RADS category	Stand-alone software (200 LDCT)	5 readers with concurrent software use (200 LDCT scans each)	5 readers without concurrent software use (200 LDCT 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 dataset, substantial management discrepancies between the 310 reader pairs (Lung-RADS category 1/2 vs. 4A/B) were reduced in half by application of the software (32/310 to 16/310).

Screening population – ClearRead CT (Riverain Technologies) (1 study)

Singh et al. included 150 patients who underwent LDCT of the chest as part of the NLST - the first 125 patients with sub-solid nodules (154 part-solid or 156 ground glass nodules between 6 and 30 mm) and the first 25 patients with no nodules detected.[54] 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 changed for these five nodules from Lung-RADS 2 to Lung-RADS 4A which would impact the management of these patients.

Surveillance population with applicability concerns – Veolity (MeVis) (1 study)

Kim et al. included 89 patients with sub-solid nodules who underwent preoperative non-enhanced CT and subsequent surgical resection at the Seoul National University Hospital.[61] In a MRMC study, nodule classification based on diameter measurements of 102 sub-solid nodules obtained by two experienced radiologists were compared with and without concurrent use of Veolity (MeVis). The sub-solid nodules were categorised according to Fleischner's guidelines[68] into (1) pure ground glass, (2) part-solid with a solid component ≥5 mm or (3) part-solid with a solid component less than 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 semi-automatic nodule measurement, Reader 1 and Reader 2 both classed more part-solid nodules as having a solid portion larger than 5 mm compared to manual measurement (59.8% versus 43.1% for Reader 1; 58.8% versus 55.9% for Reader 2-1; 61.8% versus 53.9% for Reader 2-2, see **Table 33**) which would suggest that with software use, more people would be receiving invasive procedures and less people would receive CT surveillance.

Table 33. Sub-solid nodule classification of the two readers with and without software use in patients with previously detected nodules[61]

		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%)

Unclear indication for CT scan – ClearRead CT (Riverain Technologies) (1 study)

This MRMC study by Milanese et al. included 93 consecutive patients referred to the 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).[53] One radiologist with three years of experience in chest CT and a radiology resident independently performed semi-automatic 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.[66] With vessel suppression, Reader 1 changed the nodule category from 100-250 mm^3 to <100 mm^3 for two nodules, whereas Reader 2 changed the nodule category for two nodules from the 100-250 mm^3 category to <100 mm^3 and >250 mm^3, respectively (**Table 34**).

Table 34. Risk classification based on semi-automatic volume measurement using standard CT images and vessel-suppressed CT images in consecutive LDCT with unclear indication[53]

		Reader 1 (65 solid nodules)	Reader 2 (65 solid nodules)	Total (130 solid nodules)
Standard CT images	<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 images	<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%)

3.6.4 Number of people having CT surveillance (5 studies)

Five studies reported on the number of people that were referred for CT surveillance ('intermediate nodules'),[25] people followed up as nodules suspected to be benign,[57] number of people that were classed as Lungs-RADS categories 3 or 4A[49, 50, 62] or 'intermediate' according to the NELSON criteria.[50] Four studies were performed in consecutive[25, 49, 50] or nodule-enriched screening populations,[62] and one study was performed in a random symptomatic population.[57] Of these, a MRMC study[62] and a before-after study[49] 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.

a) Comparative results – Reader with and without software (2 studies)

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

In a before-after study, Hwang et al. included 6,487 consecutive participants of the K-LUCAS project: 1,821 participants were screened before the implementation of the AVIEW Lungscreen software and 4,666 participants received screening after the implementation of the software.[49] Before software implementation, unaided single expert chest radiologists manually measured the transverse plane of the risk-dominant nodules and classed 7.6% (139/1,821) participants as Lung-RADS categories 3 or 4A. After software implementation, single expert chest radiologists classed 9.0% (418/4,666) as

Lung-RADS categories 3 or 4A based on transverse planes. Of 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) (1 study)

The study by Jacobs et al. included a nodule-enriched screening population.[62] One hundred and sixty LDCT images were selected from the US-based NLST dataset stratified by Lung-RADS category (40 Lung-RADS 1 or 2; 40 Lung-RADS 3; 40 Lung-RADS 4A; 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 nodule type and size with and without concurrent use of the software Veolity (MeVis). Without software use, the seven readers classed 32.6% (365/1,120) as Lungs-RADS categories 3 or 4A. In contrast, 43.2% (484/1,120) were classed as Lung-RADS categories 3 or 4A with concurrent software use.

Non-comparative results (3 studies) are reported in **Appendix 5, section 13.5.8.4.**

3.6.5 Number of CT scans taken as part of CT surveillance (No study)

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

3.6.6 Number of people having a biopsy or excision (5 studies)

Five studies reported on the number of people that were directly referred to MDT because of 'suspicious nodules',[25] of people with lung cancer diagnosed or followed up as nodules suspected of lung cancer,[57] the number of people that were positive on the narrow definition using Lungs-RADS (i.e. category 4B or 4X by Lung-RADS)[49, 50, 62] or 'positive' according to NELSON criteria.[50] Four studies were performed in consecutive[25, 49, 50] or nodule-enriched screening populations,[62] and one study was performed in a random symptomatic population.[57] Of these, a MRMC study[62] and a before-after study[49] 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.

a) Comparative results – Reader with and without software (2 studies)

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

In a before-after study, Hwang et al. included 6,487 consecutive participants of the K-LUCAS project.[49] Before software implementation, unaided single expert chest radiologists manually measured the transverse plane of the risk-dominant nodules and classed 2.3% (41/1,821) participants as Lung-RADS categories 4B or 4X. After software implementation, a single expert chest radiologist classed 2.0% (93/4,666) as Lung-RADS categories 4B or 4X based on transverse planes. Of 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) (1 study)

The study by Jacobs et al. included a nodule-enriched screening population.[62] One hundred and sixty LDCT images were selected from the US-based NLST dataset based on Lung-RADS category (40 Lung-RADS 1 or 2; 40 Lung-RADS 3; 40 Lung-RADS 4A; 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% (234/1,120) as Lungs-RADS category 4B. With concurrent software use, the seven readers classed 23% (259/1,120) CT images as Lung-RADS categories 4B.

Non-comparative results (3 studies) are reported in **Appendix 5, section 13.5.8.5.**

3.6.7 Stage of cancer at detection (No study)

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

3.6.8 Time to diagnosis (1 study)

One study 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 dataset.[65] 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 dataset, 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 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.

3.6.8.1 Other non-prespecified outcomes

Other outcomes not pre-specified in the protocol are reported in **Appendix 5, 13.5.8.6.** 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 or higher) of LDCT images taken and assessed in screening practice with and without the use of the AVIEW Lungscreen software (Coreline Soft).[48-50]

3.6.8.2 Sub-questions 1 to 6.

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

3.7 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 **Appendix 2 Table 67.**

4 SYSTEMATIC REVIEW OF CLINICAL EFFECTIVENESS (KEY QUESTION 2) – METHODS AND RESULTS

4.1 Methods

4.1.1 Identification and selection of studies

4.1.1.1 Search strategy

The same search strategy as described in the methods for test accuracy was used (see **section 2.1** Identification and selection of studies).

4.1.1.2 Study eligibility criteria

The study eligibility criteria were as follows:

Population	See 2.1.2
Target condition	Lung cancer
Intervention	See 2.1.2
Comparator	CT scan review by a radiologist or another healthcare professional without software for automated detection and analysis of lung nodules (using diameter or volume to measure nodule size). Where data permits, 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' acceptability of use of the software.
Study design	• Randomised controlled trials; • Quasi-randomised trials; • Cohort studies (retrospective/prospective); • Before-after studies; • Historical controlled studies. • Qualitative studies (for patient acceptability of the 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 **section 2.1.2** were used.

4.1.1.3 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 that are excluded at full text stage were documented, including the reasons for their exclusion (see **Appendix 2, Table 64** and **Table 65).**

4.2 Results

No studies on intermediate outcomes (e.g. potential benefits by earlier nodule detection and shorter time to diagnosis; potential harms of increased surveillance for patients with benign nodules) and final health outcomes were identified (see **Figure 6**). 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 analytical model and simulations using evidence from the systematic review of test accuracy evidence and additional types of evidence collected as described in **section 7**. **Figure 12** below illustrates the linked evidence approach.

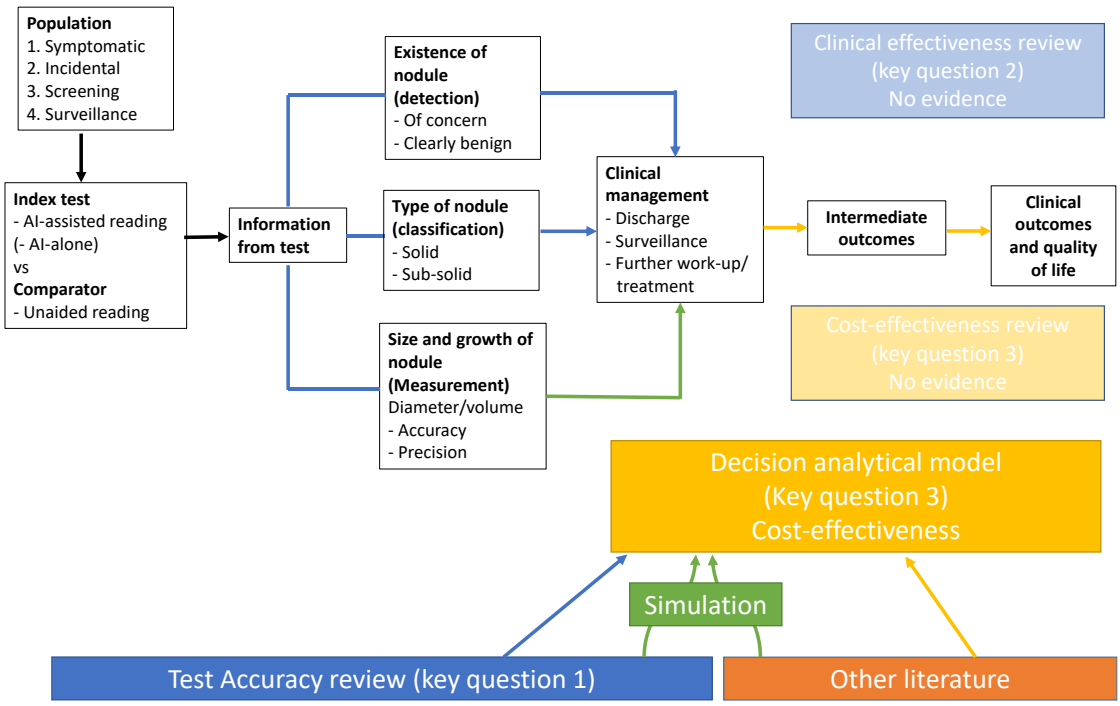


Figure 12 An illustration of linked evidence approach adopted for this diagnostic assessment

5 SYSTEMATIC REVIEW OF COST-EFFECTIVENESS (KEY QUESTION 3) – METHODS AND RESULTS

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,[69] especially with using artificial intelligence software. Algorithms designed for nodule assessment and management use information to predict malignancy and may influence screening outcomes.[69]

In this systematic review, we aimed to review all economic analyses that assessed the cost-effectiveness of using software for the automated detection and analysis of lung nodules from CT images compared with unassisted CT image analysis in people undergoing CT scans that included the chest due to symptoms suggestive of lung cancer, for purposes unrelated to suspicion of lung cancer, for surveillance of previously identified nodules or for lung cancer screening.

5.1 Methods for systematic review of cost-effectiveness

5.1.1 Identification and selection of studies

5.1.1.1 Search strategy

The searches carried out for the systematic review of test accuracy and clinical effectiveness (see section 2.1.1) 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.

As there were likely to be few, if any, economic evaluations of cost-effectiveness studies of the use of AI-based software for nodule detection and analysis in this specific population and context, broader searches for lung nodules/cancer imaging or screening (without AI terms, and not specifically CT) were undertaken to identify information on model structures, costs and utility values to inform the economic model. Where appropriate, search filters for economic evaluations and/or cost or HRQoL studies were applied.

Sources included:

MEDLINE All (Ovid);

Embase (Ovid);

National Health Service Economic Evaluation Database (NHS EED) (CRD);

Health Technology Assessment (HTA) database (CRD);

International HTA database (INAHTA);

Cost-Effectiveness Analysis (CEA) 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);

reference lists of selected highly relevant papers. Full search strategies can be found in **12.5 Appendix 6: Literature search strategies: searches to inform the economic model**.

5.1.1.2 Study eligibility criteria

Studies that satisfy the following criteria were included:

Population	See 2.1.2
Target condition	Lung cancer
Intervention	See 2.1.2
Comparator	CT scan review by a radiologist or another healthcare professional without software for automated detection and analysis of lung nodules (using diameter or volume to measure nodule size). Where data permitted, the following subgroups were considered: - 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, incremental cost effectiveness ratio, quality adjusted life years)
Study design	Full economic evaluations (including cost-effectiveness analysis, cost-utility analysis and cost-benefit analysis). Cost minimisation analysis, cost-consequence/outcome

	description, costs analysis (UK only) and cost description (UK only) may also be included if full economic evaluations are lacking.
Publication type	Peer reviewed papers. Abstracts and manufacturer data will be included, but only outcome data that have not been reported in peer-reviewed full-text papers will be extracted and reported.
Language	English

Exclusion criteria are the same as described in clinical effectiveness review section (see **2.1.2**).

5.1.1.3 Study screening and selection

All records retrieved were screened independently by two reviewers (PA/HG) at title/abstract stage, of which potentially relevant records were further examined at full-text. 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.

5.1.2 Extraction and study quality

5.1.2.1 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 and year of study), baseline characteristics (population, intervention, comparator and outcomes), methods (study perspective, time horizon, discount rate, measure of effectiveness current, 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 if an agreement could not be reached.

5.1.2.2 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.[70] Any studies using an economic model

were further assessed against the framework for the quality assessment of decision analytic modelling developed by Philips and colleagues.[71]

5.1.3 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.

5.2 Results for systematic review of cost-effectiveness

5.2.1 Results of literature search

The literature search identified 1,988 records through electronic database searches and through other sources. After removing duplicates, 1,299 studies were screened for inclusion based on title and abstract. Fifteen studies were considered potentially relevant and were reviewed at full text. All studies were excluded at the full-text stage, and the reasons for exclusion are shown in **Figure 13.**

Two potentially relevant economic analyses (Bajre et al., 2017;[72] Adams et al., 2021[73]) did not meet our inclusion criteria, but we have summarised them because these studies included interventions/comparators (AI technologies not included in this assessment) that were of interest.

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).

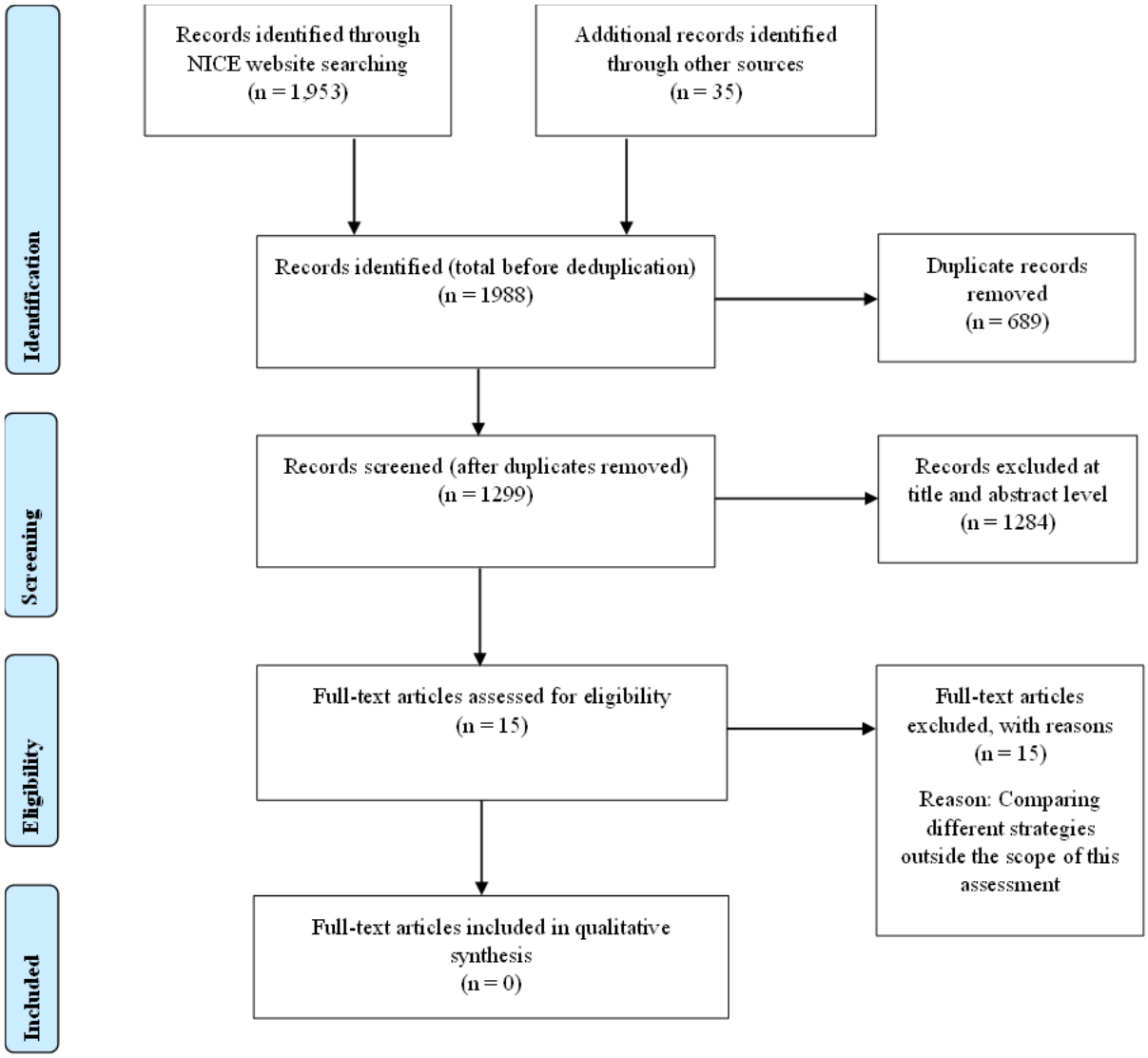


Figure 13. PRISMA Flow diagram for economic evaluation of using the AI for detection of lung nodules

5.2.2 Description of the evidence

Bajre et al. (2017)[72]

Bajre and colleagues used a decision tree structure to assess the cost-effectiveness of trained radiographers compared with radiologists for the reporting of chest x-rays in people suspected of having lung cancer. The model simulated a pathway for a hypothetical cohort of 1000 people being screened for lung cancer and the cost-effectiveness concluded at five years. The model started with a cohort of people who received a radiologist-reported chest x-ray or radiographer-reported chest x-ray. The pathway for both strategies were the same. The true disease status is known, characterised by the prevalence of lung cancer. People with lung cancer and had a positive result received a CT confirmatory test, which provided staging. The authors included stages I, II, II and IV. People with a false negative result presented later to the A&E department, where they were diagnosed with lung cancer and staged. People who had a false positive result following chest x-ray received a CT scan that confirmed no lung cancer was present. People with no lung cancer and had correctly been identified as negative by the chest x-ray received no further testing/imaging.

Information required to populate the model was obtained from the literature, and NHS reference costs. The model required information about the prevalence of lung cancer, sensitivity and specificity of radiologist-reported and radiographer-reported chest x-ray to identify lung cancer, as well as sensitivity and specificity for radiologist-reported CT scan to confirm lung cancer diagnosis and probabilities. Though not explicitly stated confirmatory diagnosis was made by the radiologist. The proportion of people diagnosed at first presentation were obtained from Cancer Research UK (CRUK) 2013.[74] Additionally, information was required about the probability of lung cancer by stage at second presentation following misdiagnosis. All costs included in the model were reported in 2014/15 prices. Costs were required for radiologist and radiographer reading of chest x-ray, cost of CT scans and total costs of treatment by stage. Authors were not explicit about which treatment people received. The benefit of the strategies was reported in terms of cases detected at first presentation and quality-adjusted life years (QALYs) yielded. Utility values by stage of diagnosis were obtained from Naik et al., 2015.

Several simplifying assumptions were made to have a workable model structure (Bajre et al., 2017 pg. 275):

- *Time taken to report a chest X-ray (CXR) is 2 min for both radiographers and radiologists*
- *False negatives present at A&E at a later date at which point disease may have advanced a stage (for patients at stage I to III)*

- *Sensitivity and specificity of radiographer reporting of CXR and radiologist reporting of both CXR and CT-scan is independent of disease stage or other patient characteristics such as age.*
- *QoL in the year following diagnosis (according to stage at diagnosis) is maintained in subsequent years*
- *There is no QoL impact arising from false positive reporting*
- *Findings for non-small cell lung cancer are representative for lung cancers in general*

The perspective and setting of the economic analysis were not clearly defined but it appears to be from the NHS and Personal Social Services (PSS) in a secondary care setting, based on the cost inputs. The results of the analysis were presented in terms of an incremental cost-effectiveness ratio (ICER), expressed as cost per QALY. The authors undertook probabilistic sensitivity analysis (PSA) to assess the joint uncertainty in key model input parameters: prevalence of lung cancer, sensitivity and specificity of radiologist and radiographer reporting of chest x-rays, lung cancer stage distribution at initial chest x-ray and stage progression following misdiagnosis. Authors stated the sampling distributions for the parameters included in the PSA but have not reported their parameters. The authors undertook threshold analysis but not one-way sensitivity analysis.

Authors reported disaggregated results for both strategies. Results were reported on the number of people expected to be diagnosed with lung cancer, QALYs yielded and treatment costs, all by stage. The QALYs yielded appeared to be high, with stage IV expected to yield more QALYs than stage III and II, respectively. There were modest QALY gains by strategy and by stage, with stage I having the greatest expected gain of 2.4 QALYs, favouring radiographer reporting. Radiographer reporting yielded more overall QALYs, but it was unclear with the inputs reported why the radiologist reporting QALYs was greater for stages II and stage IV. Radiographer reporting diagnostic and treatment costs were cheaper than radiologist reporting costs. Overall results showed that radiographer reporting of chest x-ray dominated radiologist reporting. PSA results showed that radiographer reporting continued to dominate radiologist reporting in 98% of the iterations. Based on the model structure, its inputs and assumptions, the authors concluded that trained radiographers can be used to report chest x-rays for the diagnosis of lung cancer.

Adams et al. (2021):[73]

Adams et al. study pursue two objectives:

1. To refine the categorisation at baseline time for a lung cancer screening management strategy that combines Lung-RADS and AI risk scores.

2. To determine the potential impact of such a management strategy on the recommendation for further follow-ups and costs associated with it, from a public payer perspective.

The initial hypothesis of the study is that adding an AI-based risk score can result in higher specificity, resulting in fewer follow-up investigations after baseline lung cancer screening, which might save costs.

The researchers have developed a deep learning (DL) algorithm as the core component of their lung cancer screening management strategy. Also, they used the results from studies that aimed at localising and predicting the risk of lung cancer by using the Low Dose CT-Scan (LDCT) modality from the NLST database. Developing such an algorithm included three steps: 1) training the developed algorithm, 2) tuning phase for the algorithm, and 3) testing the validity and reliability of the results for localising and predicting the risk score for lung cancer. Seventy percent of those studies were used to train the algorithm, 15% for tuning, and 15% to test the algorithm. Testing the algorithm was undertaken by six US-based experienced radiologists, who had mean experience of eight years (range 4 to 20 years).

By incorporating a deep learning algorithm in the previous section, the researchers developed a management strategy for lung cancer screening at baseline. The developed algorithm follows four principles below:

1. *"Patients with malignant nodules obtain definitive diagnosis and management in a shorter amount of time.*
2. *Patients with benign nodules are subjected to fewer (unnecessary) investigations between annual LDCT screening studies.*
3. *Patients with true lung cancer with a baseline LDCT study classified as Lung-RADS category 1 or 2 (resulting in a false-negative) were considered to have a delayed diagnosis because no additional investigations are recommended for these categories as per Lung-RADS and the nodule may not be reviewed until the next annual screening.*
4. *Patients with a benign nodule classified as Lung-RADS category 3, 4A, 4B, or 4X (resulting in a false positive) were considered to have unnecessary investigations".*

For demonstrating the advantage(s) of an AI-based lung cancer screening management strategy; the researchers upgraded lung nodules in categories 1 and 2 to category 3 if it deemed those nodules

(category 1 or 2) are at higher risk according to the AI algorithm results. The researchers believe this higher diagnostic accuracy through upgrading the nodule category to 3, causes reducing the time for follow-up from 12 months to 6 months and saving the costs associated with later diagnosing of lung cancer.

A Receiver Operator Characteristics Curve (ROC) was used to determine the sensitivity and specificity for each of six radiologists who applied Lung-RADS using confirmed lung cancers per NLST as the gold standard. The researchers have also used an operating point for the AI risk score to match the average sensitivity of the six radiologists using Lung-RADS, and the respective specificity at that point.

The costs associated with follow-up investigations were calculated based on the Medicare physician Fee Schedule non-facility National payment Amount in US dollars in 2019. Costs were required for both professional (human resources costs), and technical (non-human related costs). The costs for categories 4A, 4B, and 4X, were calculated as the minimum and maximum possible amounts because for these categories Lung-RADS allows for clinical discretion in follow-up.

The results of the study have been provided by weighted and unweighted means of sensitivity and specificities for both radiologists and AI-assisted lung cancer screening management strategies. The unweighted means of sensitivity and specificity for six radiologists were 91% (±7%) and 61% (±15%), respectively. The weighted sensitivity and specificity for six radiologists were 91% (±7%) and 66% (±16%), respectively. By using a threshold of 0.27 for the AI-assisted predicted risk score, the sensitivity for both unweighted and weighted situations was 91%, which is matched with the average of six radiologists' reads sensitivity, but the specificity only has been reported for the weighted data set and it is 96%. In addition, the AI-assisted management strategy produces a total of 41 upgraded classifications to category 3 (equal to 0.2% of all classifications). It also produces 5,750 classifications (30% of all cases) as a downgraded category. Upgrading from categories 1 and 2 to category 3 resulted in a recommendation for LDCT at six months. Also, 41 upgraded cases caused an average of 6.8 additional LDCT examinations per radiologist (= 41 classifications/6 radiologists) using the proposed AI-informed management strategy compared with the initial Lung-RADS recommendations across the weighted representative cohort.

The mean cost saving was reported as US$72 per screened patient. Given the different categories for nodule at category 4 of risk (4 A, 4B, and 4X), net cost savings were estimated to be US$242 per patient screened. The study concluded that incorporating an AI-assisted lung nodule management strategy will cause a substantial cost-saving that is related to increased specificity that results in fewer follow-up investigations in a lung cancer screening program. The authors acknowledged the following limitations:

The study results and conclusions have some limitations as below:

- Over-estimation of the cost-saving due to patients' incomplete compliance to the follow-ups.
- Differences between time intervals for follow-up between guidelines and practice.
- The results of the study cannot distinguish between diagnosis and treatment costs in subsequent years.
- The results were not informed by costs of the AI software as those costs were not established at the time of the study.
- The study used the NLST data, which targeted North America population, potentially lacking generalisability of the results to other locations with different composition of ethnicities.

6 PRELIMINARY MODEL – METHODS AND RESULTS

Two separate modelling approaches were undertaken by the EAG. This section describes a simpler approach to assessing the cost-effectiveness for AI-assisted detection of actionable nodules (which is one of the key steps in lung nodule detection and management) in the screening population. This approach allows direct use of test accuracy evidence (sensitivity and specificity) on the detection of actionable nodules as the key model parameter input. Available test accuracy evidence was insufficient for a similar analysis to be undertaken for the symptomatic and incidental populations (e.g., studies only reported sensitivity without reporting specificity).

6.1 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 14** below.

An actionable lung nodule was defined as a nodule that, when identified, would warrant further investigation and surveillance or definitive diagnostic workup according to the BTS guidelines. The key features for defining an actionable nodule in the BTS guideline are its size (≥5 mm) and lack of features strongly suggestive of a benign nature, but other factors including nodule type (solid or sub-solid), 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, ≥8 mm), which are associated with both subsequent nodule management pathways and cancer risks. We considered this appropriate as it would capture all the short-term costs and events associated with identifying and analysing actionable lung nodules.

6.2 Strategies

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

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 other healthcare professional to identify lung nodules and measure their sizes, with or without additional features such as classifying the type of the 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 measure the size of their lung nodule if present.

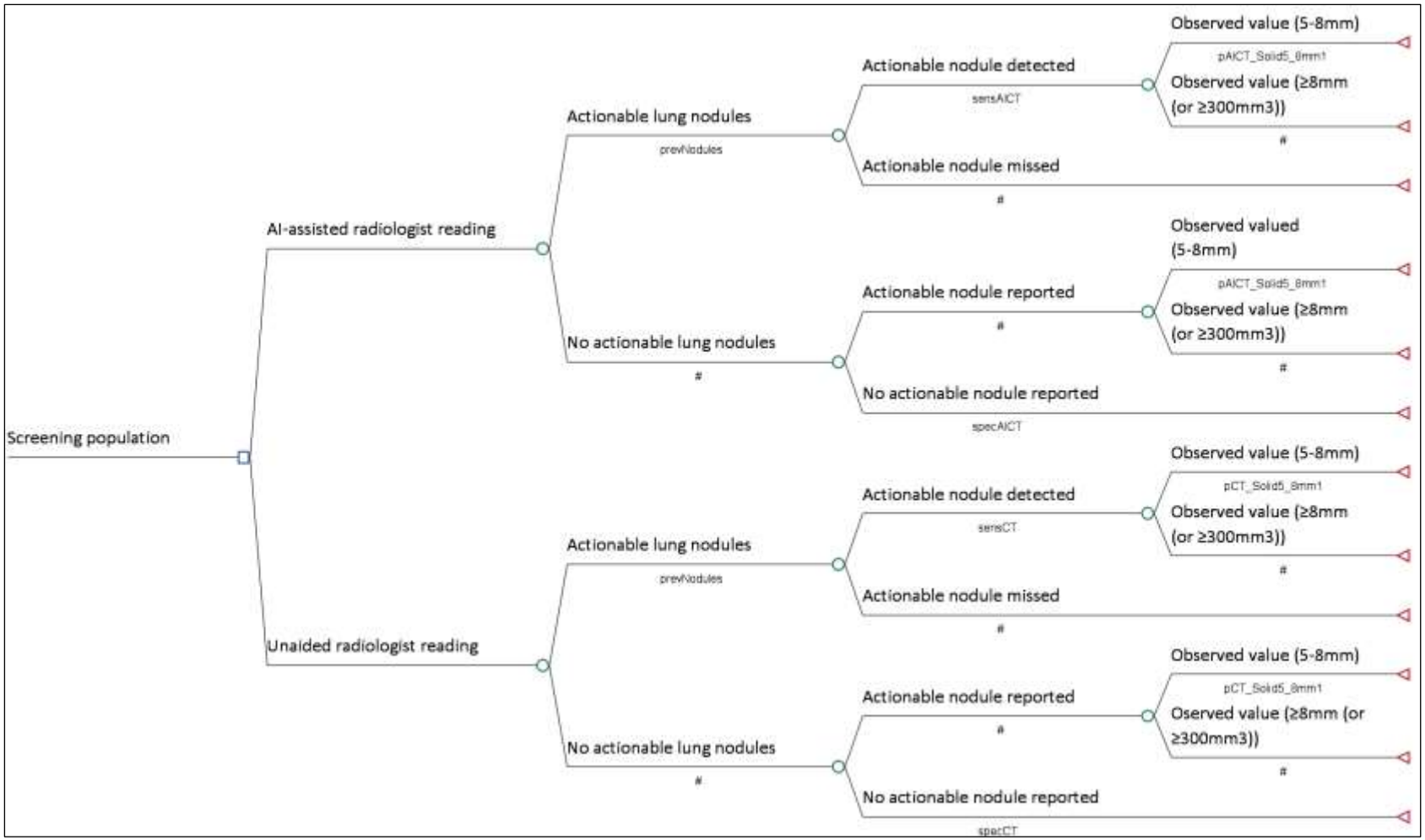


Figure 14. Illustrative model structure for the detection of actionable lung nodules

6.3 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.

6.3.1 Prevalence

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, 0.2357), obtained from the UK Lung Screen Uptake Trial, which was the largest UK study reporting this information.[75]

6.3.2 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 only from one study conducted in a screening population (Lo et al. 2018),[52] reported in section **3.3.1.2** of the test accuracy review.

Table 35. Test accuracy estimates for identifying actionable nodules by test strategy

Parameter	Value	95% confidence interval	Source
Screening population			
AI-assisted radiologist reading			
Sensitivity	72.50	69.20 – 75.80	Lo et al.[52]
Specificity	84.40	82.40 – 86.40	
Unaided radiologist reading			
Sensitivity	60.10	56.80 - 63.40	Lo et al.[52]
Specificity	89.90	87.90 - 91.90	

We define:

- True positive: actionable lung nodule detected

- True negative: actionable lung nodule neither present nor reported
- False positive: 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: actionable nodules that were not identified using each strategy

6.3.3 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 2021/22 prices.

Computer software

Cost 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 Lo et al 2018 which provided test accuracy data. Further details about our criteria used to be included in the economic analysis are reported in Section 7.4.5.

Time taken to read the CT scan and report findings

For 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 purpose.[76]

Our test accuracy review found that the time taken to read/report CT scans reduced with AI-assistance in most studies (**section 3.5.2**). 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 Lung Screen Uptake Trial[25] (assumed as mean value as the interquartile range 5-15 was symmetrical around the median) and assumed the time would be shorter for AI assisted readers. In **Table 36** 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 was 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 compared with analysing lung cancer screening images, which tend to be undertaken in batches during a protected time.

These alternative reporting times were not used here but were used in full model for respective population, to be described in **section 7**.

Table 36. Resource use associated with reporting CT scans

	Population of interest		
	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. 2022)[25]
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		
CT, computed tomography			

Table 37. 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) (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) (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 2020/21 (RD22Z- CT scan of one area, with pre- and post-contrast)
CT, computed tomography; PSSRU, Personal Social Services Research Unit		

6.3.4 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 one for people correctly identified with an actionable nodule (≥5mm and no clear benign features), and zero for all others. This was tallied to give the denominator for the incremental cost-effectiveness ratio (ICER), expressed as cost per person with an actionable lung nodule detected.

6.3.5 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 results presented on a tornado diagram.

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 unaided reading.

6.4 Results

6.4.1 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 38 presents the estimates for costs and additional people correctly identified with an actionable nodule with the use of AI-assisted radiologist reading compared to unaided radiologist reading in a screening population. These results show that AI-assisted radiologist reading (ClearRead CT) is approximately £2,900 cheaper and expected to correctly identify an additional 25.5 people with actionable nodules per 1,000 CT screens: hence, dominating the unaided reading strategy.

Table 38. Deterministic results based on expected costs and expected identification of people with actionable lung nodules (screening population of 1,000 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	2,900	123.8	-25.5	Dominated
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

6.4.2 Sensitivity analysis results

Deterministic sensitivity analysis was conducted by varying key model input parameters by their ranges or when 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 15**.

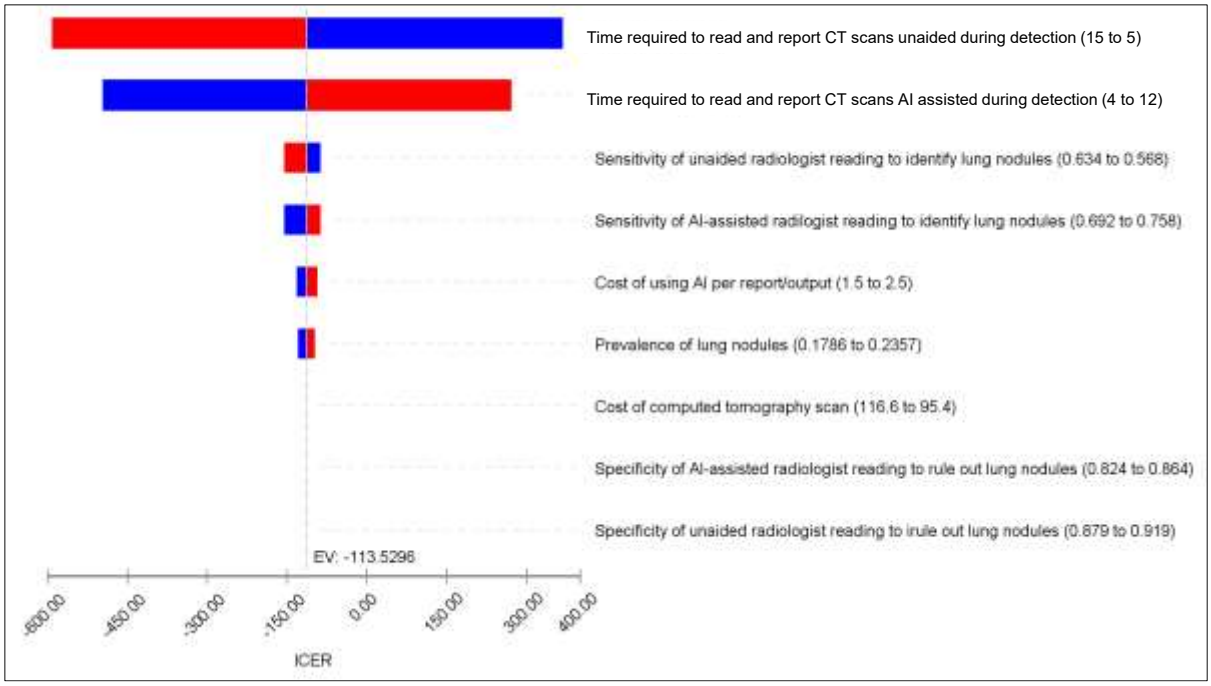


Figure 15. Tornado diagram of the impact to the cost per actionable lung nodule correctly identified by changing individual parameters (screening population)

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 versus unaided radiologist reading for identifying actionable lung nodules. However, varying these inputs within these limits are unlikely to change the ICERs outside of acceptable thresholds.

6.4.3 Scenario analysis results

Here we present the results for the scenario analyses. In these scenarios, we considered other inputs/assumptions for alternative sources to assess the impact to the results. Based on the alternative sources of evidence or assumptions made to key parameters (prevalence of detecting actionable lung nodules, assuming same reading and reporting time with/without AI assistance and an increase reading and reporting time with AI assistance), these results (see **Table 39**) were robust to changes made.

Table 39. 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	2,900	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)[3]					
AI-assisted radiologist reading (ClearRead CT)	127,600	-	204.7	-	-
Unaided radiologist reading	130,500	2,900	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	-	169.7	-	-
AI-assisted radiologist reading (ClearRead CT)	132,500	2,000	204.7	35	57
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	6,900	149.3	25.5	270
CT, computed tomography; QALY, quality adjusted life-year					

Superseded-see erratum

6.4.4 Discussion

The preliminary model provides a relatively straightforward approach to assessing cost-effectiveness of AI-assisted detection and analysis of lung nodules for chest CT scan images. However, a major

limitation for this simpler approach is that the test accuracy evidence related to 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 the impact of AI assistance on subsequent nodule management through analysis of surveillance CT scans to be evaluated. Consequently, we developed a more comprehensive decision analytical 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 link the evidence on initial nodule detection to subsequent nodule management pathway according to the BTS guidelines and then to health outcomes, the EAG further conducted simulations to fill in this major evidence gap. Further details are described in the following section.

7 DE NOVO COST-EFFECTIVENESS ANALYSIS (FULL MODEL) – METHODS

7.1 Developing the model structure

Given the limitations of the preliminary model mentioned above, we developed a full economic model that would enable the assessment of 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 vary depending on the specific population where appropriate. Further details of the population are detailed in Section 7.4.2. For people undergoing CT surveillance for previously detected nodules, the surveillance component of the model can be utilised.

The decision model follows the illustrative pathways shown in **Figure 16**. In people undergoing a CT scan in which lung nodules may be identified, 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, their type and size according to 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 was populated with appropriate information (see **Section 7.4**). 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 treatment of people with cancer.

7.2 Strategies

The model compares AI-assisted radiologist reading to 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 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 measure the size of their lung nodule if present.

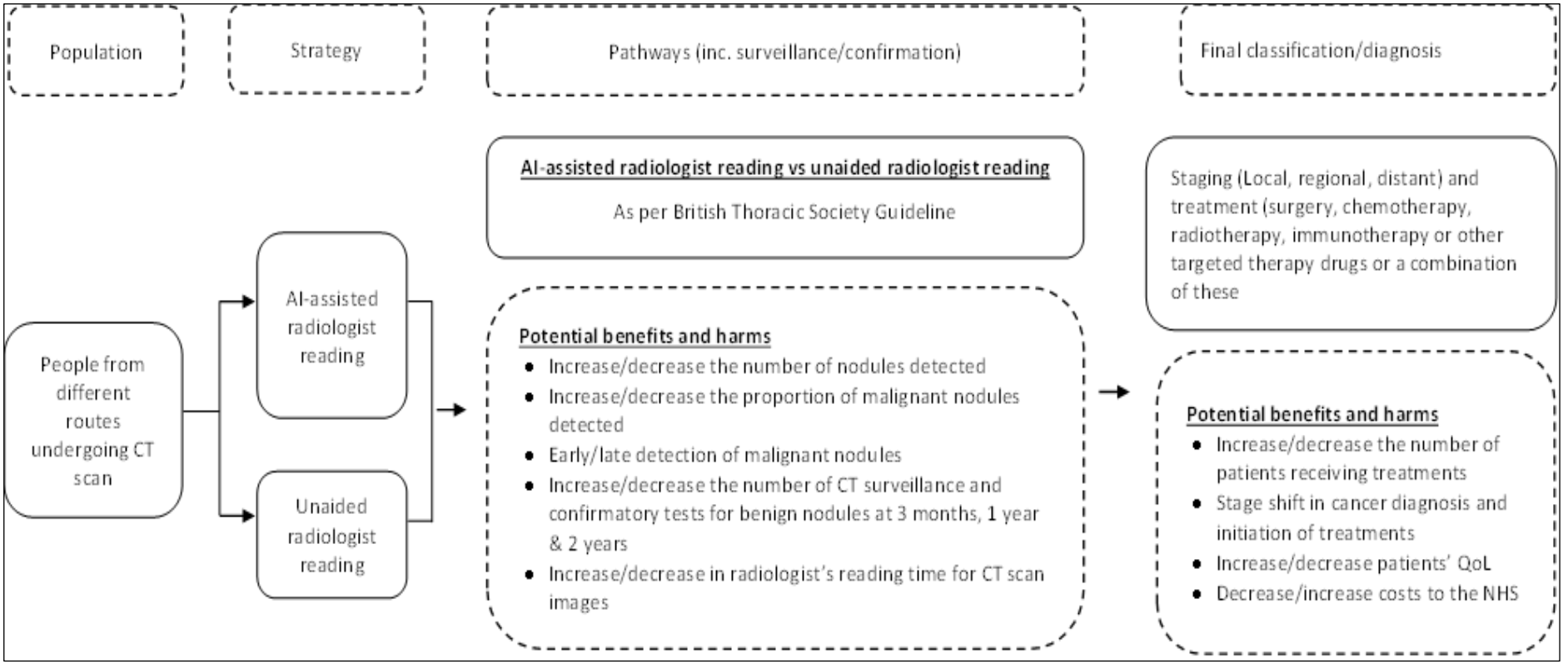


Figure 16. Illustrative structure of the clinical pathways

AI-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 (see **Figure 17**). People who have been identified as having a lung nodule, the nodule will be further assessed for its type (e.g., solid or sub-solid) as well as the size of the lung nodule. 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;[11] also called 'risk dominant nodule'). The measurement of the primary nodule would be undertaken by a radiologist (or other trained professionals) with/without assistance of AI software and categorised as follows: solid (<5mm, 5 to <8mm and ≥8mm) and sub-solid (<5mm and ≥5mm). For people with a lung nodule which 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.

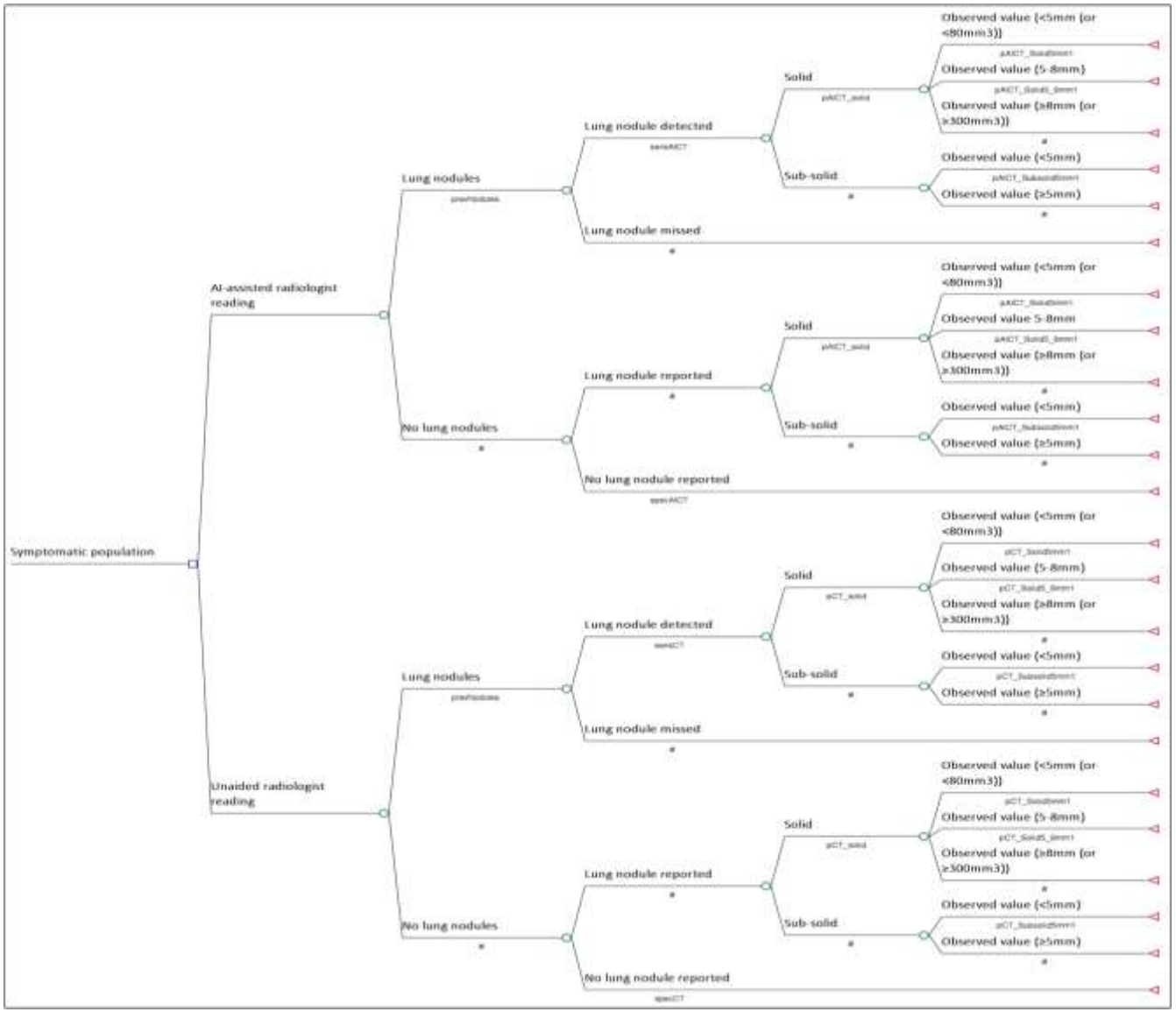


Figure 17. Illustrative model structure for the detection of lung nodules

7.3 Natural history

Our natural history model was developed to model the growth/disease progression of malignant disease, separately for solid nodules and sub-solid 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 volume doubling time (i.e., the time required for the tumour to double its volume),[77] which is based on information obtained from Treskova et al., 2017.[69] Details of our nodule growth model and its development can be found in **Table 57**.

7.4 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 section **13.6, Appendix 6**) as well as clinical expert opinion. One major challenge in the use of evidence from our test accuracy review to inform the decision analytical model arose from the mismatch between outcomes reported in the test accuracy studies (such as sensitivity and specificity for detecting nodules of various sizes and types, e.g. solid nodule ≥ 6 mm, and/or the precision, accuracy and concordance of measuring nodule size/volume, e.g. mean and standard deviation of nodule sizes measured by unaided reading and AI-assisted reading or agreement in detecting nodules ≥ 3mm between unaided readers and AI-aided readers) on the one hand, and the BTS categorisation of the primary nodule on the other hand (which is based on the combined information of nodule type and specific nodule size categories obtained from either unaided or AI-assisted reading, see **Figure 4**). In order 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 sub-solid nodules) which dictates subsequent clinical management (e.g., discharge, further CT surveillance, further 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 section below.

7.4.1 EAG 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 sub-solid nodules among identified primary nodules – this differs between different populations of interest.
- The 'true' mean sizes of the primary nodules – these differ between different population of interest and between solid and sub-solid 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/better consistency being one of the purported advantages for AI-aided reading.
- The measurement accuracy (systematic error in measurements, e.g., consistently over- or under- estimate the 'true' nodule size) – this may differ between unaided and AI-aided reading.

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 sub-solid 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 then be estimated. Comparison of results between (1) and each of (2), (3) and (4) provides information concerning mis-categorisation 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 for the simulation are presented in **Appendix 8**.

For the decision analytical 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, 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. **Figure 18** to **Figure 20** provides an overview of model parameters used and the sources of these data. Further information are detailed in the report sections below.

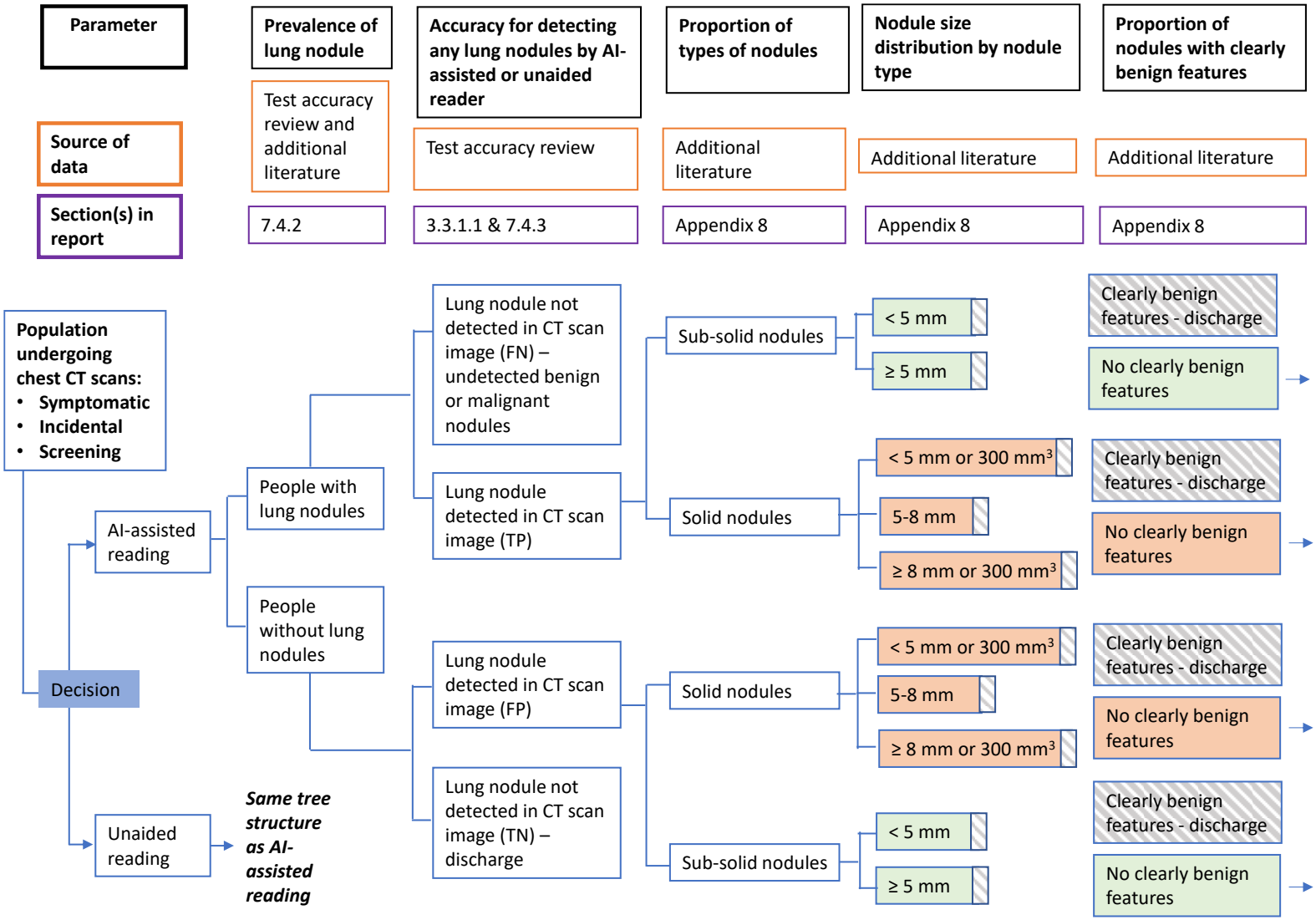


Figure 18. Abbreviated representation of the decision tree, required model parameters and data source (further parts shown in Figure 18 & 19)

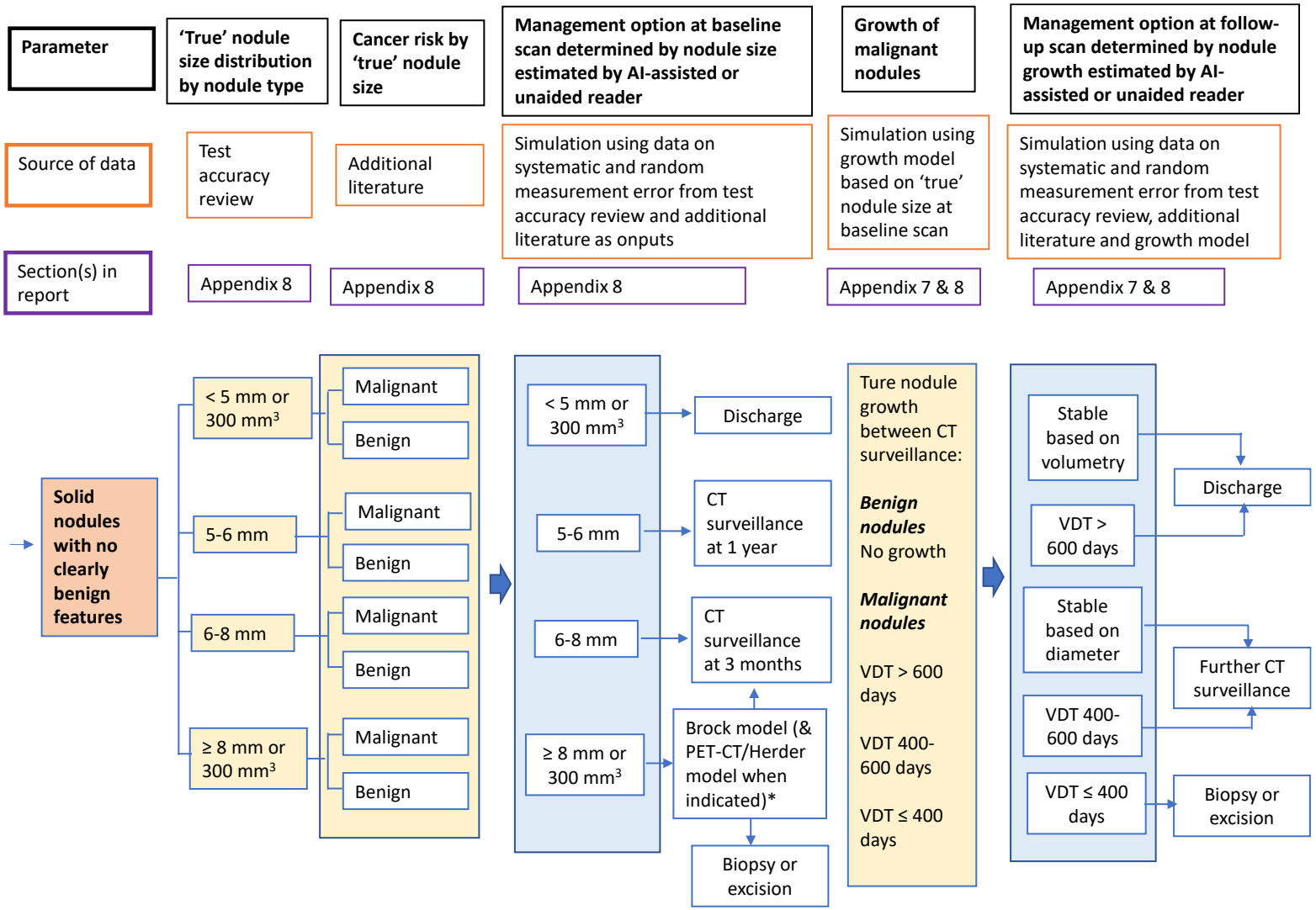


Figure 19. Abbreviated representation of the solid nodule part of the decision tree, required model parameters and data source (continued from Figure 17)

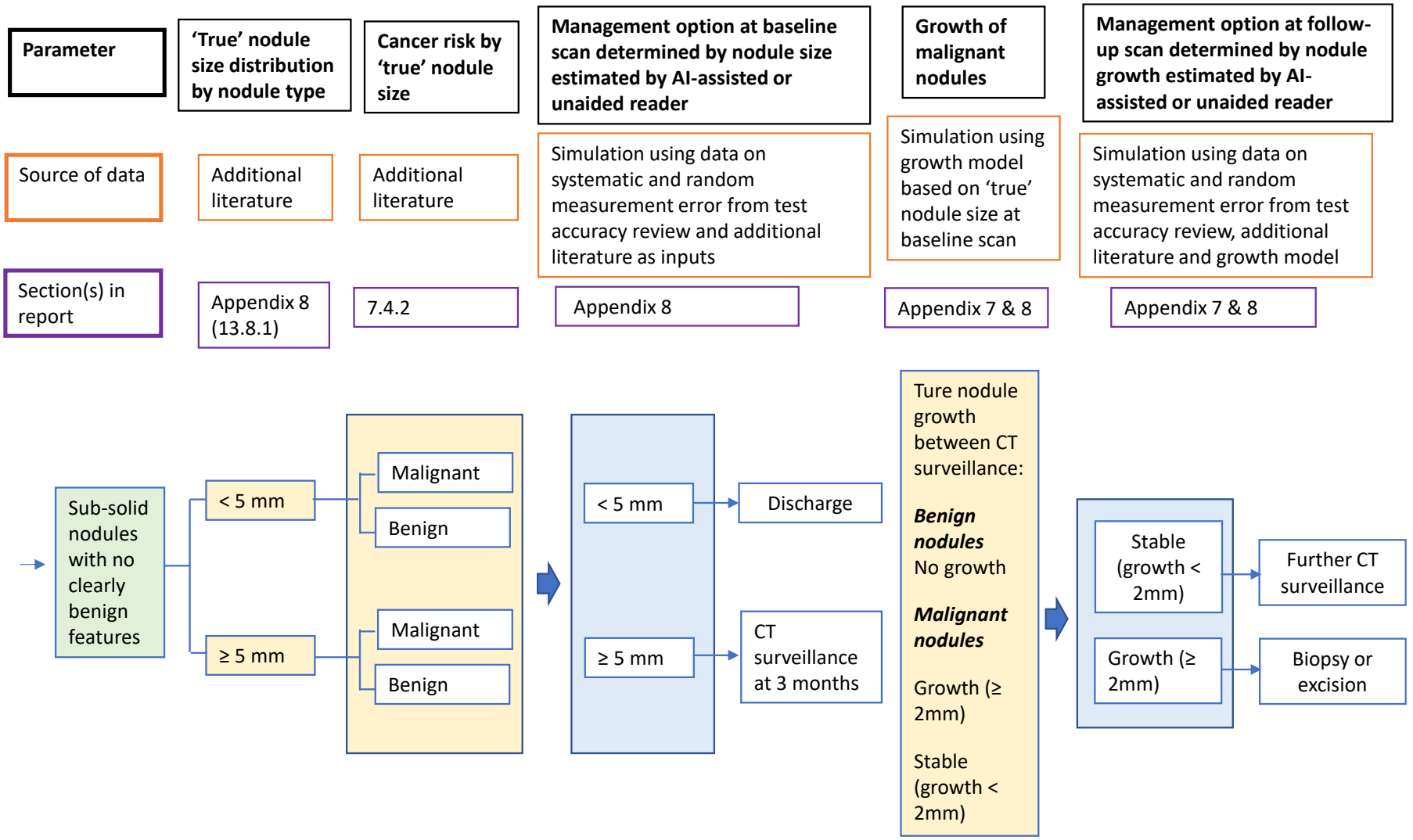


Figure 20. Abbreviated representation of the sub-solid nodule part of the decision tree, required model parameters and data source (continued from Figure 17)

7.4.2 Prevalence

Prevalence of lung nodules

The model required information about the prevalence of lung nodules in three of our four populations of interest. We presumed that the proportion of people with lung nodules would be different across different populations. Prevalence information was not required for people undergoing surveillance, as by definition, people undergoing surveillance would have a previously detected nodule. In **Table 40** we report this information. We note that people may present with more than one lung nodule; however, we assumed that people with lung nodules have one primary lung nodule, the assessment of which guides clinical management in line with the BTS guideline as described above.

Table 40. 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, 0.9763)	Kozuka et al., 2020[57]	Only study identified
Incidental (CT scan done for other reasons)	0.13 (0.02, 0.24) [a]	Callister et al., 2015[11]	Evidence review for 2015 BTS guidelines
Lung cancer screening	0.509 (0.4868, 0.5312)	Field et al., 2016[78]	Largest UK-based study that reported prevalence of any nodules
[b] CT surveillance of a previously detected nodule	Not applicable	-	-

[a] Range
[b] Not applicable because all of the people in the model would have an indeterminate lung nodule
CI, confidence interval; CT, computed tomography.

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 sub-solid, in line with the BTS guideline.[11] Here we assumed that, if a nodule was identified, then it would be correctly categorised as solid or sub-solid. We required the proportion of lung nodules by type and by reason for undergoing a CT scan. In **Table 41,** 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.

Table 41. Proportion of detected risk-dominant nodules that are solid/sub-solid

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.,[57] Table 1 518 solid nodules, 151 sub-solid nodules
Sub-solid	0.226	
Screening population		
Solid	0.939	Hwang et al.,[48] Table S3 4357 solid nodules, 285 sub-solid nodules
Sub-solid	0.061	
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).		

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 (see **Table 42)**. The information was derived from the publication by Horeweg et al.[3] Their study is based on 7,155 Dutch participants in the screening group of the NELSON trial. Lung 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 sub-solid nodules, we chose this study as model input as the population was rated as most applicable to a UK screening population.

Table 42. 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-<6mm	Assumed same as screening	Assumed same as screening	0.0089 (0.005, 0.016) (Horeweg et al., 2014)[3]	Assumed same as screening
6-8mm	Assumed same as screening	Assumed same as screening	0.011 (Horeweg et al., 2014)[3]	Assumed same as screening
≥8mm	Assumed same as screening	Assumed same as screening	0.094 (Horeweg et al., 2014)[3]	Assumed same as screening
Sub-solid				
≥5mm	Assumed same as screening	Assumed same as screening	0.036 (Horeweg et al., 2014)[3]	Assumed same as screening

7.4.3 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 **Section 1.1** 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 sensitivity and specificity of these strategies to identify people with malignant/benign nodules.

Three studies[51, 57, 59] were identified that reported these outcomes. Their study characteristics, strengths and limitations are reported in **Table 43**. The study by Zhang et al.[59] was immediately discounted as it compared double reading with software use under laboratory conditions to double reading by different readers without software use in clinical practice.

From the remaining two studies,[51, 57] we choose the study by Kozuka et al.[57] as CEA input for the symptomatic population as this study was the only identified study that was actually performed in patients suspected of having lung cancer. We also used the study by Kozuka et al.[57] as input for the incidental population as the readers were less experienced radiologists, which was judged to be applicable for general radiologists assessing CT images in A&E in UK practice. For the screening population, we decided to use the senior group (experienced chest radiologists) from the study by Hsu et al.[51] 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 43. Comparative studies reporting detection accuracy for any nodules that could be used as CEA model inputs and their advantages and disadvantages (3 studies)

Study	Study details	Sensitivity (per-subject)	Specificity (per-subject)	Advantages	Disadvantages
Hsu 2021[51]	Mixed population: 1 hospital in Taiwan; 150 consecutive cases with lung nodules ≤1 cm or no nodules: 93 clinical routine; 57 screening population. Low dose (n=57), standard dose (n=93), no contrast, slice thickness 2.5 mm. MRMC study, ClearReadCT with vessel suppression and nodule detection: 6 chest radiologists - 3 less experienced (residents in radiology with >6 months of chest CT experience) and 3 experienced chest radiologists (5, 10 ad 25 years of experience). Reference standard: Consensus expert reading (2 readers).	Per-nodule sensitivity (340 nodules) [D] Mean 64% (95% CI 62-66%) [C] Mean 80% (95% CI 81-85%) (p<0.001) Senior readers only: [D] Mean 74% (95% CI 72-77%) [C] Mean 84% (95% CI 82-86%) (p<0.001)	52 patients without nodules: [D] Mean 80% (95% CI 78-81%) [C] Mean 83% (95% CI 82-85%) (p=0.256) Senior readers only: [D] Mean 87% (95% CI 85-89%) [C] Mean 88% (95% CI 87-90%) (p=0.729)	Consecutive sampling; Mixed population but separate data for screening population reported; MRMC study included 6 readers and reports accuracy separately for 3 experienced (senior) chest radiologists (high applicability for UK screening and symptomatic populations).	Taiwan, 1 hospital (not a UK or North-Western European population, nodule prevalence might be different); 57 screening LDCT images (small sample size); Lung nodules ≤1 cm only (inclusion of only small nodules might affect sensitivity); 2.5 mm slice thickness (UK ≤2 mm, might affect accuracy); MRMC study (radiologist performance under laboratory conditions might be not representative of clinical practice); No subject-level sensitivity reported, only per-nodule sensitivity (per-subject sensitivity might be higher); Only reported mean sensitivity and mean specificity, no 2x2 data, no data for individual readers (no decimal places reported, cannot calculate exact estimates).
Kozuka 2020[57]	Symptomatic population (suspected lung cancer): Random 120 chest CT images from 1 hospital in Japan. Standard dose; no contrast; 1 mm slice thickness. MRMC study, InferRead CT Lung (Infervision);	111 subjects with nodules, pooled reader A + reader B: [D] 68.0% (151/222) (95% CI 61.4-74.1%);	6 subjects without nodules, pooled reader A + reader B [D] 91.7% (11/12) (95% CI 61.5-99.8%); [C] 83.3% (10/12)	Only study on symptomatic population; random selection; 1 mm slice thickness (applicable to the UK); reported 2x2 data individually for Reader A and Reader B.	Japan, 1 hospital (not a UK or North-Western European population, nodule prevalence might be different); 117 CT images included in analyses (small sample size); MRMC study (radiologist performance under laboratory

Study	Study details	Sensitivity (per-subject)	Specificity (per-subject)	Advantages	Disadvantages
	2 less experienced radiologists (1 and 5 years of diagnostic experience); Reference standard: Consensus expert reading (3 readers).	[C] 85.1% (189/222) (95% CI 79.8-89.5%) ($p < 0.001$)	(95% CI 51.6-97.9%) (no level of significance reported)		conditions might be not representative of clinical practice); 2 less experienced radiologists (1 year and 5 years of experience) (applicability concerns to UK reading practice for symptomatic population); Only 6 CT images without nodules (wide 95% CI for specificity; 1 additional FP case in 1 reader resulted in an apparently big difference in pooled point estimates).
Zhang 2021[59]	Screening population: 860 consecutive patients from 1 hospital in China (part of NELCIN-B3 project); Low dose; no contrast; 0.625-1.0 mm; InferRead CT Lung (Infervision) 1 radiology resident with supervision of 1 experienced radiologist - With software (MRMC study); Without software (clinical practice); Reference standard: Consensus expert reading (2 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, 1 hospital (not a UK or North-Western European population, nodule prevalence might be different); Different readers with and without software use: [C] Performance of 1 resident and 1 radiologist only; [E] 14 different residents and 15 different radiologists; 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; LDCT, Low-dose computed tomography; MRMC, Multi-reader, multi-case study.

Table 44. Test accuracy estimates to identify any lung nodule by reason for undergoing CT scan

Parameter	Value	95% confidence interval	Source
People with symptoms suggestive of lung cancer			
AI-assisted radiologist reading			
Sensitivity	85.14	79.80 - 89.50	Kozuka et al., 2020[57]
Specificity	83.33	51.60 - 97.90	
Unaided radiologist reading			
Sensitivity	68.02	61.40 - 74.10	Kozuka et al., 2020[57]
Specificity	91.67	61.55 - 99.88	
Incidental (CT scan done of other reasons)			
AI-assisted radiologist reading			
Sensitivity	85.14	79.80 - 89.50	Kozuka et al., 2020[57]
Specificity	83.33	51.60 - 97.90	
Unaided radiologist reading			
Sensitivity	68.02	61.40 - 74.10	Kozuka et al., 2020[57]
Specificity	91.67	61.55 - 99.88	
Screening			
AI-assisted radiologist reading			
Sensitivity	83	79 - 86	Hsu et al., 2021[51]
Specificity	88	85 - 91	
Unaided radiologist reading			
Sensitivity	73	69 - 77	Hsu et al., 2021[51]
Specificity	86	83 - 90	
CT, computed tomography			

From as far as possible, we extracted information from individual studies identified form our test accuracy systematic review to populate 2 x 2 tables to derive study specific test performance for both strategies. We define:

- 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 incorrected 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, thus 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.

7.4.4 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 due to 'watchful waiting' when people are referred to CT surveillance. During surveillance, people undergo imaging aimed at measuring the growth of lung nodules, which is characterised by its volume doubling time (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 outside of this threshold may be referred to further surveillance or discharged.

7.4.5 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, software technologies, and treatment associated with lung cancer. All costs are presented in 2021/22 prices and after the first year, both costs and benefits were discounted at a rate of 3.5% per annum. Costs obtained from the literature through systematic reviewing were uprated to current prices where necessary using the Hospital and Community Health Services (HCHS) index from Unit Costs of Health and Social Care 2022.

Computer software

There is paucity of test accuracy and cost data for some of the technologies included in the final scope of this assessment. In order 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 45**, we outlined 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 in the symptomatic population, we included InferRead CT Lung (Infervision) technology. It was noted that there were different costing structures in place, so attempts were made to obtain/derive a per scan cost.

Table 45. Technologies outlined in scope against our selection criteria for the base-case economic analysis

Technology (Company)	Criteria		
	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[45])
AVIEW LCS+ (Coreline Soft)	Not available	No	No
ClearRead CT (Riverain Technologies)	Yes	Yes	Yes (Accuracy, Screening population[54] and unclear indication[53]) Yes (Concordance, Mixed population[56])
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[61])
Veye Lung Nodules (Aidence)	Yes	No	Yes (Accuracy, Mixed populations[31, 64]) Yes (Concordance, Mixed population[31])
VUNO Med-LungCT AI (VUNO)	Not available	No	No

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 mins, 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 there would be additional costs incurred (visit to multidisciplinary team, further CT scans and biopsy).

Treatment costs

Total treatment costs by stage of disease were obtained from Bajre et al.,2017[72] which were originally from Cancer Research UK 2014.[79] Total costs included retreatment costs and were reported in price year of 2014/15. These costs were obtained from the literature and uprated to current prices (2020/21) using the Hospital and Community Health Services (HCHS) index from Unit Costs of Health and Social Care 2022.[76]

Table 46. 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) (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) (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 2020/21 (RD22Z- CT scan of one area, with pre- and post-contrast)

Multidisciplinary team	£146	National schedule of NHS costs 2020/21 (CDMT_OTH other cancer MDT meetings)
Guided needle biopsy	£1670	NHS reference schedule (DZ71Z- minor thoracic procedure, guided needle biopsy)
Bronchoscopy		
PET scan	£1161	RN01a- PET-CT of one area, 19 years and over
Treatment		
Stage I	£16,740	Bajre et al., 2017[72]
Stage II	£19,072	
Stage III	£21,408	
Stage IV	£13,342	
CT, computed tomography; PET-CT, positron emission tomography and computed tomography; PSSRU, Personal Social Services Research Unit		

Superseded-see erratum

7.4.6 Utility values

The utility values that were used to derive the quality adjusted life years (QALYs) for people with lung cancer were mainly obtained from Bajre et al.,[72] which were originally obtained from Naik et al., 2015. Briefly, these authors collected health-related quality of life information using the 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), patients 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 (Rickets et al., 2020).

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 three months or one year). Also, we assumed that 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 UK-specific general population norms.

We assumed that a disutility of –0.2 associated with undergoing a biopsy with a duration of three months.

7.4.7 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 for people without lung cancer was obtained from the Office of National Statistics (ONS) and an average of the mortality rate for males and females 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, (Jacobs et al., 1999)[80] but we did not apply any increase to mortality for individuals with benign lung nodules.

7.4.8 Outcomes

Three different outcome/effectiveness measures were used in the analysis: correct identification of actionable nodules, cancer correctly detected and treated and quality-adjusted life years (QALYs).

Cost per correct identification of actionable nodules

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

Cost per cancer correctly detected and treated

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

Cost per QALY

Four sets of QALY values were estimated for use in the model. First, the QALY values for people who do not have any lung nodules. Second, the QALY values for people with benign lung nodules. Third, QALY values for treated for lung cancer and fourth, people who have undiagnosed lung cancer.

7.4.9 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
- For lung nodules that were not identified at initial CT scan but were later detected or diagnosed as cancer, we assumed that these lung nodules 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 screening population and hence 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 would not grow afterwards.
- For the AI-assisted reading strategy, we assumed that 95% of people with benign nodules would be discharged at the one-year CT surveillance and 5% would be discharged at the two-year CT surveillance. For the unaided reading strategy, we assumed that 95% of people would be discharged at the two-year CT surveillance and 5% at the one-year CT surveillance.
- Among false negative cases at initial CT scan, we assumed 0.04% will be malignant(Horeweg et al.)
- We assumed a utility decrement associated with undergoing a biopsy as –0.2.[81, 82]
- There would be no cancers caused by radiation exposure.

7.4.10 Analysis

The economic analysis was undertaken from the perspective of the NHS and PSS and according to the Consolidated Health Economic Evaluation Reporting Standards (CHEERS).[70] The results of the analysis are presented in terms of an incremental cost-effectiveness ratio (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.[20] A deterministic analysis was undertaken for the base-case for the primary and secondary outcome measures.

We undertook probabilistic sensitivity analysis (PSA) to determine the joint uncertainty in model input parameters. We undertook the PSA based on the outcome of cost per QALY gained only. In the PSA, 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 are 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.. 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 (WTP) thresholds per QALY gained, with the results plotted on a cost-effectiveness acceptability curve (CEAC).

Additionally, we undertook several 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 results presented on a tornado diagram.

Scenario analyses

Given the limited evidence available, we had to use information from different studies and sources, often with some concerns related to risk of bias and applicability, to link evidence on diagnostic accuracy of AI-assisted reading compared to 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 through undertaking scenario analyses for different values for each variable, and structures of the economic model. We identified four parameters which 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 CT scans

In the detection phase of the model, we explored using prevalence of any lung nodules detectable at 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 47** shows the prevalence information that we used in scenario analysis.

Table 47. 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)	0.33 (Callister et al., 2015)	0.13 (Callister et al., 2015)	0.380 (Lancaster et al., 2021)

Accuracy for identifying actionable nodules

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

Time taken to read CT scans

The time taken to read CT scans reduced with AI-assistance in most studies included in our review (references). However, these studies were predominantly conducted under research conditions and there is uncertainty with regard to how AI assistance may impact on read/reporting time in real clinical practice. In the base-case we assumed that time required to read and report a CT scan image would be shortened from 10 minutes for unaided readers to eight minutes for AI assisted reading. In **Table 48**, we report the time taken (expert opinion), by population. In scenario analyses, we explored the possibility of varying this time for different strategies.

Table 48. Resource use associated with reading and reporting CT scans

	Population of interest			
	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			

CT, computed tomography

7.4.11 Areas beyond the scope of the assessment

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 where evidence on radiologist's reading time and/or radiology turnaround time related to the use of the software is found, it will be taken into account in the estimation of costs.

8 DE NOVO COST-EFFECTIVENESS ANALYSIS (FULL MODEL) - RESULTS

8.1 Base-case results

The full model comprising two stages provides a quantitative framework to link the diagnostic accuracy using AI-assisted reading compared to unaided reading of CT scans for identifying any lung nodules, to determining those requiring further actions, 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 terms of QALYs. We first present findings related to intermediate outcomes in **Table 49**, then summarise deterministic results for the following outcomes: cost per correct identification of actionable nodules, cost per cancers detected and treated, and cost per quality adjusted life year (QALY). 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.

Table 49. Summary of intermediate outcomes from the full model

Results	Symptomatic		Incidental		Screening		Surveillance	
	AI-assisted	Unaided	AI-assisted	Unaided	AI-assisted	Unaided	AI-assisted	Unaided
Correct detection of any lung nodules	808.0	645.5	110.7	88.4	422.5	371.6		
Correct detection of actionable nodules	481.8	333.4	58.6	42.5	223.8	178.7		
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 (<5mm)	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	22.489	22.489	3.2203	3.2204	12.345	12.345		

Superseded
see erratum

8.1.1 Symptomatic population

Deterministic results are reported in **Table 50** to **Table 52** for the symptomatic population.

Cost per correct identification of people with actionable nodules

Table 50 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 a symptomatic population. These results show that AI-assisted radiologist reading (InferRead CT Lung) is approximately £4,000 cheaper and expected to correctly identify an additional 148.4 people with actionable nodules: hence, dominating the unaided reading strategy.

Table 50. 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	4,010	333.4	-148.4	Dominated
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

Cost per cancer correctly detected and treated

Results from **Table 51** 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.

Table 51. 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
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

Cost per QALY

Results in **Table 52** show that unaided reading strategy dominates by being less costly and more effective than AI-assisted radiologist reading (InferRead CT Lung) when QALYs are taken into account.

Table 52. 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
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

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 21 shows the impact to 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.

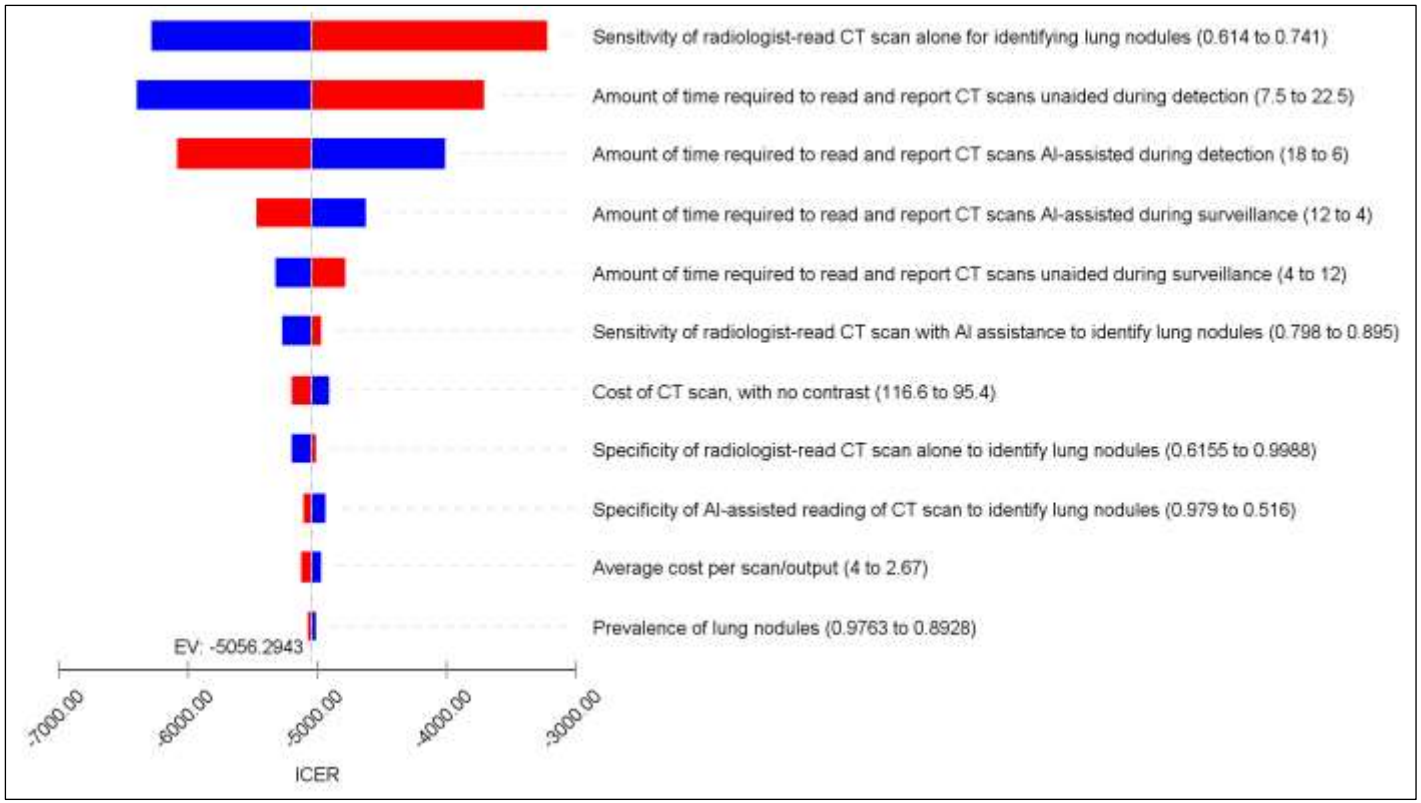


Figure 21. Tornado diagram of the impact to the cost per QALY by changing individual parameters (symptomatic population)

Scenario analyses

Table 53. 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 to 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 (sub-solid nodules) in both strategies					
Unaided radiologist reading	717,470	-	6349.5	-	-
AI-assisted radiologist reading	860,190	142,720	6320.5	-29.00	Dominated

(InferRead CT Lung)					
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
CT, computed tomography; QALY, quality adjusted life-year					

Superseded-
see erratum

8.1.2 Incidental population

Cost per correct identification of a person with actionable lung nodules

Table 54 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 a symptomatic population. These results show that AI-assisted radiologist reading (InferRead CT Lung) is approximately £4,000 cheaper and expected to correctly identify an additional 16.1, resulting in the unaided reading strategy being dominated.

Table 54. Deterministic results based on expected costs and expected cases appropriately identified (incidental population of 1,000 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	4,010	42.5	-16.1	Dominated
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

Results from **Table 55** 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.

Table 55. Deterministic results based on expected costs and expected cancer correctly detected and treated (incidental population of 1,000 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	2,430	1.844	-0.34	Dominated
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

Cost per QALY

Results in **Table 56** 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.

Table 56. 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	2,430	6573.63	2.44	996
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

Incidental population

Sensitivity analysis

Figure 22 shows the impact to the cost per QALY by 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.

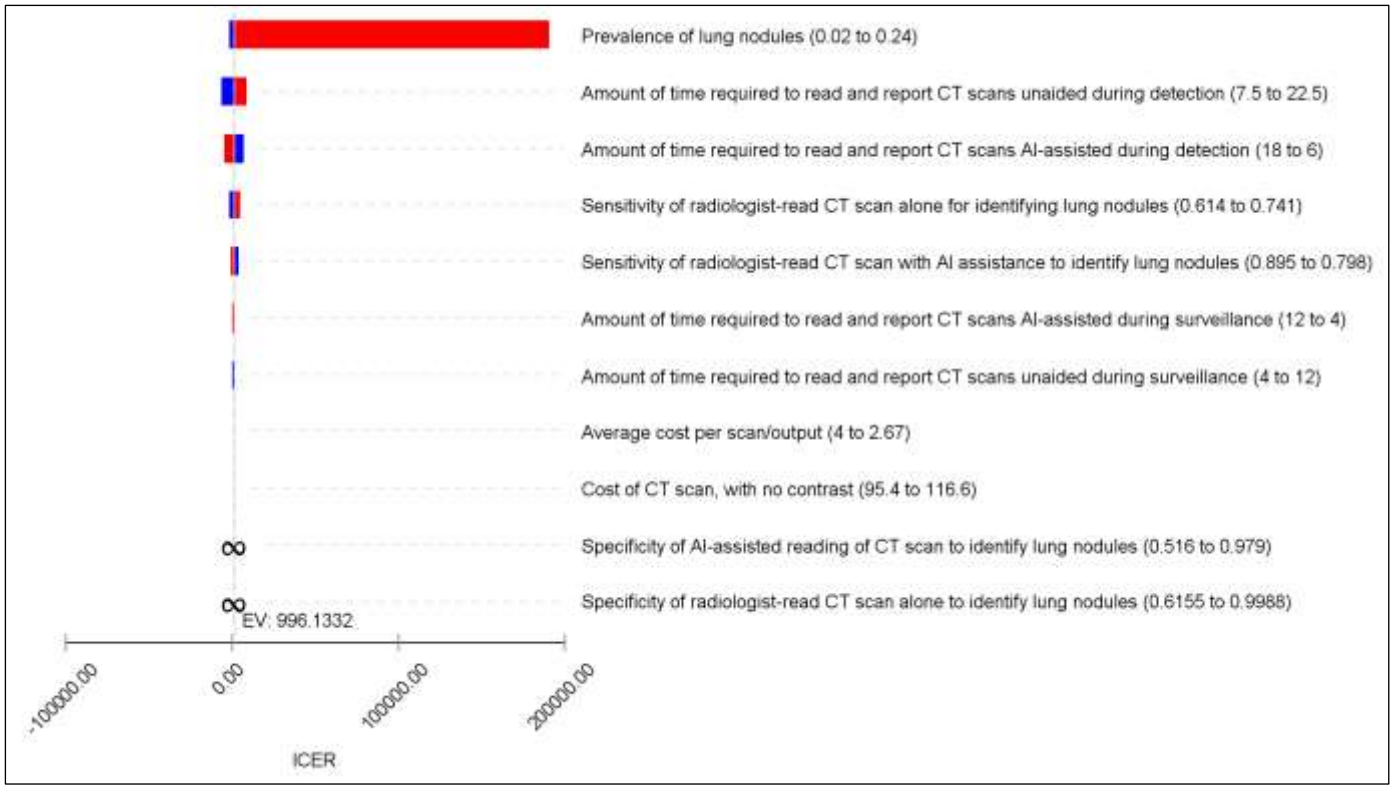


Figure 22. Tornado diagram of the impact to the cost per QALY Identified by changing individual parameters (incidental population)

Scenario analyses

Table 57. 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	2,430	6573.63	2.44	996
Prevalence of detecting any lung nodules (0.1300 to 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	229,210	5,300	6571.19	-24.43	Dominated

Superseded-
see erratum

(InferRead CT Lung)					
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	-24.43	Dominated
People with benign nodules discharged at 2-year CT surveillance (solid nodules) and 4-year CT surveillance (sub-solid 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	2,430	6582.69	-0.89	Dominated
CT, computed tomography; QALY, quality adjusted life-year					

Superseded-
see erratum

8.1.3 Screening population

Deterministic results are reported in **Table 58 t**o **Table 60** for the incidental population.

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

Table 58 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 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. Use of AI-assistance software strategy is cheaper compared to unaided reading strategy resulting in the latter being dominated.

Table 58. Deterministic results based on expected costs and expected correct identification of people with actionable nodules (screening population of 1,000 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	2,900	178.7	-45.1	Dominated
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

Cost per cancer correctly detected and treated

Results from **Table 59** show 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.

Table 59. 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
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

Cost per QALY

Results from **Table 60** 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.

Table 60. 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
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

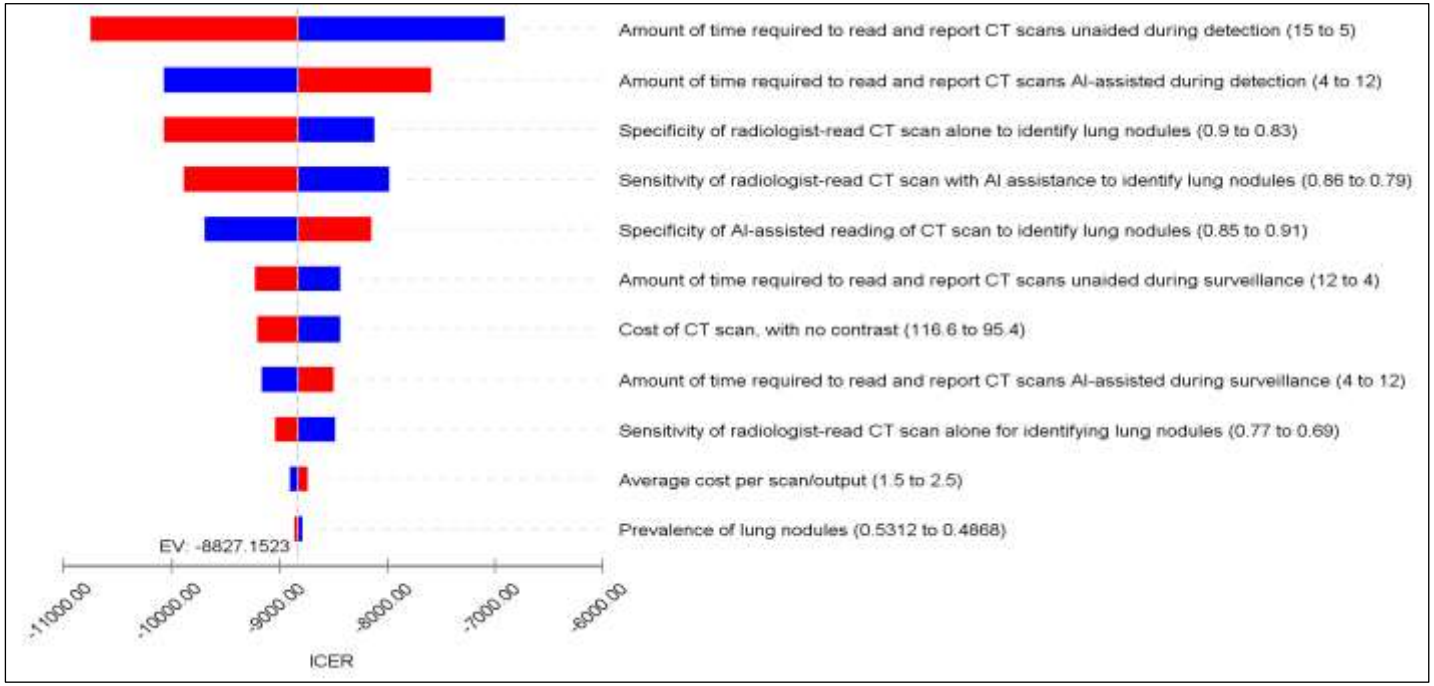


Figure 23. Tornado diagram of the impact to the cost per QALY by changing individual parameters (screening population)

Scenario analyses

Table 61. 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 to 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 (sub-solid 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
CT, computed tomography; QALY, quality adjusted life-year					

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** (section **13.9**). 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 may not have been fully captured in these analyses.

8.1.4 Surveillance population

In addition to exploring the cost-effectiveness of AI-assisted image analysis in the symptomatic, incidental and screening populations, the EAG further undertook a cost-effectiveness analysis for 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 with improved reliability in 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. 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 symptomatic or screening population at the initial scan) and assume that the same reading strategy are used at both scans.

Results in **Table 62** 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 <5mm at the initial scan. Information used to undertake these analyses were obtained from our simulation used to information 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), 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.

These results show that the AI-assisted strategy is less costly and more effective; thus, dominating the unaided strategy.

Cost per QALY

Table 62. Deterministic results based on expected costs and QALYs (screening population of 1,000 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
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

We undertook further scenario analysis where we assumed that people whose cancers were missed during surveillance would present with stage IV disease instead (**Table 63**). These results showed that the AI-assisted strategy continued to dominate unaided reading in this patient population.

Cost per QALY

Table 63. Deterministic results based on expected costs and QALYs (screening population of 1,000 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
CT, computed tomography; ICER, incremental cost-effectiveness ratio Exact results have been obtained from TreeAge but were rounded by the authors and presented.					

8.2 Discussion

8.2.1 Summary of key results

AI-assistance increases the number of lung nodules detected at first presentation. The number of extra nodules detected per 1000 persons screened are 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 are 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 it being malignant remains. However, we assume that a proportion of these additional nodules will be malignant, and therefore detected early as a result of the nodule being correctly identified. 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 5mm and 8mm undergo surveillance and are only investigated 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, though improving measurement accuracy, also introduces a systematic over-estimation of size. The way this is modelled implies that, when repeated measurements are taken to estimate VDTs, these will be systematically under-estimated. 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 from missing cancers through failing to detect a nodule.

In terms of cost per QALY, use of AI in the screening population was estimated to be cost effective, 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 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. 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 to 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 cause of AI assistance not being 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 affected results. In the incidental population there were higher QALYs overall for the unaided reader than the AI-assisted strategy, so unaided reading had a cost per QALY of £996 compared to AI assistance, indicating 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 which was unaffected by sensitivity and scenario analyses. Many of the data inputs for the screening population differed from those from 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 are 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 for the unaided readers. In the symptomatic and incidental populations there are more people experiencing these harms and costs when AI assistance is used compared to unaided readers. 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 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 suggests AI is more specific whereas the symptomatic and incidental data used suggested the unaided reader was more specific (see section 7.4.3). Whilst there was 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 through 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.

8.2.2 Generalisability of results

A key limitation is the paucity of data with major concerns regarding generalisability. For example, while our base case analysis indicates 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.[51] The results are not consistent with findings from other studies which suggested that the specificity for AI-assisted reading tends to be worse compared with unaided reading. The risk of bias and applicability concerns commonly found in studies included in our systematic review and highlighted in section **3.2** 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 incidental population. Consequently, a large number of model parameters have to be 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 generalisability to technologies (ClearRead CT and InferRead CT Lung) that have directly contributed to model parameter inputs related to test accuracy and costs. Generalisability of the findings to other technologies would be dependent on the demonstration of equivalent or more favourable evidence. However, it is worth reiterating that overall our findings are highly uncertain due to paucity of evidence and other issues explicated below.

8.2.3 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 according to 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 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 analytical 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.
- While the decision analytical model that we created is likely to require further refinement and validation and the findings are highly uncertain due to the paucity of data, it provides a useful framework that will allow further evaluation to be undertaken when more evidence emerges

While 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 highlighted below.

- 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 under unaided read to ensure equal cancer prevalence between the populations subject to different detection strategies.
- We assume that all nodules presented are 3-30mm in starting size, and come from a log-normal distribution. These cut-offs are plausible, but it is possible there may be 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.[83]
- Each patient's solid nodule growth is assume to follow a single Gompertz growth rate as reported in the literature. Whilst 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 at a ratio of 4:5. The linear growth assumption, while following the way growth is not capped and means that for some patients, their nodules may grow much faster than would occur in real life.
- There is no mortality 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 a small number of patients had a large measurement error which 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 being benign are in fact benign.
- When categorising patients at the 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 VDT) was independent of starting nodule size.

The large number of simplifying assumptions indicates that there is a high level of structural and methodological uncertainty associated with the decision analytical model which 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 heterogenous evidence to evaluate clinical 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.

9 ASSESSMENT OF FACTORS RELEVANT TO THE NHS AND OTHER PARTIES

This technology assessment focuses on evaluation of 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 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 AI software capable of detecting and analysing multiple disease conditions
- Integration of the technologies into existing picture archiving and communication system (PACS) and workflow; compatibility with existing CT scanners and workstations
- Different costs and costing structure 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 wellbeing 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

10 DISCUSSION

10.1 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 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 datasets 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 were also reviewed to provide supplementary information. The remaining three studies reported on prospective screening experiences based on the same screening pilot trial conducted in South 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 with AI. Evidence from research settings suggest 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 in 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 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 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. Paucity of data and methodological challenges mean the findings of the linked evidence approach are highly uncertain and needs to be interpreted with great caution.
- Acknowledging the above caveats, 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 reader.

10.2 Strengths and limitations of the assessment

10.2.1 Strengths

The strengths of this technology assessment include:

- Comprehensive and systematic searches of relevant literature, supplemented by requests of 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.

10.2.2 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.

10.2.2.1 Limitations of the review

- Excluded literature before 2012; but studies on AI software published before this date are unlikely to be relevant to current assessment.
- Only 7 companies (Aidence, contextflow, Infervision, JLK, MeVis, Riverain, Siemens Healthineers) submitted information and/or replied to our questions for clarification.
- 15 records were excluded on full text level as the software name was unclear, and we have received no author reply.
- MeVis: Excluded studies using the research software CIRRUS as well as studies on the AI software Visia.
- Siemens Healthineers: Excluded studies on any other technologies, e.g. syngo.
- Due to the limited evidence and heterogeneity, meta-analysis as well as subgroup analysis by ethnicity, nodule type, dose or reader speciality were not performed.
- The review did not specifically consider differences in test accuracy of different AI software because no evidence with direct comparisons between different software was identified, and 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 over-diagnosis of lung cancer was not considered in this technology assessment.

10.2.2.2 Limitations of the evidence

<u>Volume and nature of available evidence</u>

- No studies on 5/13 technologies were identified. 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. These 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 for Veolity (MeVis) and Veye Lung Nodules (Aidence). Of the eight technologies with at least one study available, only AI-RAD Companion (Siemens Healthineers), ClearRead CT (Riverain Technologies), Veolity (MeVis) and Veye Lung Nodules (Aidence) had at last two studies conducted in Western Europe or North America. These impose major limitations in the applicability of evidence to the UK setting. The number of studies available for each technologies ranges from 6 (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 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 very high level of uncertainty for test accuracy estimates related to individual technologies.
- There is 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 available evidence in a way that provides an overview of the potential impact of using an AI-derived software to support nodule detection and analysis compared to current practice (without AI assistance) rather than focusing on the performance of individual technologies, for which evidence is still immature for most of the technologies. 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 provide equal or better evidence, and as indicators for undertaking further research in many areas of major uncertainty.
- Inconsistencies in numbers and results between the journal article by Murchison (2022)[31] and the clinical evaluation report by Aidence (2020)[28] In the DAR results section, we have only reported the results by Murchison et al. (2022) as this publication was newer and published in a peer-reviewed journal.

<u>Applicability concerns, risk of bias and data inconsistency</u>

- This review focused on the identification of evidence which 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.[49]

- 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 focussed on only a separate software function like nodule detection, nodule measurement or nodule type determination.
- There were no prospective test accuracy studies of a consecutive cohorts in clinical practice. The majority of studies were small and used enriched datasets.
- In addition to study location, most studies had additional applicability concerns regarding the target population: e.g. nodule- and or cancer-enriched, undertaken retrospectively in research settings (further discussed below), and slice thickness of CT scans.
- Many studies evaluated AI algorithms as stand-alone systems rather than as an aid to radiologists - raising applicability concerns.
- Reference standard for nodule detection was usually based on majority or consensus of two or more expert chest radiologists.[84]
- 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, where readers under study conditions behave differently to how they would under routine clinical conditions.[85]
- MRMC studies were mainly performed with US or Asian radiologists with different reading experience and speciality. Consequently, study results have limited applicability to the UK context.
- Further methodological issues of the included studies include the focus on single centres studies, reporting per-nodule sensitivity and number of FP 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;

in contrast with clinical practice, readers in included studies usually had no access to relevant prior CT images.

- Inconsistencies in numbers and results between the journal article by Murchison (2022)[31] and the clinical evaluation report by Aidence (2020)[28] In the DAR results section, we have only reported the results by Murchison et al. (2022) as this publication was newer and published in a peer-reviewed journal.

10.3 Uncertainties

Uncertainties associated with high risk of bias and applicability concerns of available evidence:

- All the issues related to risk of bias and applicability presented in Section 3.2 and highlighted above increase the uncertainty in the estimated test accuracy, clinical effectiveness and cost-effectiveness of the technologies being evaluated in this technology assessment.
- Per person vs 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 guideline recommends the management of lung nodule based on the one with the largest size (risk dominating nodule), in practice other nodules with sizes or features that are not safe to ignore may also be measured and analyses during the same reading session and be followed up during surveillance. Consequently, per person analysis of clinical management decision would reflect the real impact of AI assistance on clinical practice more closely. Nevertheless results from per person analyses or per nodule analysis based on the risk dominating nodule are infrequently presented, and in some cases we have to use data from per nodule analyses to inform our model. The impact of this is uncertain and is difficult to estimate using sensitivity analysis.

Uncertainties associated with the long causal chain modelled using linked evidence approach

- One of the main purported benefits for AI-assisted image analysis is the improved precision and accuracy in the measurement of nodule size (diameter) and volume, and by extension in the estimation of nodule growth. Evidence on the impact of AI assistance on these was reviewed and presented in **Sections 3.3.3** (diameter measurement), **3.3.4** (volume measurement) and **3.4** (nodule growth monitoring) respectively. While there is good evidence on improved consistency in nodule measurement between different readers when assisted by AI, evidence on measurement accuracy (e.g. whether measurements assisted by

AI systematically over- or under-estimate the sizes/volumes of the nodules) is less clear. Furthermore, while separate evidence on measurement precision and accuracy was reported in some studies, evidence on 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 associated with other methodological challenges

- Difficulties in defining reference standard for nodule detection.[84]

10.4 Other relevant factors

AI has increasingly been applied to directly predict 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 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 its impact related to multiple conditions is beyond the scope of this DAR. Such 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 evaluation of the impact of timing and frequency of low-dose CT scans on cancer detection, it was not designed for assessing 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 **Appendix 10** (**section 13.10**).

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. Current evidence suggests AI-assisted reading tends to reduce specificity and results in nodules being classified into higher risk categories based on current clinical guidelines although it may not always be the case. The reported performance of AI assisted reading varies substantially among published studies, possibly attributed to heterogeneous study population, reader experience, speciality 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 analysis of CT scan images assisted by 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 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, EAG's assessment suggested that for the symptomatic and incidental populations AI-assisted CT image analysis dominates the 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 the unaided reader. 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 estimates of improved test specificity for AI. 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 are likely to be important factors in addition to accuracy in nodule detection in driving cost-effectiveness.

11.1 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 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 and cost-effectiveness evidence is lacking. Potential adoption of these technologies will need to consider uncertainties associated with quality, quantitative and applicability of available evidence of individual technologies in addition to their functionality, relevant costs and costing structure. Undertaking further research to generate evidence may be needed to inform decisions on 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 algorithm. Ongoing audit of potential impact of these updates on test accuracy and service provision may be desirable.

11.2 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 of high risk of bias and have multiple applicability concerns for the UK settings. No prospective studies evaluating intermediate clinical process and downstream clinical outcomes were identified. Further prospective studies of use of software derived from AI algorithm to aid chest CT image analysis that adopts per-person analysis for estimating test accuracy, incorporates clinical process and outcome measures, and that 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.

Further methodological research that may be required 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).

.

12 REFERENCES

1. Larici AR, Farchione A, Franchi P, Ciliberto M, Cicchetti G, Calandriello L*, et al.* Lung nodules: size still matters. *Eur Respir Rev* 2017;**26**(146):170025. http://dx.doi.org/10.1183/16000617.0025-2017

2. 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**(2):584-600. http://dx.doi.org/10.1148/radiol.2017162894

3. 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**(12):1332-41. http://dx.doi.org/10.1016/s1470-2045(14)70389-4

4. Cancer Research UK. *Lung cancer incidence statistics*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence (Accessed 4 October 2022).

5. National Disease Registration Service. *Staging data in England*. Public Health England; 2020. URL: https://www.cancerdata.nhs.uk/stage_at_diagnosis (Accessed 5 October 2022).

6. NHS. *NHS long term plan*. URL: https://www.longtermplan.nhs.uk/ (Accessed 5 October 2022).

7. National Institute for Health and Care Excellence. *Suspected cancer: recognition and referral. NICE guideline [NG12]*. NICE; 2015. URL: https://www.nice.org.uk/guidance/ng12 (Accessed 5 October 2022).

8. National Institute for Health and Care Excellence. *Lung cancer: diagnosis and management. NICE guideline [NG122]*. NICE; 2019. URL: https://www.nice.org.uk/guidance/ng122 (Accessed 5 October 2022).

9. UK National Screening Committee. *Adult screening programme: lung cancer*. Gov.uk. URL: https://view-health-screening-recommendations.service.gov.uk/lung-cancer/ (Accessed 6 October 2022).

10. NHS England. *Evaluation of the Targeted Lung Health Check programme*. URL: https://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).

11. 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-ii54. http://dx.doi.org/10.1136/thoraxjnl-2015-207168

12. 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**(10):910-9. http://dx.doi.org/10.1056/NEJMoa1214726

13. British Thoracic Society. *PN risk calculator*. URL: https://www.brit-thoracic.org.uk/quality-improvement/guidelines/pulmonary-nodules/pn-risk-calculator/ (Accessed 5 October 2022).

14. Herder GJ, van Tinteren H, Golding RP, Kostense PJ, Comans EF, Smit EF*, et al.* Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. *Chest* 2005;**128**(4):2490-6. http://dx.doi.org/10.1378/chest.128.4.2490

15. British Thoracic Society. *BTS guidelines for the investigation and management of pulmonary nodules*. URL: https://www.brit-thoracic.org.uk/quality-improvement/guidelines/pulmonary-nodules/ (Accessed 1 December 2021).

16. American College of Radiology. *Lung CT Screening Reporting & Data System (Lung-RADS). Version 1.1*. 2019. URL: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads (Accessed 24 October 2022).

17. 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: https://www.england.nhs.uk/wp-content/uploads/2019/02/B1646-standard-protocol-targeted-lung-health-checks-programme-v2.pdf (Accessed 5 December 2022).

18. British Society of Thoracic Imaging, The Royal College of Radiologists. *Considerations to ensure optimum roll-out of targeted lung cancer screening over the next five years*. The Royal College of Radiologists; 2020. URL: https://www.rcr.ac.uk/posts/considerations-ensure-optimum-roll-out-targeted-lung-cancer-screening-over-next-five-years (Accessed 5 October 2022).

19. 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).

20. National Institute for Health and Clinical Excellence. *Diagnostics Assessment Programme manual*. Manchester: NICE; 2011. URL: https://www.nice.org.uk/media/default/about/what-we-do/nice-guidance/nice-diagnostics-guidance/diagnostics-assessment-programme-manual.pdf (Accessed 5 October 2022).

21. 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).

22. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB*, et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**(8):529-36. http://dx.doi.org/10.7326/0003-4819-155-8-201110180-00009

23. 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**(11):1592-9. http://dx.doi.org/10.7326/m21-2234

24. 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**(1):293. http://dx.doi.org/10.1186/s12874-020-01179-5

25. 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**(10):6891-9. http://dx.doi.org/10.1007/s00330-022-08824-1

26. 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. http://dx.doi.org/10.26044/ecr2019/C-3685

27. 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. http://dx.doi.org/10.26044/ecr2019/C-3686

28. Wakkie J, Doorn L. *Clinical Evaluation Report: Veye Lung Nodules (and Veye Chest 2.15). LN-CER-003. Document version 3.1.* Aidence [unpublished; company submission]; 2020.

29. 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* 2022:Jul 3. http://dx.doi.org/10.1007/s00330-022-08973-3

30. 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. http://dx.doi.org/10.1016/j.lungcan.2022.01.002

31. Murchison JT, Ritchie G, Senyszak D, Nijwening JH, van Veenendaal G, Wakkie J, *et al.* 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**(5):e0266799. http://dx.doi.org/10.1371/journal.pone.0266799

32. 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. http://dx.doi.org/10.1016/j.ejro.2022.100435

33. 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**(Suppl 2):A131. http://dx.doi.org/10.1136/thorax-2019-BTSabstracts2019.221

34. Ahn Y, Lee SM, Noh HN, Kim W, Choe J, Do KH, *et al.* 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**(1):202-10. http://dx.doi.org/10.1148/radiol.2021202803

35. 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**(12):4465-74. http://dx.doi.org/10.1007/s00330-016-4317-3

36. 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**(6):1174-80. http://dx.doi.org/10.1016/j.ejrad.2016.03.027

37. Garzelli L, Goo JM, Ahn SY, Chae KJ, Park CM, Jung J, *et al.* 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. http://dx.doi.org/10.1016/j.ejrad.2018.01.016

38. 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**(7):988-94. http://dx.doi.org/10.1016/j.acra.2020.01.014

39. Park S, Park G, Lee SM, Kim W, Park H, Jung K, *et al.* Deep learning-based differentiation of invasive adenocarcinomas from preinvasive or minimally invasive lesions among pulmonary subsolid nodules. *Eur Radiol* 2021;**31**(8):6239-47. http://dx.doi.org/10.1007/s00330-020-07620-z

40. 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. http://dx.doi.org/10.1016/j.compbiomed.2022.105538

41. 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**(3):58. http://dx.doi.org/10.1007/s10916-019-1180-1

42. 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**(Suppl 2):s108-s17. http://dx.doi.org/10.1016/j.acra.2021.02.007

43. 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**(8):3629-42. http://dx.doi.org/10.21037/qims-20-1314

44. 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. http://dx.doi.org/10.1016/j.ebiom.2019.05.040

45. 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**(3):154-61. http://dx.doi.org/10.1097/RTI.0000000000000613

46. 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**(1):55. http://dx.doi.org/10.1186/s12916-021-01928-3

47. Rueckel 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**(6):2486-98. http://dx.doi.org/10.21037/qims-20-1037

48. 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**(9):7202-12. http://dx.doi.org/10.1007/s00330-021-07827-8

49. 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**(1):475-85. http://dx.doi.org/10.1007/s00330-020-07151-7

50. Hwang EJ, Goo JM, Kim HY, Yoon SH, Jin GY, Yi J*, et al.* 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**(5):2845-55. http://dx.doi.org/10.1007/s00330-020-07424-1

51. 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**(8):626. http://dx.doi.org/10.1016/j.crad.2021.04.006

52. 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**(3):480-8. http://dx.doi.org/10.2214/AJR.17.18718

53. 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. http://dx.doi.org/10.1016/j.ejrad.2018.02.020

54. 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**(4):1134-43. http://dx.doi.org/10.21037/qims-20-630

55. Takaishi T, Ozawa Y, Bando Y, Yamamoto A, Okochi S, Suzuki H*, et al.* 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**(2):159-64. http://dx.doi.org/10.1007/s11604-020-01043-y

56. 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**(8):2211. http://dx.doi.org/10.3390/cancers12082211

57. 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**(11):1052-61. http://dx.doi.org/10.1007/s11604-020-01009-0

58. 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**(3):e180084. http://dx.doi.org/10.1148/ryai.2019180084

59. 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**(3):327-34. http://dx.doi.org/10.2174/1573405617666210806125953

60. 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**(8):3266-74. http://dx.doi.org/10.1007/s00330-016-4716-5

61. 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**(5):2124-33. http://dx.doi.org/10.1007/s00330-017-5171-7

62. 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**(5):e200160. http://dx.doi.org/10.1148/rycan.2021200160

63. 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. http://dx.doi.org/10.1016/j.ejrad.2021.109526

64. 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**(11):838-45. http://dx.doi.org/10.1016/j.crad.2021.07.012

65. Park S, Park H, Lee SM, Ahn Y, Kim W, Jung K*, et al.* 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. http://dx.doi.org/10.1007/s00330-021-08202-3

66. 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**(1):228-43. http://dx.doi.org/10.1148/radiol.2017161659

67. Singh R, Nitiwarangkul C, Shepard J-AO, 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.

68. 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**(1):304-17. http://dx.doi.org/10.1148/radiol.12120628

69. 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**(1):162. http://dx.doi.org/10.1186/s12916-017-0924-3

70. 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**(3):367-72. http://dx.doi.org/10.1007/s10198-013-0471-6

71. 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**(36):iii-iv, ix-xi, 1-158. http://dx.doi.org/10.3310/hta8360

72. Bajre MK, Pennington M, Woznitza N, Beardmore C, Radhakrishnan M, Harris R*, et al.* Expanding the role of radiographers in reporting suspected lung cancer: a cost-effectiveness analysis using a decision tree model. *Radiography (Lond)* 2017;**23**(4):273-8. http://dx.doi.org/10.1016/j.radi.2017.07.011

73. 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**(5):741-51. http://dx.doi.org/https://dx.doi.org/10.1016/j.jacr.2020.11.014

74. Cancer Research UK. *Lung cancer statistics*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer (Accessed 5 December 2022).

75. 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**(10):908-12. http://dx.doi.org/10.1136/thoraxjnl-2020-214703

76. Jones K, Burns A. *Unit costs of health and social care 2021.* Personal Social Services Research Unit, University of Kent, Canterbury; 2021. http://dx.doi.org/10.22024/UniKent/01.02.92342

77. Steele JD, Buell P. Asymptomatic solitary pulmonary nodules: host survival, tumor size, and growth rate. *J Thorac Cardiovasc Surg* 1973;**65**(1):140-51.

78. 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**(2):161-70. http://dx.doi.org/10.1136/thoraxjnl-2015-207140

79. Birtwistle M, Earnshaw A. *Saving lives, averting costs: an analysis of the financial implications of achieving earlier diagnosis of colorectal, lung and ovarian cancer.* Incisive Health, Cancer Research UK; 2014.

80. Jacobs DR, Jr., Adachi H, Mulder I, Kromhout D, Menotti A, Nissinen A*, et al.* Cigarette smoking and mortality risk: twenty-five-year follow-up of the Seven Countries Study. *Arch Intern Med* 1999;**159**(7):733-40. http://dx.doi.org/10.1001/archinte.159.7.733

81. Sutton AJ, Sagoo GS, Jackson L, Fisher M, Hamilton-Fairley G, Murray A*, et al.* 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**(9):e0237492. http://dx.doi.org/10.1371/journal.pone.0237492

82. 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**(4):1-174. http://dx.doi.org/10.3310/hta16040

83. Davies L, Petitti DB, Martin L, Woo M, Lin JS. Defining, estimating, and communicating overdiagnosis in cancer screening. *Ann Intern Med* 2018;**169**(1):36-43. http://dx.doi.org/10.7326/m18-0694

84. Armato 3rd SG, 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**(1):28-38. http://dx.doi.org/10.1016/j.acra.2008.05.022

85. 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**(1):47-53. http://dx.doi.org/10.1148/radiol.2491072025

86. Riverain Technologies. *National Institute for Health and Care Excellence Diagnostics Assessment Programme: artificial intelligence for analysing chest CT images (provisional title): request for information.* Riverain Technologies [unpublished; company submission]; 2021.

87. Contextflow. *National Institute for Health and Care Excellence Diagnostics Assessment Programme: artificial intelligence for analysing chest CT images (provisional title): request for information.* Contextflow [unpublished; company submission]; 2021.

88. *Clinical validation of InferRead Lung CT.AI. NCT04119960*. Bethesda, MD: ClinicalTrials.gov, U.S. National Library of Medicine; 2019. URL: https://clinicaltrials.gov/ct2/show/NCT04119960 (Accessed 15 November 2022).

89. *International Lung Screen Trial (ILST). NCT02871856*. Bethesda, MD: ClinicalTrials.gov, U.S. National Library of Medicine; 2016. URL: https://clinicaltrials.gov/ct2/show/NCT02871856 (Accessed 16 November 2022).

90. Lim KP, Marshall H, Tammemägi M, Brims F, McWilliams A, Stone E*, et al.* Protocol and rationale for the International Lung Screening Trial. *Ann Am Thorac Soc* 2020;**17**(4):503-12. http://dx.doi.org/10.1513/AnnalsATS.201902-102OC

91. *Clinical performance evaluation of Veye Lung Nodules (CPEVLN). NCT04792632*. Bethesda, MD: ClinicalTrials.gov, U.S. National Library of Medicine; 2021. URL: https://clinicaltrials.gov/ct2/show/NCT04792632 (Accessed 15 November 2022).

92. *AI in Health and Care Award: scoping plan.* Aidence [unpublished; company submission]; 2021.

93. *A multi-center, retrospective pivotal trial to evaluate the efficacy of artificial intelligence-based pulmonary nodule detection software 'VUNO Med – Lung CAD' in thoracic CT. KCT0005065*. Seoul, Korea: CRIS: Clinical Research Information Service; 2020. URL: http://cris.nih.go.kr/cris/en/search/search_result_st01.jsp?seq=16420 (Accessed 15 November 2022).

94. 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**(12):e754-e66. http://dx.doi.org/10.1016/s1470-2045(17)30861-6

95. 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**(3):135-51. http://dx.doi.org/10.1038/s41571-020-00432-6

96. 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**(3):340-9. http://dx.doi.org/10.1016/j.lungcan.2008.01.001

97. 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**(1):9139. http://dx.doi.org/10.1038/s41598-021-88328-y

98. 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**(1):318-26. http://dx.doi.org/10.1002/acm2.13142

99. 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**(1):139-47. http://dx.doi.org/10.1007/s00330-012-2570-7

13 APPENDICES

13.1 Appendix 1: Literature search strategies: 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	17/01/22	2,740
Embase	17/01/22	3,495
Cochrane Library (CENTRAL and Cochrane Database of Systematic reviews)	17/01/22	131 (all from CENTRAL; 0 results from CDSR)
Science Citation Index and Conference Proceedings – Science (Web of Science)	19/01/22	3,210
HTA database (CRD)	19/01/22	1
INAHTA HTA database	19/01/22	3
medRxiv	19/01/22	7
clinicaltrials.gov	19/01/22	17
WHO ICTRP	19/01/22	22
Total number of records retrieved: 9,626 **Duplicates removed (EndNote): 3,296** **Final number for screening: 6,330**		
Other sources		
Source	**Date searched**	**Documents retrieved**
National Institute for Health and Care Excellence (NICE) website	24/01/22	3
Canadian Agency for Drugs and Technologies in Health (CADTH) website	24/01/22	7
ISPOR conference presentations	25/01/22	0
HTAi annual meetings	25/01/22	1
SPIE proceedings	27/01/22	14
IEEE Engineering in Medicine & Biology Society annual conference	27/01/22	1
European Congress of Radiology	31/01/22	47
Radiological Society of North America annual meetings	01/02/22	55
FDA devices databases	14/02/22	5
Device / manufacturer websites	15-16/02/22	15 documents, plus 1 link to video presentation
Forwards citation tracking: Science Citation Index (Web of Science) and Google Scholar	26/05/22 & 30/05/22	44
Total: 192		

Search strategies used:

MEDLINE ALL
Date searched: 17/01/22
Ovid MEDLINE(R) ALL <1946 to January 14, 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/ 134273
2 ai.kf,tw.34062
3 ((artificial or machine or deep) adj5 (intelligence or learning or reasoning)).kf,tw.89902
4 exp Neural Networks, Computer/ 42235
5 (neural network* or convolutional or CNN or CNNs).kf,tw. 73835
6 exp Diagnosis, Computer-Assisted/ 85513
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. 31141
9 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 322906
10 exp Lung Neoplasms/di, dg or Solitary Pulmonary Nodule/di, dg 56493
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. 274199
12 ((pulmonary or lung) adj2 lesion*).kf,tw. 14782
13 10 or 11 or 12 302352
14 Tomography, X-Ray Computed/ or exp Tomography, Spiral Computed/ 418962
15 (comput* adj2 tomograph*).kf,tw. 348023
16 (CT or LDCT).kf,tw. 388825
17 (CAT adj2 (scan* or x-ray* or xray*)).kf,tw. 1342
18 Mass Screening/ 111594
19 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw. 4813
20 "Early Detection of Cancer"/ 31774
21 14 or 15 or 16 or 17 or 18 or 19 or 20 893125
22 9 and 13 and 21 2767
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 27 2867
29 exp animals/ not humans/ 4943529
30 28 not 29 2851
31 limit 30 to english language 2740

The artificial intelligence search terms (lines 1-4 & 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 doi:10.1136/bmj.n1872 (see online supplementary appendix 1)

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. Available from: https://www.nice.org.uk/guidance/indevelopment/gid-dg10041/documents (accessed 9 November 2021)

Embase
Date searched: 17/01/22
Embase <1974 to 2022 January 14>

1 exp artificial intelligence/ or exp machine learning/ 304838
2 ai.kf,tw.45921
3 ((artificial or machine or deep) adj5 (intelligence or learning or reasoning)).kf,tw.105922
4 (neural network* or convolutional or CNN or CNNs).kf,tw. 89201
5 computer assisted diagnosis/ 40877
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. 38837
8 1 or 2 or 3 or 4 or 5 or 6 or 7 420312
9 exp lung cancer/di or lung nodule/di 46922
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. 392765
11 ((pulmonary or lung) adj2 lesion*).kf,tw. 21058
12 9 or 10 or 11 420629
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/ 931594
14 (comput* adj2 tomograph*).kf,tw. 445065
15 (CT or LDCT).kf,tw. 664348
16 (CAT adj2 (scan* or x-ray* or xray*)).kf,tw. 2036
17 mass screening/ or cancer screening/ 142872
18 screening/ 184110
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 20 1643282
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 28 3692
30 (exp animal/ or exp animal experiment/) not (exp human/ or exp human experiment/ or conference abstract.pt.)4770834
31 29 not 30 3673
32 limit 31 to english language 3495

Cochrane Library (via www.cochranelibrary.com)
Date searched: 17/01/22
Cochrane Central Register of Controlled Trials, Issue 12 of 12, December 2021
Cochrane Database of Systematic Reviews, Issue 1 of 12, January 2022

ID Search Hits
#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,kw 4506
#3 ((artificial OR machine OR deep) NEAR/5 (intelligence OR learning OR reasoning)):ti,ab,kw 2857
#4 [mh "Neural Networks, Computer"] 148
#5 ((neural NEXT network*) OR convolutional OR CNN OR CNNs):ti,ab,kw 1479
#6 [mh "Diagnosis, Computer-Assisted"] 1931
#7 (("computer aided" OR "computer assisted") NEAR/1 (diagnosis OR detection)):ti,ab,kw 1001
#8 (("support vector" NEXT machine*) OR (random NEXT forest*) OR "black box learning"):ti,ab,kw 776
#9 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 10964
#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,kw 25143
#12 ((pulmonary OR lung) NEAR/2 lesion*):ti,ab,kw 533
#13 #10 OR #11 OR #12 25426
#14 [mh ^"Tomography, X-Ray Computed"] OR [mh "Tomography, Spiral Computed"] 4555
#15 (comput* NEAR/2 tomograph*):ti,ab,kw 20680
#16 (CT OR LDCT):ti,ab,kw 81013
#17 (CAT NEAR/2 (scan* OR x-ray* OR xray*)):ti,ab,kw 34
#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,kw 758
#20 [mh ^"Early Detection of Cancer"] 1384
#21 #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20 96454

#22 #9 AND #13 AND #21 125
#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, Trials 131

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(2):195-207. http://dx.doi.org/10.5195/jmla.2020.834

Science Citation Index and Conference Proceedings - Science (via Web of Science)
Date searched: 19/01/2022
SCI-EXPANDED: 1970-present
CPCI-S: 1990-present

23 (((#17) OR #18) OR #19) OR #22 and English (Languages) 3,210
22 (#20 OR #21) AND #7 AND #16 216
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*)) 2,633
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*)) 3,964
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 #16 3,085
16 #10 or #11 or #12 or #13 or #14 or #15 655,436
15 TS=("Early Detection of Cancer") 2,106
14 TS=((lung OR lungs OR pulmon*) NEAR/3 (nodule* OR cancer* OR tumor* OR tumour*) NEAR/3 screen*) 6,299
13 TS=("Mass Screening") 5,559
12 TS=(CAT NEAR/2 (scan* OR x-ray* OR xray*)) 1,067

11 TS=(CT OR LDCT) 455,518
10 TS=(comput* NEAR/2 tomograph*) 361,422
9 #7 OR #8 380,001
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 #5 901,467
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(2):195-207. http://dx.doi.org/10.5195/jmla.2020.834

HTA Database (via CRD https://www.crd.york.ac.uk/CRDWeb/)
Date searched: 19/01/22

1 MeSH DESCRIPTOR Artificial Intelligence EXPLODE ALL TREES 290
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 TREES 108
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 HTA 148
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 TREES 1151
11 MeSH DESCRIPTOR Solitary Pulmonary Nodule EXPLODE ALL TREES 27
12 (#9 OR #10 OR #11) IN HTA 341
13 MeSH DESCRIPTOR Tomography, X-Ray Computed 896
14 MeSH DESCRIPTOR Tomography, Spiral Computed EXPLODE ALL TREES 75
15 (comput* ADJ2 tomograph*) 1395
16 (CT OR LDCT) 1231
17 (CAT ADJ2 (scan* OR x-ray* OR xray*)) 6
18 MeSH DESCRIPTOR Mass Screening 2103
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 TREES 277

21 (#13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20) IN HTA 953
22 #8 AND #12 AND #21 1

International HTA database (via INAHTA https://database.inahta.org/)
Date searched: 19/01/22

21 #20 AND #14 AND #8 3
20 #19 OR #16 OR #15 417
19 #18 AND #17 383
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 #9 2443
13 tomograph* OR radiograph* OR CT OR x-ray* OR xray* OR MRI OR PET 813
12 screening 1234
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 #1 189
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 CNNs 5
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: 19/1/22

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: 19/01/22

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: 19/01/22

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 records for 1 trials 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 https://www.nice.org.uk/

Date searched: 24/01/22

Browsed: NICE Guidance > Conditions and diseases > Cancer > Lung cancer:
https://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: https://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 cancer 51 results, of which 1 potentially relevant, already identified above

nodule 3 results, of which 1 potentially relevant, already identified above

Searched published guidance: https://www.nice.org.uk/guidance/published?sp=on
No filters.

artificial intelligence	3 results, of which 1 potentially relevant, already identified above
machine learning	0 results
deep learning	0 results
ai	1 result, of which 0 relevant
neural network	0 results

Browsed guidance In consultation: https://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 https://www.cadth.ca/

Date searched: 24/01/22

Search screen: https://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

ai 20 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 https://www.ispor.org/heor-resources/presentations-database/search
Date searched: 25/01/22

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: 25/01/22

HTAi 2021 Virtual (Manchester). Full program 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 https://www.spiedigitallibrary.org/)

Date searched: 26/01/22
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: 27/01/22
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 https://www.myesr.org/congress/about-ecr/past-congresses)

Date searched: 31/1/22

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: https://www.rsna.org/annual-meeting/future-and-past-meetings)

Date searched: 01/02/22

RSNA 2020 meeting program available at: https://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 program available at:
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 program 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: 14/02/22
Search interfaces:

- Premarket Approval (PMA) database, 'Device' field
  https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma.cfm

- 510(k) Premarket Notification database, 'Device Name' field https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm
- Device Classification Under Section 513(f)(2)(De Novo) database, 'device name' field https://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: 15-16/02/22

AI-Rad Companion Chest CT / Siemens Healthineers
https://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. Browsed:
https://www.corelinesoft.com/aview-lcs-2/aview-lcs-plus/
https://www.corelinesoft.com/aview-lcs-2/
https://www.corelinesoft.com/newsroom-eng/
0 documents to download

ClearRead CT / Riverain Technologies
https://www.riveraintech.com/clearread-ai-solutions/clearread-ct/ 1 reference on page, already identified via database searches

https://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. Browsed:
https://global.infervision.com/product/19/
https://global.infervision.com/news/5/
https://global.infervision.com/news/6/
0 documents to download

JLD-01K / JLK Inc
https://www.jlkgroup.com/en/medihub.html 0 documents to download

Lung AI / Arterys
https://www.arterys.com/clinicalapp/lungapp - references 'Arterys Lung AI Nodule Detection study - University of California, San Diego' – unable to find this via Google search
https://www.arterys.com/clinical-evidence - nothing on Lung AI; 0 documents to download

ung Nodule AI / Fujifilm. Browsed:
https://www.fujifilm.com/uk/en/healthcare/healthcare-it
https://synapse.fujifilm.eu/ai-lab/#(grid|filter)=.radiology;
0 documents to download

qCT-Lung / Qure.ai. Browsed:
https://qure.ai/product/qct-lung/
https://qure.ai/evidences/
0 documents to download

SenseCare-Lung Pro / Sensetime. Browsed:
https://www.sensetime.com/en/product-detail?categoryId=32629
https://www.sensetime.com/en/news-index
0 documents to download

MeVis / Veolity. Browsed
https://www.veolity.com/
https://www.veolity.com/news-events
0 documents to download

Aidence / Veye Lung Nodules
https://www.aidence.com/veye-lung-nodules/
https://www.aidence.com/development-clinical-validation/ **2 conference posters and 1 unpublished manuscript downloaded**

https://www.aidence.com/clinical-research/ 5 articles/reports, of which 1 CQC report not identified via previous searches; **1 document downloaded**
https://www.aidence.com/resources/
https://www.aidence.com/articles/ **6 articles downloaded** (including 3 from an external site, 2 of which are in Dutch)

VUNO Med-LungCT AI / VUNO
https://www.vuno.co/en/lung
https://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 26/05/22**	**Google Scholar, searched 30/05/22**
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	
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.

13.2 Appendix 2: Table of excluded studies with rationale

Table 64. Publications excluded after review of full-text articles – Electronic database searches (n=150)

Reference	Main reason for exclusion
Excluded on population: >10% oncologic patients (n=10)	
1. Ahn Y, Lee SM, Noh HN, et al. 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(1):202-10. doi: https://dx.doi.org/10.1148/radiol.2021202803	>10% oncologic patients
2. Martini K, Bluthgen C, Eberhard M, et al. Impact of Vessel Suppressed-CT on Diagnostic Accuracy in Detection of Pulmonary Metastasis and Reading Time. Acad Radiol 2021;28(7):988-94. doi: https://dx.doi.org/10.1016/j.acra.2020.01.014	>10% oncologic patients
3. Meybaum C, Graff M, Fallenberg EM, et al. Contribution of CAD to the Sensitivity for Detecting Lung Metastases on Thin-Section CT - A Prospective Study with Surgical and Histopathological Correlation. ROFO Fortschr Geb Rontgenstr Nuklearmed 2020;192(1):65-73. doi: https://dx.doi.org/10.1055/a-0977-3453	>10% oncologic patients
4. Park S, Lee SM, Kim W, et al. Computer-aided Detection of Subsolid Nodules at Chest CT: Improved Performance with Deep Learning-based CT Section Thickness Reduction. Radiology 2021;299(1):211-19. doi: https://dx.doi.org/10.1148/radiol.2021203387	>10% oncologic patients
5. Shaffer K. Deep Learning and Lung Cancer: AI to Extract Information Hidden in Routine CT Scans. Radiology 2020;296(1):225-26. doi: https://dx.doi.org/10.1148/radiol.2020201366	>10% oncologic patients
6. Vassallo L, Traverso A, Agnello M, et al. A cloud-based computer-aided detection system improves identification of lung nodules on computed tomography scans of patients with extra-thoracic malignancies. Eur Radiol 2019;29(1):144-52. doi: https://dx.doi.org/10.1007/s00330-018-5528-6	>10% oncologic patients
7. Wagner AK, Hapich A, Psychogios MN, et al. Computer-Aided Detection of Pulmonary Nodules in Computed Tomography Using ClearReadCT. J Med Syst 2019;43(3):58. doi: https://dx.doi.org/10.1007/s10916-019-1180-1	>10% oncologic patients
8. Weikert T, Akinci D'Antonoli T, Bremerich J, et al. Evaluation of an AI-Powered Lung Nodule Algorithm for Detection and 3D Segmentation of Primary Lung Tumors. Contrast Media Mol Imaging 2019;2019:1545747. doi: https://dx.doi.org/10.1155/2019/1545747	>10% oncologic patients
9. Yacoub B, Kabakus I, Schoepf J, et al. Performance of an Artificial Intelligence-Based Platform Against Clinical Radiology Reports for the Evaluation of Non-contrast Chest CT. J Thorac Imaging 2021;36(6):W123. doi: http://dx.doi.org/10.1097/RTI.0000000000000619	>10% oncologic patients
10. Yacoub B, Kabakus IM, Schoepf UJ, et al. Performance of an Artificial Intelligence-Based Platform Against Clinical Radiology Reports for the Evaluation of Noncontrast Chest CT. Acad Radiol 2021;10:10. doi: https://dx.doi.org/10.1016/j.acra.2021.02.007	>10% oncologic patients
Excluded on population: Chest phantoms (n=3)	

Reference	Main reason for exclusion
11. Ebner L, Roos JE, Christensen JD, et al. Maximum-Intensity-Projection and Computer-Aided-Detection Algorithms as Stand-Alone Reader Devices in Lung Cancer Screening Using Different Dose Levels and Reconstruction Kernels. AJR Am J Roentgenol 2016;207(2):282-8. doi: https://dx.doi.org/10.2214/AJR.15.15588	Chest phantom
12. Peters AA, Decasper A, Munz J, et al. Performance of an AI based CAD system in solid lung nodule detection on chest phantom radiographs compared to radiology residents and fellow radiologists. J 2021;13(5):2728-37. doi: https://dx.doi.org/10.21037/jtd-20-3522	Chest phantom
13. Schwyzer M, Messerli M, Eberhard M, et al. Impact of dose reduction and iterative reconstruction algorithm on the detectability of pulmonary nodules by artificial intelligence. Diagn Interv Imaging 2022;03:03. doi: https://dx.doi.org/10.1016/j.diii.2021.12.002	Chest phantom
Excluded on population: Other image type (n=6)	
14. Lee JH, Sun HY, Park S, et al. Performance of a Deep Learning Algorithm Compared with Radiologic Interpretation for Lung Cancer Detection on Chest Radiographs in a Health Screening Population. Radiology 2020;297(3):687-96. doi: https://dx.doi.org/10.1148/radiol.2020201240	Other image type
15. Rajagopalan K, Babu S. The detection of lung cancer using massive artificial neural network based on soft tissue technique. BMC Med Inf Decis Mak 2020;20(1):282. doi: https://dx.doi.org/10.1186/s12911-020-01220-z	Other image type
16. Schultheiss M, Schmette P, Bodden J, et al. Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance. Sci 2021;11(1):15857. doi: https://dx.doi.org/10.1038/s41598-021-94750-z	Other image type
17. Ueda D, Yamamoto A, Shimazaki A, et al. Artificial intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: a retrospective clinical validation study. BMC Cancer 2021;21(1):1120. doi: https://dx.doi.org/10.1186/s12885-021-08847-9	Other image type
18. Yamada Y, Shiomi E, Hashimoto M, et al. Value of a Computer-aided Detection System Based on Chest Tomosynthesis Imaging for the Detection of Pulmonary Nodules. Radiology 2018;287(1):333-39. doi: https://dx.doi.org/10.1148/radiol.2017170405	Other image type
19. Yoo H, Lee SH, Arru CD, et al. AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset. Eur Radiol 2021;31(12):9664-74. doi: https://dx.doi.org/10.1007/s00330-021-08074-7	Other image type
Excluded on technology: Language processing tool (n=1)	
20. Hunter B, Reis S, Campbell D, et al. Development of a Structured Query Language and Natural Language Processing Algorithm to Identify Lung Nodules in a Cancer Centre. Front Med (Lausanne) 2021;8:748168. doi: https://dx.doi.org/10.3389/fmed.2021.748168	Language processing tool
Excluded on technology: Malignancy risk prediction (n=12)	
21. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954-61. doi: https://dx.doi.org/10.1038/s41591-019-0447-x	Malignancy risk prediction
22. Heuvelmans MA, Oudkerk M. Deep learning to stratify lung nodules on annual follow-up CT. Lancet Digit Health 2019;1(7):e324-e25. doi: https://dx.doi.org/10.1016/S2589-7500(19)30156-6	Malignancy risk prediction

Reference	Main reason for exclusion
23. Huang P, Park S, Yan R, et al. Added Value of Computer-aided CT Image Features for Early Lung Cancer Diagnosis with Small Pulmonary Nodules: A Matched Case-Control Study. Radiology 2018;286(1):286-95. doi: https://dx.doi.org/10.1148/radiol.2017162725	Malignancy risk prediction
24. Jacobs C, Setio AAA, Scholten ET, et al. Deep Learning for Lung Cancer Detection on Screening CT Scans: Results of a Large-Scale Public Competition and an Observer Study with 11 Radiologists. Radiol Artif Intell 2021;3(6):e210027. doi: https://dx.doi.org/10.1148/ryai.2021210027	Malignancy risk prediction
25. Lassau N, Bousaid I, Chouzenoux E, et al. Three artificial intelligence data challenges based on CT and MRI. Diagn Interv Imaging 2020;101(12):783-88. doi: https://dx.doi.org/10.1016/j.diii.2020.03.006	Malignancy risk prediction
26. Pickup L, Arteta C, Declerck J, et al. P1.11-02 Acceleration of Lung Cancer Diagnosis: Utility Study for AI-Based Stratification of Pulmonary Nodules. Journal of Thoracic Oncology 2019;14(10 Supplement):S515. doi: http://dx.doi.org/10.1016/j.jtho.2019.08.1075	Malignancy risk prediction
27. Tsakok MT, Mashar M, Pickup L, et al. The utility of a convolutional neural network (CNN) model score for cancer risk in indeterminate small solid pulmonary nodules, compared to clinical practice according to British Thoracic Society guidelines. Eur J Radiol 2021;137:109553. doi: https://dx.doi.org/10.1016/j.ejrad.2021.109553	Malignancy risk prediction
28. Wels M, Lades F, Muehlberg A, et al. General Purpose Radiomics for Multi-Modal Clinical Research. Medical Imaging 2019: Computer-Aided Diagnosis 2019;10950 doi: 10.1117/12.2511856	Malignancy risk prediction
29. Xu T, Huang C, Liu Y, et al. Artificial intelligence based on deep learning for differential diagnosis between benign and malignant pulmonary nodules: A real-world, multicenter, diagnostic study. Journal of Clinical Oncology 2020;38(15) doi: https://dx.doi.org/10.1200/JCO.2020.38.15-suppl.9037	Malignancy risk prediction
30. Zeng JY, Ye HH, Yang SX, et al. Clinical application of a novel computer-aided detection system based on three-dimensional CT images on pulmonary nodule. Int J Clin Exp Med 2015;8(9):16077-82.	Malignancy risk prediction
31. Zhao L, Bai C, Zhu Y. Preliminary study on diagnostic value of artificial intelligence in early-stage lung cancer. American Journal of Respiratory and Critical Care Medicine 2020;201(1)	Malignancy risk prediction
32. Zhao L, Bai C-X, Zhu Y. Diagnostic value of artificial intelligence in early-stage lung cancer. Chin Med J 2020;133(4):503-04. doi: http://dx.doi.org/10.1097/CM9.0000000000000634	Malignancy risk prediction
Excluded on technology: Software not commercially available (n=37)	
33. Akter O, Moni MA, Islam MM, et al. Lung cancer detection using enhanced segmentation accuracy. Appl Intell 2021;51(6):3391-404. doi: 10.1007/s10489-020-02046-y	Software not commercially available
34. Aresta G, Jacobs C, Araujo T, et al. iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. Sci 2019;9(1):11591. doi: https://dx.doi.org/10.1038/s41598-019-48004-8	Software not commercially available
35. Cui X, Zheng S, Heuvelmans MA, et al. Performance of a deep learning-based lung nodule detection system as an alternative reader in a Chinese lung cancer screening program. Eur J Radiol 2022;146:110068. doi: https://dx.doi.org/10.1016/j.ejrad.2021.110068	Software not commercially available

Reference	Main reason for exclusion
36. Huang W, Xue Y, Wu Y. A CAD system for pulmonary nodule prediction based on deep three-dimensional convolutional neural networks and ensemble learning. PLoS ONE 2019;14(7):e0219369. doi: https://dx.doi.org/10.1371/journal.pone.0219369	Software not commercially available
37. Iwasawa T, Matsumoto S, Aoki T, et al. A comparison of axial versus coronal image viewing in computer-aided detection of lung nodules on CT. Jpn J Radiol 2015;33(2):76-83. doi: https://dx.doi.org/10.1007/s11604-014-0383-0	Software not commercially available
38. Jacobs C, van Rikxoort EM, Scholten ET, et al. Solid, part-solid, or non-solid?: classification of pulmonary nodules in low-dose chest computed tomography by a computer-aided diagnosis system. Invest Radiol 2015;50(3):168-73. doi: https://dx.doi.org/10.1097/RLI.0000000000000121	Software not commercially available
39. Jacobs C, van Rikxoort EM, Twellmann T, et al. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. Med Image Anal 2014;18(2):374-84. doi: https://dx.doi.org/10.1016/j.media.2013.12.001	Software not commercially available
40. Kuo C-FJ, Barman J, Hsieh CW, et al. Fast fully automatic detection, classification and 3D reconstruction of pulmonary nodules in CT images by local image feature analysis. Biomedical Signal Processing and Control 2021;68:102790. doi: http://dx.doi.org/10.1016/j.bspc.2021.102790	Software not commercially available
41. Lassen BC, Jacobs C, Kuhnigk JM, et al. Robust semi-automatic segmentation of pulmonary subsolid nodules in chest computed tomography scans. Phys Med Biol 2015;60(3):1307-23. doi: https://dx.doi.org/10.1088/0031-9155/60/3/1307	Software not commercially available
42. Liang F, Li C, Fu X. Evaluation of the Effectiveness of Artificial Intelligence Chest CT Lung Nodule Detection Based on Deep Learning. J 2021;2021:9971325. doi: https://dx.doi.org/10.1155/2021/9971325	Software not commercially available
43. Liang J, Ye G, Guo J, et al. Reducing False-Positives in Lung Nodules Detection Using Balanced Datasets. Front 2021;9:671070. doi: https://dx.doi.org/10.3389/fpubh.2021.671070	Software not commercially available
44. Liu C, Hu SC, Wang C, et al. Automatic detection of pulmonary nodules on CT images with YOLOv3: development and evaluation using simulated and patient data. Quant 2020;10(10):1917-29. doi: https://dx.doi.org/10.21037/qims-19-883	Software not commercially available
45. Liu JB, Liu LH, He W, et al. Computer-aided detection of pulmonary nodules in computed tomography images: Effect on observer performance. Journal of Medical Imaging and Health Informatics 2017;7(6):1205-11. doi: http://dx.doi.org/10.1166/jmihi.2017.2201	Software not commercially available
46. Liu JK, Jiang HY, Gao MD, et al. An Assisted Diagnosis System for Detection of Early Pulmonary Nodule in Computed Tomography Images. J Med Syst 2017;41(2):30. doi: https://dx.doi.org/10.1007/s10916-016-0669-0	Software not commercially available
47. Long C, Hackett T, Yang D, et al. Automatic detection and diagnosis of pulmonary nodule using deep convolutional neural network. Canadian Journal of Respiratory, Critical Care, and Sleep Medicine 2019;3(Supplement 1):11. doi: http://dx.doi.org/10.1080/24745332.2019.1623590	Software not commercially available
48. Masood A, Yang P, Sheng B, et al. Cloud-Based Automated Clinical Decision Support System for Detection and Diagnosis of Lung Cancer in Chest CT. IEEE J Transl Eng Health Med 2020;8:4300113. doi: https://dx.doi.org/10.1109/JTEHM.2019.2955458	Software not commercially available
49. Nguyen CC, Tran GS, Nguyen VT, et al. Pulmonary Nodule Detection Based on Faster R-CNN With Adaptive Anchor Box. IEEE Access 2021;9:154740-51. doi: 10.1109/ACCESS.2021.3128942	Software not commercially available

Reference	Main reason for exclusion
50. Nomura Y, Higaki T, Fujita M, et al. Effects of Iterative Reconstruction Algorithms on Computer-assisted Detection (CAD) Software for Lung Nodules in Ultra-low-dose CT for Lung Cancer Screening. Acad Radiol 2017;24(2):124-30. doi: https://dx.doi.org/10.1016/j.acra.2016.09.023	Software not commercially available
51. Paing MP, Hamamoto K, Tungjitkusolmun S, et al. Automatic Detection and Staging of Lung Tumors using Locational Features and Double-Staged Classifications. Appl Sci-Basel 2019;9(11) doi: 10.3390/app9112329	Software not commercially available
52. Pereira FR, De Andrade JMC, Escuissato DL, et al. Classifier Ensemble Based on Computed Tomography Attenuation Patterns for Computer-Aided Detection System. IEEE Access 2021;9:123134-45. doi: 10.1109/ACCESS.2021.3109860	Software not commercially available
53. Qiu Z, Wu Q, Wang S, et al. Development of a deep learning-based method to diagnose pulmonary ground glass nodules by sequential computed tomography imaging. Thorac Cancer 2022;06:06. doi: https://dx.doi.org/10.1111/1759-7714.14305	Software not commercially available
54. Savic M, Ma Y, Ramponi G, et al. Lung Nodule Segmentation with a Region-Based Fast Marching Method. Sensors (Basel) 2021;21(5):09. doi: https://dx.doi.org/10.3390/s21051908	Software not commercially available
55. Seito AAA, Jacobs C, Ciompi F, et al. Computer-Aided Detection of Lung Cancer: Combining Pulmonary Nodule Detection Systems with a Tumor Risk Prediction Model. Medical Imaging 2015: Computer-Aided Diagnosis 2015;9414 doi: 10.1117/12.2080955	Software not commercially available
56. Silva M, Capretti G, Sverzellati N, et al. Non-solid and part-solid nodules: Comparison between visual and computer aided detection. J Thorac Imaging 2017;32(4):W19. doi: http://dx.doi.org/10.1097/RTI.0000000000000288	Software not commercially available
57. Silva M, Schaefer-Prokop CM, Jacobs C, et al. Detection of Subsolid Nodules in Lung Cancer Screening: Complementary Sensitivity of Visual Reading and Computer-Aided Diagnosis. Invest Radiol 2018;53(8):441-49. doi: https://dx.doi.org/10.1097/RLI.0000000000000464	Software not commercially available
58. Song J, Huang SC, Kelly B, et al. Automatic lung nodule segmentation and intra-nodular heterogeneity image generation. IEEE j 2021;15:15. doi: https://dx.doi.org/10.1109/JBHI.2021.3135647	Software not commercially available
59. Tammemagi M, Ritchie AJ, Atkar-Khattra S, et al. Predicting Malignancy Risk of Screen-Detected Lung Nodules-Mean Diameter or Volume. J Thorac Oncol 2019;14(2):203-11. doi: https://dx.doi.org/10.1016/j.jtho.2018.10.006	Software not commercially available
60. Tan JR, Cheong EHT, Chan LP, et al. Implementation of an Artificial Intelligence-Based Double Read System in Capturing Pulmonary Nodule Discrepancy in CT Studies. Curr Probl Diagn Radiol 2021;50(2):119-22. doi: https://dx.doi.org/10.1067/j.cpradiol.2020.07.006	Software not commercially available
61. Terasawa T, Aoki T, Murakami S, et al. Detection of lung carcinoma with predominant ground glass opacity on CT using temporal subtraction method. Eur Radiol 2018;28(4):1594-99. doi: https://dx.doi.org/10.1007/s00330-017-5085-4	Software not commercially available
62. Wang YQ, Yue SH, Li Q, et al. Research on Technologies of Computer Aided Diagnosis for Solitary Pulmonary Nodule Based on CT Images. 2019 Ieee International Instrumentation and Measurement Technology Conference (I2mtc) 2019:724-28.	Software not commercially available
63. Woo M, Devane AM, Lowe SC, et al. Deep learning for semi-automated unidirectional measurement of lung tumor size in CT. Cancer Imaging 2021;21(1):43. doi: https://dx.doi.org/10.1186/s40644-021-00413-7	Software not commercially available
64. Xu YM, Zhang T, Xu H, et al. Deep Learning in CT Images: Automated Pulmonary Nodule Detection for Subsequent Management Using Convolutional Neural Network. Cancer Manag Res 2020;12:2979-92. doi: https://dx.doi.org/10.2147/CMAR.S239927	Software not commercially available

Reference	Main reason for exclusion
65. Yen A, Pfeffer Y, Blumenfeld A, et al. Use of a Dual Artificial Intelligence Platform to Detect Unreported Lung Nodules. J Comput Assist Tomogr 2021;45(2):318-22. doi: https://dx.doi.org/10.1097/RCT.0000000000001118	Software not commercially available
66. Young S, Lo P, Kim G, et al. The effect of radiation dose reduction on computer-aided detection (CAD) performance in a low-dose lung cancer screening population. Med Phys 2017;44(4):1337-46. doi: https://dx.doi.org/10.1002/mp.12128	Software not commercially available
67. Yu H, Li J, Zhang L, et al. Design of lung nodules segmentation and recognition algorithm based on deep learning. BMC Bioinformatics 2021;22(Suppl 5):314. doi: https://dx.doi.org/10.1186/s12859-021-04234-0	Software not commercially available
68. Zhang QH, Kong XJ. Design of Automatic Lung Nodule Detection System Based on Multi-Scene Deep Learning Framework. IEEE Access 2020;8:90380-89. doi: 10.1109/ACCESS.2020.2993872	Software not commercially available
69. Zuo W, Zhou F, He Y. An Embedded Multi-branch 3D Convolution Neural Network for False Positive Reduction in Lung Nodule Detection. J Digit Imaging 2020;33(4):846-57. doi: https://dx.doi.org/10.1007/s10278-020-00326-0	Software not commercially available
Excluded on technology: Software not specified as relevant by NICE (n=32)	
70. Benzakoun J, Bommart S, Coste J, et al. Computer-aided diagnosis (CAD) of subsolid nodules: Evaluation of a commercial CAD system. Eur J Radiol 2016;85(10):1728-34. doi: https://dx.doi.org/10.1016/j.ejrad.2016.07.011	Software not specified as relevant by NICE
71. Brown M, Browning P, Wahi-Anwar MW, et al. Integration of Chest CT CAD into the Clinical Workflow and Impact on Radiologist Efficiency. Acad Radiol 2019;26(5):626-31. doi: https://dx.doi.org/10.1016/j.acra.2018.07.006	Software not specified as relevant by NICE
72. Buls N, Watte N, Nieboer K, et al. Performance of an artificial intelligence tool with real-time clinical workflow integration - Detection of intracranial hemorrhage and pulmonary embolism. Phys Med 2021;83:154-60. doi: http://dx.doi.org/10.1016/j.ejmp.2021.03.015	Software not specified as relevant by NICE
73. Chen K, Lai YC, Vanniarajan B, et al. Clinical impact of a deep learning system for automated detection of missed pulmonary nodules on routine body computed tomography including the chest region. Eur Radiol 2022;09:09. doi: https://dx.doi.org/10.1007/s00330-021-08412-9	Software not specified as relevant by NICE
74. Chen L, Gu D, Chen Y, et al. An artificial-intelligence lung imaging analysis system (ALIAS) for population-based nodule computing in CT scans. Comput Med Imaging Graph 2021;89:101899. doi: https://dx.doi.org/10.1016/j.compmedimag.2021.101899	Software not specified as relevant by NICE
75. Cho J, Kim J, Lee KJ, et al. Incidence Lung Cancer after a Negative CT Screening in the National Lung Screening Trial: Deep Learning-Based Detection of Missed Lung Cancers. J 2020;9(12):02. doi: https://dx.doi.org/10.3390/jcm9123908	Software not specified as relevant by NICE
76. Ctri. To Study how accurately of Predible Lung detects Lung Nodules. https://trialsearchwhoint/Trial2aspx?TrialID=CTRI/2019/07/020120 2019	Software not specified as relevant by NICE
77. Cui X, Ye Z, Zheng S, et al. Validation of a deep learning-based computer-aided system for lung nodule detection in a Chinese lung cancer screening program. Eur Respir J 2020;56(Supplement 64) doi: https://dx.doi.org/10.1183/13993003.congress-2020.4168	Software not on NICE list
78. Cui X, Ye Z, Zheng S, et al. P42.02 Evaluating the Feasibility of a Deep Learning-Based Computer-Aided Detection System for Lung Nodule Detection in a Lung Cancer Screening Program. Journal of Thoracic Oncology 2021;16(3 Supplement):S477-S78. doi: https://dx.doi.org/10.1016/j.jtho.2021.01.826	Software not specified as relevant by NICE

Reference	Main reason for exclusion
79. Den Harder AM, Willemink MJ, van Hamersvelt RW, et al. Effect of radiation dose reduction and iterative reconstruction on computer-aided detection of pulmonary nodules: Intra-individual comparison. Eur J Radiol 2016;85(2):346-51. doi: https://dx.doi.org/10.1016/j.ejrad.2015.12.003	Software not specified as relevant by NICE
80. Guo X, Li Y, Yang C, et al. Deep Learning-Based Computed Tomography Imaging to Diagnose the Lung Nodule and Treatment Effect of Radiofrequency Ablation. J 2021;2021:6556266. doi: https://dx.doi.org/10.1155/2021/6556266	Software not specified as relevant by NICE
81. Han D, Heuvelmans M, Rook M, et al. Evaluation of a novel deep learning-based classifier for perifissural nodules. Eur Radiol 2021;31(6):4023-30. doi: https://dx.doi.org/10.1007/s00330-020-07509-x	Software not specified as relevant by NICE
82. Jacobs C, van Ginneken B. Google's lung cancer AI: a promising tool that needs further validation. Nat Rev Clin Oncol 2019;16(9):532-33. doi: https://dx.doi.org/10.1038/s41571-019-0248-7	Software not specified as relevant by NICE
83. Jeon KN, Goo JM, Lee CH, et al. Computer-aided nodule detection and volumetry to reduce variability between radiologists in the interpretation of lung nodules at low-dose screening computed tomography. Invest Radiol 2012;47(8):457-61. doi: https://dx.doi.org/10.1097/RLI.0b013e318250a5aa	Software not specified as relevant by NICE
84. Jprn U. Japanese Multi-Center Study for Utility of 3D Computer-Aided Detection System at Lung Cancer Screening with Low-dose CT Protocol. https://trialsearchwhoint/Trial2aspx?TrialID=JPRN-UMIN000030415 2018	Software not specified as relevant by NICE
85. Jurkovic IA, Papanikolaou N, Stathakis S, et al. Objective assessment of the quality and accuracy of deformable image registration. J 2020;45(3):156-67. doi: https://dx.doi.org/10.4103/jmp.JMP_47_19	Software not specified as relevant by NICE
86. Li L, Liu Z, Huang H, et al. Evaluating the performance of a deep learning-based computer-aided diagnosis (DL-CAD) system for detecting and characterizing lung nodules: Comparison with the performance of double reading by radiologists. Thorac Cancer 2019;10(2):183-92. doi: https://dx.doi.org/10.1111/1759-7714.12931	Software not specified as relevant by NICE
87. Li X, Guo F, Zhou Z, et al. Performance of Deep-learning-based Artificial Intelligence on Detection of Pulmonary Nodules in Chest CT. Zhongguo fei ai za zhi [Chinese journal of lung cancer] 2019;22(6):336-40. doi: 10.3779/j.issn.1009-3419.2019.06.02	Software not specified as relevant by NICE
88. Liu Z, Li L, Li T, et al. Does a Deep Learning-Based Computer-Assisted Diagnosis System Outperform Conventional Double Reading by Radiologists in Distinguishing Benign and Malignant Lung Nodules? Front 2020;10:545862. doi: http://dx.doi.org/10.3389/fonc.2020.545862	Software not specified as relevant by NICE
89. Matsumoto S, Ohno Y, Aoki T, et al. Computer-aided detection of lung nodules on multidetector CT in concurrent-reader and second-reader modes: a comparative study. Eur J Radiol 2013;82(8):1332-7. doi: https://dx.doi.org/10.1016/j.ejrad.2013.02.005	Software not specified as relevant by NICE
90. Ohno Y, Aoyagi K, Chen Q, et al. Comparison of computer-aided detection (CADe) capability for pulmonary nodules among standard-, reduced- and ultra-low-dose CTs with and without hybrid type iterative reconstruction technique. Eur J Radiol 2018;100:49-57. doi: https://dx.doi.org/10.1016/j.ejrad.2018.01.010	Software not specified as relevant by NICE
91. Ohno Y, Aoyagi K, Takenaka D, et al. Machine learning for lung CT texture analysis: Improvement of inter-observer agreement for radiological finding classification in patients with pulmonary diseases. Eur J Radiol 2021;134:109410. doi: https://dx.doi.org/10.1016/j.ejrad.2020.109410	Software not specified as relevant by NICE

Reference	Main reason for exclusion
92. Ohno Y, Aoyagi K, Yaguchi A, et al. 3D CADv system with and without CNN: Comparison of nodule component measurement accuracy and differentiation in routine clinical practice data. International Journal of Computer Assisted Radiology and Surgery 2020;15(1 Supplement):S114-S15. doi: http://dx.doi.org/10.1007/s11548-020-02171-6	Software not specified as relevant by NICE
93. Ohno Y, Aoyagi K, Yaguchi A, et al. Differentiation of Benign from Malignant Pulmonary Nodules by Using a Convolutional Neural Network to Determine Volume Change at Chest CT. Radiology 2020;296(2):432-43. doi: https://dx.doi.org/10.1148/radiol.2020191740	Software not specified as relevant by NICE
94. Ohri B, Smith D, Melville P, et al. Use of "artificial Intelligence" to Aid Pulmonary Nodule Assessment. Respirology 2020;25:177. doi: https://dx.doi.org/10.1111/resp.13778	Software not specified as relevant by NICE
95. Prakashini K, Babu S, Rajgopal KV, et al. Role of Computer Aided Diagnosis (CAD) in the detection of pulmonary nodules on 64 row multi detector computed tomography. Lung India 2016;33(4):391-7. doi: https://dx.doi.org/10.4103/0970-2113.184872	Software not specified as relevant by NICE
96. Qi LL, Wu BT, Tang W, et al. Long-term follow-up of persistent pulmonary pure ground glass nodules with deep learning-assisted nodule segmentation. Eur Radiol 2020;30(2):744-55. doi: https://dx.doi.org/10.1007/s00330-019-06344-z	Software not specified as relevant by NICE
97. Wang YW, Wang JW, Yang SX, et al. Proposing a deep learning-based method for improving the diagnostic certainty of pulmonary nodules in CT scan of chest. Eur Radiol 2021;31(11):8160-67. doi: https://dx.doi.org/10.1007/s00330-021-07919-5	Software not specified as relevant by NICE
98. Wu N, Li X, Luo X. P62.10 AI-Based Three-Dimension Reconstruction for Pulmonary Nodules -New Auxiliary Exploration for Thoracic Surgery. Journal of Thoracic Oncology 2021;16(10 Supplement):S1181. doi: http://dx.doi.org/10.1016/j.jtho.2021.08.655	Software not specified as relevant by NICE
99. Yanagawa M, Honda O, Kikuyama A, et al. Pulmonary nodules: effect of adaptive statistical iterative reconstruction (ASIR) technique on performance of a computer-aided detection (CAD) system-comparison of performance between different-dose CT scans. Eur J Radiol 2012;81(10):2877-86. doi: https://dx.doi.org/10.1016/j.ejrad.2011.09.011	Software not specified as relevant by NICE
100. Yang D, Bai C, Hu J, et al. Deep convolutional neutral networks based artificial intelligence system for pulmonary nodule detection and diagnosis in United States and Chinese dataset. American Journal of Respiratory and Critical Care Medicine 2018;197(MeetingAbstracts)	Software not specified as relevant by NICE
101. Yuan R, Mayo J, Streit I, et al. MA10.06 Randomized Clinical Trial with Computer Assisted Diagnosis (CAD) Versus Radiologist as First Reader of Lung Screening LDCT. Journal of Thoracic Oncology 2019;14(10 Supplement):S287-S88. doi: http://dx.doi.org/10.1016/j.jtho.2019.08.578	Software not specified as relevant by NICE
Excluded on technology: Manufacturer eligible but other software (n=13)	
102. Azour L, Moore WH, O'Donnell T, et al. Inter-Reader Variability of Volumetric Subsolid Pulmonary Nodule Radiomic Features. Acad Radiol 2021;17:17. doi: https://dx.doi.org/10.1016/j.acra.2021.01.026	Manufacturer eligible but other software
103. Bogoni L, Ko JP, Alpert J, et al. Impact of a computer-aided detection (CAD) system integrated into a picture archiving and communication system (PACS) on reader sensitivity and efficiency for the detection of lung nodules in thoracic CT exams. J Digit Imaging 2012;25(6):771-81. doi: https://dx.doi.org/10.1007/s10278-012-9496-0	Manufacturer eligible but other software

Reference	Main reason for exclusion
104. Godoy MC, Kim TJ, White CS, et al. Benefit of computer-aided detection analysis for the detection of subsolid and solid lung nodules on thin- and thick-section CT. AJR Am J Roentgenol 2013;200(1):74-83. doi: https://dx.doi.org/10.2214/AJR.11.7532	Manufacturer eligible but other software
105. Heuvelmans M, Oudkerk M, Zhao YR, et al. Comparison of three software systems for semi-automatic volumetry of pulmonary nodules on baseline and follow-up CT examinations. Acta Radiol 2014;55(6):691-98. doi: http://dx.doi.org/10.1177/0284185113508177	Manufacturer eligible but other software
106. Jacobs C, van Rikxoort EM, Murphy K, et al. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. Eur Radiol 2016;26(7):2139-47. doi: https://dx.doi.org/10.1007/s00330-015-4030-7	Manufacturer eligible but other software
107. Larici AR, Amato M, Ordonez P, et al. Detection of noncalcified pulmonary nodules on low-dose MDCT: comparison of the sensitivity of two CAD systems by using a double reference standard. Radiol Med (Torino) 2012;117(6):953-67.	Manufacturer eligible but other software
108. Liang M, Tang W, Xu DM, et al. Low-Dose CT Screening for Lung Cancer: Computer-aided Detection of Missed Lung Cancers. Radiology 2016;281(1):279-88. doi: https://dx.doi.org/10.1148/radiol.2016150063	Manufacturer eligible but other software
109. Messerli M, Kluckert T, Knitel M, et al. Computer-aided detection (CAD) of solid pulmonary nodules in chest x-ray equivalent ultralow dose chest CT - first in-vivo results at dose levels of 0.13mSv. Eur J Radiol 2016;85(12):2217-24. doi: https://dx.doi.org/10.1016/j.ejrad.2016.10.006	Manufacturer eligible but other software
110. Mozaffary A, Trabzonlu TA, Lombardi P, et al. Integration of fully automated computer-aided pulmonary nodule detection into CT pulmonary angiography studies in the emergency department: effect on workflow and diagnostic accuracy. Emerg 2019;26(6):609-14. doi: https://dx.doi.org/10.1007/s10140-019-01707-x	Manufacturer eligible but other software
111. Nair A, Gartland N, Barton B, et al. Comparing the performance of trained radiographers against experienced radiologists in the UK lung cancer screening (UKLS) trial. Br J Radiol 2016;89(1066):20160301. doi: https://dx.doi.org/10.1259/bjr.20160301	Manufacturer eligible but other software
112. Nair A, Screaton NJ, Holemans JA, et al. The impact of trained radiographers as concurrent readers on performance and reading time of experienced radiologists in the UK Lung Cancer Screening (UKLS) trial. Eur Radiol 2018;28(1):226-34. doi: https://dx.doi.org/10.1007/s00330-017-4903-z	Manufacturer eligible but other software
113. Takahashi EA, Koo CW, White DB, et al. Prospective Pilot Evaluation of Radiologists and Computer-aided Pulmonary Nodule Detection on Ultra-low-Dose CT With Tin Filtration. J Thorac Imaging 2018;33(6):396-401. doi: https://dx.doi.org/10.1097/RTI.0000000000000348	Manufacturer eligible but other software
114. Zhao Y, de Bock GH, Vliegenthart R, et al. Performance of computer-aided detection of pulmonary nodules in low-dose CT: comparison with double reading by nodule volume. Eur Radiol 2012;22(10):2076-84. doi: https://dx.doi.org/10.1007/s00330-012-2437-y	Manufacturer eligible but other software
Excluded on technology: Software name unclear, no author reply (n=15)	
115. Arteta C, Novotny P, Santos C, et al. Automatic Nodule Size Measurements Can Improve Prediction Accuracy Within a Brock Risk Model. Journal of Thoracic Oncology 2018;13(10 Supplement):S429. doi: http://dx.doi.org/10.1016/j.jtho.2018.08.490	Software name unclear; no author reply received
116. Brown MS, Kim HJ, Lo P, et al. Automated tumor size assessment: Consistency of computer measurements with an expert panel. Journal of Clinical Oncology 2013;31(15 SUPPL. 1)	Software name unclear; no author reply received

Reference		Main reason for exclusion
117.	Brown MS, Lo P, Barnoy E, et al. Clinically usable computer-aided detection (CAD) system for lung cancer screening with CT. American Journal of Respiratory and Critical Care Medicine 2013;187(MeetingAbstracts)	Software name unclear; no author reply received
118.	Gu X, Xie W, Fang Q, et al. The effect of pulmonary vessel suppression on computerized detection of nodules in chest CT scans. Med Phys 2020;47(10):4917-27. doi: https://dx.doi.org/10.1002/mp.14401	Software name unclear; no author reply received
119.	Gu XM, Chai YL, Weiyang X, et al. Effect of CAD system with a vessel suppression function on clinical lung nodule detection in chest CT scans. Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment 2021;11599 doi: 10.1117/12.2582059	Software name unclear; no author reply received
120.	Gu XM, Xie WY, Fang QM, et al. Lung vessel suppression and its effect on nodule detection in chest CT scans. Medical Imaging 2020: Computer-Aided Diagnosis 2020;11314 doi: 10.1117/12.2549405	Software name unclear; no author reply received
121.	Lieman-Sifry J, Brouha S, Weihe E, et al. Deep learning-based cad may improve detection of pulmonary nodules while preserving a low false-positive rate. J Thorac Imaging 2019;34(4):W61. doi: http://dx.doi.org/10.1097/RTI.0000000000000421	Software name unclear; no author reply received
122.	Liu Y, Luo H, Qing H, et al. Screening baseline characteristics of early lung cancer on low-dose computed tomography with computer-aided detection in a Chinese population. Cancer epidemiol 2019;62:101567. doi: https://dx.doi.org/10.1016/j.canep.2019.101567	Software name unclear; no author reply received
123.	Miki S, Nomura Y, Hayashi N, et al. Prospective Study of Spatial Distribution of Missed Lung Nodules by Readers in CT Lung Screening Using Computer-assisted Detection. Acad Radiol 2021;28(5):647-54. doi: https://dx.doi.org/10.1016/j.acra.2020.03.015	Software name unclear; no author reply received
124.	Ohno Y, Seki S, Yoshikawa T, et al. Convolutional Neural Network for 3D CADv Systems: Utility for Differentiation of Malignant from Benign Pulmonary Nodules. International Journal of Computer Assisted Radiology and Surgery 2019;14(Supplement 1):S67. doi: http://dx.doi.org/10.1007/s11548-019-01969-3	Software name unclear; no author reply received
125.	Setio AA, Jacobs C, Gelderblom J, et al. Automatic detection of large pulmonary solid nodules in thoracic CT images. Med Phys 2015;42(10):5642-53. doi: https://dx.doi.org/10.1118/1.4929562	Software name unclear; no author reply received
126.	Tokunaga S, Hazeki N, Tamura D, et al. Computer-aided detection (CAD) as concurrent vs. second reader for lung nodules on CT in a Japanese multicenter study: Evaluation of reading time and observer performance in radiologists and pulmonologists. Chest 2013;144(4 MEETING ABSTRACT) doi: http://dx.doi.org/10.1378/chest.1703410	Software name unclear; no author reply received
127.	Werner S, Gast R, Horger M, et al. Accuracy and Reproducibility of a Software Prototype for Semi-Automated Computer-Aided Volumetry of the solid and subsolid Components of part-solid Pulmonary Nodules. RoFo Fortschritte auf dem Gebiet der Rontgenstrahlen und der Bildgebenden Verfahren 2021 doi: http://dx.doi.org/10.1055/a-1656-9834	Software name unclear; no author reply received
128.	Zheng S, Cui X, Vonder M, et al. Deep learning-based pulmonary nodule detection: Effect of slab thickness in maximum intensity projections at the nodule candidate detection stage. Comput Methods Programs Biomed 2020;196:105620. doi: https://dx.doi.org/10.1016/j.cmpb.2020.105620	Software name unclear; no author reply received
129.	Zheng S, Cui X, Ye Z, et al. P42.06 Automatic Lung Nodule Detection by a Deep Learning-Based CAD System: The Value of Slab Thickness in the Maximum Intensity Projection Technique. Journal of Thoracic Oncology 2021;16(3 Supplement):S479-S80. doi: https://dx.doi.org/10.1016/j.jtho.2021.01.830	Software name unclear; no author reply received

Reference	Main reason for exclusion
Excluded on outcomes – Clinical trial register, no outcomes yet (n=4)	
130. fdeeeKCT0005065. A multi-center, retrospective pivotal trial to evaluate the efficacy of artificial intelligence-based pulmonary nodule detection software 'VUNO Med – Lung CAD' in thoracic CT: http://cris.nih.go.kr/cris/en/search/search_result_st01.jsp?seq=16420, 2020.	Clinical trial register, no outcomes yet
131. Institute VP, University S, Technologies R. Evaluation of Computer-Aided Lung Nodule Detection Software in Thoracic CT for Riverain Technologies LLC: https://ClinicalTrials.gov/show/NCT02440139, 2015.	Clinical trial register, no outcomes yet
132. NCT04119960. Clinical Validation of InferRead Lung CT.AI: https://clinicaltrials.gov/show/NCT04119960, 2019.	Clinical trial register, no outcomes yet
133. NCT04792632. Clinical Performance Evaluation of Veye Lung Nodules: https://clinicaltrials.gov/show/NCT04792632, 2021.	Clinical trial register, no outcomes yet
Excluded on outcomes – No relevant outcomes reported (n=11)	
134. Buckler AJ, Danagoulian J, Johnson K, et al. Inter-Method Performance Study of Tumor Volumetry Assessment on Computed Tomography Test-Retest Data. Acad Radiol 2015;22(11):1393-408. doi: https://dx.doi.org/10.1016/j.acra.2015.08.007	No relevant outcomes reported
135. ChiCTR1900021144. Evaluation of AI-assisted detection of lung nodules in low dose CT images: http://www.chictr.org.cn/showproj.aspx?proj=35698, 2019.	No relevant outcomes reported
136. ChiCTR2000029278. A blinded, self-control trial to evaluate an AI based CAD system for Lung Nodule Diagnosis: http://www.chictr.org.cn/showproj.aspx?proj=48219, 2020.	No relevant outcomes reported
137. Ganti S. Radiological lessons, tips and tricks from UK's first lung cancer screening site. Lung Cancer 2020;139(Supplement 1):S6. doi: http://dx.doi.org/10.1016/S0169-5002%2820%2930041-6	No relevant outcomes reported
138. Heuvelmans MA, Walter JE, Vliegenthart R, et al. Disagreement of diameter and volume measurements for pulmonary nodule size estimation in CT lung cancer screening. Thorax 2018;73(8):779-81. doi: https://dx.doi.org/10.1136/thoraxjnl-2017-210770	No relevant outcomes reported
139. Kisby G, Dentry M. Use of computer-aided detection (CAD) in CT Chest imaging for the diagnosis of lung nodules. Journal of Medical Imaging and Radiation Oncology 2021;65(SUPPL 1):143. doi: http://dx.doi.org/10.1111/1754-9485.13301	No relevant outcomes reported
140. Lee J, Kim Y, Kim HY, et al. Feasibility of implementing a national lung cancer screening program: Interim results from the Korean Lung Cancer Screening Project (K-LUCAS). Transl 2021;10(2):723-36. doi: https://dx.doi.org/10.21037/tlcr-20-700	No relevant outcomes reported
141. Lee J, Lim J, Kim Y, et al. Development of Protocol for Korean Lung Cancer Screening Project (K-LUCAS) to Evaluate Effectiveness and Feasibility to Implement National Cancer Screening Program. Cancer Res 2019;51(4):1285-94. doi: https://dx.doi.org/10.4143/crt.2018.464	No relevant outcomes reported
142. Nct. Evaluation of Use of Diagnostic AI for Lung Cancer in Practice. https://clinicaltrialsgov/show/NCT03780582 2018	No relevant outcomes reported
143. Park S, Lee SM, Do KH, et al. Deep Learning Algorithm for Reducing CT Slice Thickness: Effect on Reproducibility of Radiomic Features in Lung Cancer. Korean J Radiol 2019;20(10):1431-40. doi: https://dx.doi.org/10.3348/kjr.2019.0212	No relevant outcomes reported
144. Schreuder A, van Ginneken B, Scholten ET, et al. Classification of CT Pulmonary Opacities as Perifissural Nodules: Reader Variability. Radiology 2018;288(3):867-75. doi: https://dx.doi.org/10.1148/radiol.2018172771	No relevant outcomes reported

Reference	Main reason for exclusion
Excluded on publication type – Conference abstract with no additional data reported (n=2)	
145. Hwang EJ, Yoon SH, Goo JM, et al. P2.11-16 Variability in Reading Low-Dose Chest CT: Individual Readers vs. Central Review in a Nationwide Lung Cancer Screening Project. Journal of Thoracic Oncology 2019;14(10 Supplement):S798-S99. doi: http://dx.doi.org/10.1016/j.jtho.2019.08.1716	Conference abstract with no additional data reported
146. Lo S, Freedman M, Mun SK. The application of a vessel suppressed function incorporated with lung opacity analysis for the significant increase of nodule detectability in CT. International Journal of Computer Assisted Radiology and Surgery 2017;12(1 Supplement 1):S150. doi: http://dx.doi.org/10.1007/s11548-017-1588-3	Conference abstract with no additional data reported
Excluded on publication type – no primary research article (n=2)	
147. Crosby D, Lyons N, Greenwood E, et al. A roadmap for the early detection and diagnosis of cancer. The Lancet Oncology 2020;21(11):1397-99. doi: http://dx.doi.org/10.1016/S1470-2045%2820%2930593-3	No primary research article
148. Svoboda E. Artificial intelligence is improving the detection of lung cancer. Nature 2020;587(7834):S20-S22. doi: https://dx.doi.org/10.1038/d41586-020-03157-9	No primary research article
Excluded - Duplicate (n=2)	
149. Mun SK, Lo SB, Freedman MT, et al. Computer-aided detection of lung nodules on CT with a computerized pulmonary vessel suppressed function. American Journal of Roentgenology 2018;210(3):480-88. doi: http://dx.doi.org/10.2214/AJR.17.18718	Duplicate
150. Yuan R, Mayo J, Streit I, et al. Randomized Clinical Trial with Computer Assisted Diagnosis (CAD) Versus Radiologist as First Reader of Lung Screening LDCT. Journal of Thoracic Oncology 2019;14(10):S287-S88. doi: 10.1016/j.jtho.2019.08.578	Duplicate

Table 65. Publications excluded after review of full-text articles – Studies provided by companies (n=99)

Reference	Main reason for exclusion
Aidence B.V. (n=21)	
1. Accelerated Access Collaborative. AI in Health and Care Award – Scoping Plan	No relevant data reported.
2. Aidence (2021). Veye Lung Nodules - Instructions for Use. Software version 3.6.0, Document ID: SIGN-234	Same data as in Clinical Evaluation Report.
3. Aidence (2021). Integrating Veye Bridge 2.x	No relevant data reported.
4. Aidence (2021). Veye Lung Nodules - Version 3.6 Change Impact Assessment, Document ID: SIGN-268	No relevant data reported.
5. Aidence (2021). Competitor overview	No relevant data reported.
6. Aidence (2021). NICE DAP60 – AI for chest CT. Checklist of confidential information. 8th October 2021	No relevant data reported.
7. Aidence (2021). NICE DAP60 – AI for chest CT. Request for information	No relevant data reported.
8. Aidence (2022). NICE DAP60 – AI for chest CT. Checklist of confidential information. 15th June 2022.	No relevant data reported.
9. Aidence (2022). NICE DAP60 – AI for chest CT. Request for information (revised)	No relevant data reported.
10. Aidence. Veye Chest – Datasets & Validation	Reported data already included.
11. Blazis SP, Dickerscheid DBM, Linsen PVM, et al. Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system. Eur J Radiol 2021;136:109526. doi: https://dx.doi.org/10.1016/j.ejrad.2021.109526	Already included
12. DEKRA (2021). EU Quality Management System Certificate	No relevant data reported.
13. De Monye W & Wakkie J. Efficiency Study Veye Chest (EFFEY STUDY) – Collaboration between Spaarne Gasthuis and Aidence B.V.	No relevant data reported.
14. Jacobs C, Setio AAA, Scholten ET, et al. Deep Learning for Lung Cancer Detection on Screening CT Scans: Results of a Large-Scale Public Competition and an Observer Study with 11 Radiologists. Radiol Artif Intell. 2021;3(6):e210027. Published 2021 Oct 27. doi:10.1148/ryai.2021210027	Malignancy risk prediction.
15. Martins Jarnalo CO, Linsen PVM, Blazis SP, et al. 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(11):838-45. doi: https://dx.doi.org/10.1016/j.crad.2021.07.012	Already included
16. Murchison JT, Ritchie G, Senyszak D, Nijwening JH, van Beek EJR. Evaluation of a deep learning software tool for CT based lung nodule growth assessment. ECR 2019 (Poster C-3685)	Already included
17. Murchison JT, Ritchie G, Senyszak D, Nijwening JH, van Beek EJR. Evaluation of a deep learning software tool for CT based lung segmentation. ECR 2019 (Poster C-3686)	Already included
18. Murchison JT, Ritchie G, Senyszak D, Nijwening JH, van Beek EJR. Validation of a deep learning computer aided system for CT based lung nodule detection, classification and quantification and growthrate estimation in a routine clinical population. Manuscript 2021	Already included.

Reference	Main reason for exclusion
19. Murchison JT, Ritchie G, Senyszak D, et al. 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(5):e0266799. Published 2022 May 5. doi:10.1371/journal.pone.0266799	Already included.
20. Rezazade Mehrizi MH & Algra P. AI Implementation Stories- Lessons learned in NWZ hospital. MEMO RAD 2021;26(1):17-18.	No methods reported.
21. Wakkie J & van Veenendaal G (2020). 510(k) Study Protocol Clinical Performance Evaluation of Veye Lung Nodules - Standalone performance and reader study	No relevant data reported.
contextflow GmbH (n=20)	
22. Agarwal P, et al. (2022) Combining Content-Based Image Retrieval with a knowledge-based diagnostic decision support system in chest-CT. ECR 2022 (July, Vienna)	Conference abstract without related full journal article.
23. Calhaun ME, Hofmanninger J, Wood C, Langs G, Makropoulos A. Combining automated malignancy risk estimation with lung nodule detection may reduce physician effort and increase diagnostic accuracy. Draft abstract. Lung cancer conference IASLC 2022 (August, Vienna)	Draft conference abstract. IN if full text article (below) becomes available.
24. Calhaun ME, Hofmanninger J, Wood C, Langs G, Makropoulos A. A publication validating malignancy score prediction in a large cohort. (to be published 14 July 2022)	No full text available by 31/08/2022
25. contextflow (2021). TD Intended Use - contextflow SEARCH Lung CT	No relevant data reported.
26. contextflow (2021). TD Intended Use - contextflow SEARCH Lung CT. Version 2.0	Not enough information on study methods provided by 31/08/2022
27. contextflow (2021). EU Declaration of Conformity contextflow SEARCH Lung CT	No relevant data reported.
28. contextflow (2021). NICE DAP60 – AI for chest CT. Request for information	No relevant data reported
29. contextflow (2021). NICE DAP60 – AI for chest CT. Checklist of confidential information. 29th December 2021.	No relevant data reported
30. contextflow (2022). NICE DAP60 – AI for chest CT. Checklist of confidential information. 24th May 2022	No relevant data reported
31. Contextflow (2022). contextflow - scientific evidence and publications	No relevant data reported.
32. Pan J & Langs G. (2022). Prediction of disease severity in COVID-19 patients identifies predictive disease patterns in lung CT. European Society of Thoracic Imaging/ESTI 2022 (June, Oxford)	Conference abstract without related full journal article. Disease severity in COVID-19 patients.
33. Pan J & Langs G. (2022). Evaluation of diagnosing diffuse parenchymal lung disease in pulmonary CTs. European Society of Thoracic Imaging/ESTI 2022 (June, Oxford)	Conference abstract without related full journal article.
34. Pan J & Langs G. (2022). Comparing predictive values for patterns in patients with/without ICU treatment. Abstract submitted to RSNA 2022.	Conference abstract without related full journal article.

Reference	Main reason for exclusion
35. Pieler M, Hofmanninger J, Donner R, Sikka A, Jiménez Arroyo E, Prosch H, Zhang R, Krenn M, Langs G, Makropoulos A. Evaluation of automatic volumetry of honeycombing and ground glass opacity patterns in lung CT scans (abstract submitted to ECR 2022, July/Vienna)	Conference abstract without related full journal article.
36. Prayer, F., Röhrich, S., Pan, J. et al. Künstliche Intelligenz in der Bildgebung der Lunge. Radiologe 60, 42–47 (2020). https://doi.org/10.1007/s00117-019-00611-2	Full text in German language.
37. Preyer F. et al. (2022). Dermatomyositis - description of the cohort, in terms of quantitative profiles. Abstract submitted to RSNA 2022.	Conference abstract without related full journal article.
38. Röhrich, S., Schlegl, T., Bardach, C. et al. Deep learning detection and quantification of pneumothorax in heterogeneous routine chest computed tomography. Eur Radiol Exp 4, 26 (2020). https://doi.org/10.1186/s41747-020-00152-7	Technology used for automated, volume-level pneumothorax grading (presence and size). No relevant outcomes reported.
39. Röhrich S, et al. (2022). Evaluation of diagnosing diffuse parenchymal lung disease in pulmonary CTs. European Society of Thoracic Imaging/ESTI 2022 (June, Oxford)	Conference abstract without related full journal article.
40. Röhrich S, et al. (2022). Results of the Big Medilytics Study (BML). ECR 2022 (July, Vienna)	Conference abstract with now additional data to included full article.
41. TÜV SÜD (2021). EU Quality Management System Certificate (MDR)	No relevant data reported.
Infervision Medical Technology Co., Ltd. (n=13)	
42. BSI (2020). EC Certificate – Full Quality Assurance System	No relevant data reported.
43. Hu Q, et al. Application of computer-aided detection (CAD) software to automatically detect nodules under SDCT and LDCT scans with different parameters. Computers in Biology and Medicine. Vol. 146, July 2022, 105538.	Population not eligible (>10% patients with extrathoracic cancers; see Table 65).
44. Infervision (2021). NICE DAP60 – AI for chest CT. Checklist of confidential information. 20th October 2021	No relevant data reported.
45. Infervision (2021). NICE DAP60 – AI for chest CT. Request for information	No relevant data reported.
46. Infervision (2022). NICE DAP60- AI for chest CT. Checklist for confidential information. 13th June 2022.	No relevant data reported.
47. Infervision. InferRead CT Lung V2 – Instructions for Use	No relevant data reported.
48. Kozuka T. et al. (2020) Efficiency of a computer-aided diagnosis system with deep learning in detection of pulmonary nodules on 1-mm-thick images of computed tomography. Japanese Journal of Radiology. https://link.springer.com/article/10.1007/s11604-020-01009-0	Already IN
49. Li K, 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 Aug;11(8):3629-3642. doi: 10.21037/qims-20-1314.	Population not eligible (>10% patients with previously diagnosed lung cancer; see Table 65).

Reference	Main reason for exclusion
50. Liu K. et al. (2019) Evaluating a Fully Automated Pulmonary Nodule Detection Approach and Its Impact on Radiologist Performance. Radiology. https://pubs.rsna.org/doi/full/10.1148/ryai.2019180084	Already IN
51. Ma J. et al (2020). Survey on deep learning for pulmonary medical imaging. Frontiers in Medicine. https://pubmed.ncbi.nlm.nih.gov/31840200/	Review
52. Wang L. et al. (2020) Toward standardized premarket evaluation of computer aided diagnosis/detection products: insights from FDA-approved products. Expert Review of Medical Devices. https://pubmed.ncbi.nlm.nih.gov/32842797/	Review
53. Wang Y. et al. (2019) IILS: Intelligent imaging layout system for automatic report standardization and intra-interdisciplinary clinical workflow optimization. EBioMedicine. https://pubmed.ncbi.nlm.nih.gov/31129095/	Population not eligible (>10% patients with extrathoracic cancers; see Table 65).
54. Yang K. et al. (2020). Identification of benign and malignant pulmonary nodules on chest CT using improved 3D U-Net deep learning framework. European Journal of Radiology. https://pubmed.ncbi.nlm.nih.gov/32505895/	Malignancy risk prediction.
JLK Inc. (n=8)	
55. Advena Limited (2021). Certificate of Registration	No relevant data reported.
56. JLK (2021). NICE DAP60 – AI for chest CT. Request for information	No relevant data reported.
57. JLK (2021). NICE DAP60 – AI for chest CT. Checklist of confidential information. 26th November 2021	No relevant data reported.
58. JLK (2022). NICE DAP60 – AI for chest CT. Checklist of confidential information. 27th May 2022.	No relevant data reported.
59. JLK (2021). EC Declaration of Conformity (Self-Certification)	No relevant data reported.
60. Psaila A (Advena Ltd) (2021). Device Description & Specification for JLD-01K. Issue 1.0, Document ref. DD500	No relevant data reported.
61. Xiao, Zhitao, Bowen Liu, Lei Geng, Fang Zhang, and Yanbei Liu. 2020. "Segmentation of Lung Nodules Using Improved 3D-UNet Neural Network" Symmetry 12, no. 11: 1787. https://doi.org/10.3390/sym12111787	Software not commercially available.
62. Wang J, Wang JW, Wens YF, et al. Pulmonary Nodule Detection in Volumetric Chest CT Scans Using CNNs-Based Nodule-Size-Adaptive Detection and Classification. IEEE Access 2019;7:46033-44. doi: 10.1109/ACCESS.2019.2908195	Software not commercially available.
MeVis Medical Solutions (n=22)	
63. Cohen JG, Goo JM, Yoo RE, Park CM, Lee CH, van Ginneken B, Chung DH, Kim YT. Software performance in segmenting ground glass and solid components of subsolid nodules in pulmonary adenocarcinomas. Eur Radiol. 2016 Dec;26(12):4465-4474.	Already IN

Reference	Main reason for exclusion
64. Cohen JG, Goo JM, Yoo RE, Park SB, van Ginneken B, Ferretti GR, Lee CH, Park CM. 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 Jun;85(6):1174-80.	Population not eligible (53 adenocarcinomas presenting as part-solid nodules in 50 patients; report in extra table)
65. Cohen JG, Kim H, Park SB, van Ginneken B, Ferretti GR, Lee CH, Goo JM, Park CM. Comparison of the effects of model-based iterative reconstruction and filtered back projection algorithms on software measurements in pulmonary subsolid nodules. Eur Radiol. 2017 Aug;27(8):3266-3274.	Already IN
66. Jacobs C, Sánchez CI, Saur SC, Twellmann T, de Jong PA, van Ginneken B. Computer-aided detection of ground glass nodules in thoracic CT images using shape, intensity and context features. Med Image Comput Comput Assist Interv. 2011;14(Pt 3):207-14.	Published before 2012.
67. Jacobs C, van Rikxoort EM, Twellmann T, Scholten ET, de Jong PA, Kuhnigk JM, Oudkerk M, de Koning HJ, Prokop M, Schaefer-Prokop C, van Ginneken B. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. Med Image Anal. 2014 Feb;18(2):374-84.	Software not commercially available - the tested algorithm for detection of SSNs is not included in Veolity (MeVis).
68. Jacobs C, van Rikxoort EM, Murphy K, Prokop M, Schaefer-Prokop CM, van Ginneken B. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. Eur Radiol. 2016 Jul;26(7):2139-47.	Visia (MeVis Medical Solutions) not Veolity
69. Jacobs C, Schreuder A, van Riel SJ, 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(5):e200160. doi: https://dx.doi.org/10.1148/rycan.2021200160	Already included.
70. MEDCERT (2018). EC-Certificate of Conformity	No relevant data reported.
71. MeVis (2021). VeolityTM LungCAD 1.7 User Guide (Instructions for Use)	Company did not provide further details on population by 31/08/2022
72. MeVis (2021). VeolityTM LungRead 1.7 User Guide (Instructions for Use)	No relevant data reported.
73. MeVis (2021). Veolity 1.7 Scanning Protocol Recommendations	No relevant data reported.
74. Ritchie AJ, Sanghera C, Jacobs C, Zhang W, Mayo J, Schmidt H, Gingras M, Pasian S, Stewart L, Tsai S, Manos D, Seely JM, Burrowes P, Bhatia R, Atkar-Khattra S, van Ginneken B, Tammemagi M, Tsao MS, Lam S; Pan-Canadian Early Detection of Lung Cancer Study Group. Computer Vision Tool and Technician as First Reader of Lung Cancer Screening CT Scans. J Thorac Oncol. 2016 May;11(5):709-717.	CIRRUS Lung Screening
75. Scholten ET, Jacobs C, van Ginneken B, Willemink MJ, Kuhnigk JM, van Ooijen PM, Oudkerk M, Mali WP, de Jong PA. Computer-aided segmentation and volumetry of artificial ground glass nodules at chest CT. AJR Am J Roentgenol. 2013 Aug;201(2):295-300.	Artificial ground glass nodules
76. Scholten ET, de Hoop B, Jacobs C, van Amelsvoort-van de Vorst S, van Klaveren RJ, Oudkerk M, Vliegenthart R, de Koning HJ, van der Aalst CM, Mali WT, Gietema HA, Prokop M, van Ginneken B,	Author contacted about software used but no reply

Reference		Main reason for exclusion
	de Jong PA. Semi-automatic quantification of subsolid pulmonary nodules: comparison with manual measurements. PLoS One. 2013 Nov 21;8(11):e80249.	
77.	Scholten ET, de Jong PA, Jacobs C, van Ginneken B, van Riel S, Willemink MJ, Vliegenthart R, Oudkerk M, de Koning HJ, Horeweg N, Prokop M, Mali WP, Gietema HA. Interscan variation of semi-automated volumetry of subsolid pulmonary nodules. Eur Radiol. 2015 Apr;25(4):1040-7.	CIRRUS Lung Screening
78.	Scholten ET, Jacobs C, van Ginneken B, van Riel S, Vliegenthart R, Oudkerk M, de Koning HJ, Horeweg N, Prokop M, Gietema HA, Mali WP, de Jong PA. Detection and quantification of the solid component in pulmonary subsolid nodules by semiautomatic segmentation. Eur Radiol. 2015 Feb;25(2):488-96.	CIRRUS Lung Screening
79.	Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard CVD, Cerello P, Chen H, Dou Q, Fantacci ME, Geurts B, Gugten RV, Heng PA, Jansen B, de Kaste MMJ, Kotov V, Lin JY, Manders JTMC, Sóñora-Mengana A, García-Naranjo JC, Papavasileiou E, Prokop M, Saletta M, Schaefer-Prokop CM, Scholten ET, Scholten L, Snoeren MM, Torres EL, Vandemeulebroucke J, Walasek N, Zuidhof GCA, Ginneken BV, Jacobs C. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal. 2017 Dec;42:1-13.	Visia CT Lung CAD system
80.	Silva M, Prokop M, Jacobs C, Capretti G, Sverzellati N, Ciompi F, van Ginneken B, Schaefer-Prokop CM, Galeone C, Marchianò A, Pastorino U. Long-Term Active Surveillance of Screening Detected Subsolid Nodules is a Safe Strategy to Reduce Overtreatment. J Thorac Oncol. 2018 Oct;13(10):1454-1463.	CIRRUS Lung Screening
81.	Silva M, Schaefer-Prokop CM, Jacobs C, Capretti G, Ciompi F, van Ginneken B, Pastorino U, Sverzellati N. Detection of Subsolid Nodules in Lung Cancer Screening: Complementary Sensitivity of Visual Reading and Computer-Aided Diagnosis. Invest Radiol. 2018 Aug;53(8):441-449.	CIRRUS Lung Screening (research version with prototype Veolity)
82.	SynApps Solutions Ltd. (2021). NICE DAP60 – AI for chest CT. Checklist of confidential information. 17th September 2021.	No relevant data reported.
83.	SynApps Solutions Ltd. (2022). NICE DAP60 – AI for chest CT. Checklist of confidential information. 20th May 2022.	No relevant data reported.
84.	SynApps Solutions Ltd. (2021). NICE DAP60 – AI for chest CT. Request for information	No relevant data reported.
Riverain Technologies (n=12)		
85.	Intertek Semko AB (2020). EC Certification – Full Quality Assurance System	No relevant data reported.
86.	Lo SB, Freedman MT, Gillis LB, et al. JOURNAL CLUB: Computer-Aided Detection of Lung Nodules on CT With a Computerized Pulmonary Vessel Suppressed Function. AJR Am J Roentgenol 2018;210(3):480-88. doi: https://dx.doi.org/10.2214/AJR.17.18718	Already IN

Reference	Main reason for exclusion
87. Martini K, Bluthgen C, Eberhard M, et al. Impact of Vessel Suppressed-CT on Diagnostic Accuracy in Detection of Pulmonary Metastasis and Reading Time. Acad Radiol 2021;28(7):988-94. doi: https://dx.doi.org/10.1016/j.acra.2020.01.014	OUT on population (>10% with extrathoracic cancer, see Table 65).
88. Milanese G, Eberhard M, Martini K, et al. Vessel suppressed chest Computed Tomography for semi-automated volumetric measurements of solid pulmonary nodules. Eur J Radiol 2018;101:97-102. doi: https://dx.doi.org/10.1016/j.ejrad.2018.02.020	Already IN
89. Riverain Technologies. ClearRead CT Abridged User Guide Version 5.0	No relevant data reported.
90. Riverain Technologies (2021). ClearRead CT Compare Sample	No relevant data reported.
91. Riverain Technologies (2021). ClearRead CT Detect Sample	No relevant data reported.
92. Riverain Technologies (2021). NICE DAP60 – AI for chest CT. Checklist of confidential information. 9th November 2021.	No relevant data reported.
93. Riverain Technologies (2022). NICE DAP60 – AI for chest CT. Checklist of confidential information. 23rd May 2022.	No relevant data reported.
94. Riverain Technologies (2021). NICE DAP60 – AI for chest CT. Request for information	No relevant data reported.
95. Singh R, Kalra MK, Homayounieh F, et al. Artificial intelligence-based vessel suppression for detection of sub-solid nodules in lung cancer screening computed tomography. Quant 2021;11(4):1134-43. doi: https://dx.doi.org/10.21037/qims-20-630	Already IN
96. van Leeuwen, K.G., Schalekamp, S., Rutten, M.J.C.M. et al. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 31, 3797–3804 (2021). https://doi.org/10.1007/s00330-021-07892-z	Overview
Siemens Healthineers (n=3)	
97. Siemens Healthineers (2020). AI-Rad Companion (Pulmonary). Addendum – Usage of syngo.CT Lung CAD VD20. VA12	No additional information on study methods provided by 31/08/2022
98. Siemens Healthineers (2021). AI-Rad Companion. Instructions for Use – AI-Rad Companion (Pulmonary) VA13	No relevant data reported.
99. Siemens Healthineers (2021). AI-Rad Companion (Pulmonary). Addendum – Lung CAD VD20. VA13	No additional information on study methods provided by 31/08/2022

Table 66. Study characteristics and main outcomes of records excluded on study population only (n=11)

Reference, country	Study design	Aim	Population	Index test	Comparator	Reported outcomes
Ahn 2021,[34] Korea	MRMC study: Stand-alone AI vs readers with concurrent AI.	To evaluate the performance of a commercially available DL-algorithm for automatic measurement of the solid portion of surgically proven lung adenocarcinomas manifesting as subsolid lesions.	Asan medical Centre (Seoul, Korea), January to December 2018; 448 patients with 448 SSNs ≥6 mm who had undergone curative resection of non-small lung cancer with chest CT performed <30 days of surgery.	VUNO MedLungCT AI, version 1.0.0 (VUNO); Nodule segmentation and measurement of solid portion (maximal axial diameter; maximal diameter multiplanar); [A] Stand-alone AI; [C] Concurrent AI: 5 radiologists with 4-26 years of experience.	None	Segmentation adequacy and failure rate [A]; Measurement inter-observer variability [C]; Measurement agreement between [A] vs [C]; [A] vs invasive size at pathologic examination; [C] vs invasive size at pathologic examination.
Cohen 2016a,[35] Korea	MRMC study: Concurrent AI vs unaided readers.	To evaluate the performance of computer-aided segmentation of ground glass and solid components in SSNs and to compare the software and pathology measurements in pulmonary adenocarcinomas manifesting as SSNs.	Thoracic surgery database (Seoul National University College of Medicine) review for surgically resected GGNs between 2013 and 2015. 23 patients with 73 resected pulmonary adenocarcinomas manifesting as SSNs.	Veolity version 1.1 (MeVis); Nodule segmentation and measurement (maximal transverse diameter) of the entire nodule and its solid component (9 different attenuation thresholds); [C] Concurrent AI: 1 radiologist with 4 years of experience.	[D] Unaided readers: 2 radiologists with 24 and 4 years of experience, respectively, using electronic clippers (entire nodule: lung window; solid component: lung and mediastinal windows).	Segmentation adequacy and failure rate [A]; Agreement between [C]/[D] and tumour sizes on pathology; Agreement between [C]/[D] and invasive component sizes on pathology; Diagnostic accuracy of [C] and [D] in predicting AIS or MIA; Agreement between [C] and [D] measurements for ground glass components and solid components, respectively.

Reference, country	Study design	Aim	Population	Index test	Comparator	Reported outcomes
Cohen 2016b[36], Korea	MRMC study: Concurrent software use - unenhanced vs enhanced CT images.	To evaluate the differences in semi-automatic measurements of CT attenuation and volume of part-solid nodules between unenhanced and enhanced CT scans.	Retrospectively reviewed all preoperative CT scans for the part-solid nodules with consecutive pre-and post-enhancement acquisitions between July 2014 and May 2015 (Seoul National University College of Medicine). 53 lung adenocarcinomas presenting as part-solid nodules in 50 patients.	Veolity version 1.1 (MeVis); Nodule segmentation and measurement (largest diameter, volume, mass and attenuation) of ground glass and solid components; [C] Concurrent AI: 1 radiologist with 4 years of experience.	None	Segmentation adequacy and failure rate [A]; Difference between [C] unenhanced and [C] enhanced acquisitions (largest diameter, volume, mass, attenuation) for the whole nodule and solid component, respectively; Difference of measures between [C] unenhanced and [C] enhanced according to adenocarcinoma category (AIS/MIA and IA).
Garzelli 2018,[37] Korea	MRMC study: Concurrent software use - with and without vessel removal.	To evaluate the value of vessel removal algorithm in semi-automatic segmentation of subsolid nodule by comparing the software measurements of the solid component on CT with and without vessel removal, with the measurement of invasive tumour on pathology in lung	Medical record review of all patients who had undergone surgical resection for lung adenocarcinomas and pre-invasive lesions that manifested as subsolid nodules at Seoul National University hospital between January 2014 and June 2015. 73 patients were included.	AVIEW LungScreen (Coreline Soft); Nodule segmentation and measurement of the ground glass and solid components, with and without vessel removal function (3D longest, axial longest and effective diameters); [C] Concurrent AI: 2 radiologists with 3 and 26 years of experience.	None	Segmentation adequacy and failure rate [C]; Comparison of [C] with the pathology measurements (whole tumour size and invasive component) with and without vessel removal; Inter-reader variability and intra-reader variability of [C] with and without vessel removal; Diagnostic accuracy of [C] (3D longest diameter of solid component with vessel

Reference, country	Study design	Aim	Population	Index test	Comparator	Reported outcomes
		adenocarcinomas manifesting as subsolid nodules.		1 repeated all measurements at a 3-week interval.		removal) in predicting AAH/AIS/MIA.
Hu 2022,[40] China	Prospective test accuracy study: Stand-alone AI performance – effect of dose, blending level and definition modes.	To compare the CAD detectable sensitivity on pulmonary nodules between SDCT and LDCT scans with different parameters including definition modes and blending levels of adaptive statistical iterative reconstruction.	117 patients with extra-thoracic malignancies who were scheduled for chest CT examination to detect pulmonary metastasis or for follow-up to monitor their conditions and determine the treatment response. July to December 2017.	InferRead CT lung (InferVision); Blending levels (0%, 60%, 80%); Definition modes (HD, non-HD; [A] Stand-alone AI.	None	Sensitivity and FP rate for LDCT and SDCT at various blending levels and HD and non-HD modes.
Li 2021,[43] China	MRMC study: Stand-alone AI vs unaided readers	To evaluate the accuracy of an AI-driven commercial CAD product (InferRead CT Lung Research) in malignancy risk prediction using a real-world database.	442 consecutive patients with 525 lesions who underwent lung resection at Fifth Affiliated Hospital of Sun Yat-sen University between September 2015 and November 2018.	InferRead CT Lung Research (Infervision); Nodule detection and risk prediction (low, moderate, high). [A] Stand-alone AI.	[D] Unaided reader: 2 radiology residents (3 years of chest radiology and general radiology experience, respectively) independently reviewed and graded each lesion (high-risk, >70%), moderate-risk, 50-70%), low-risk, <50%).	Technical failure rate of [A] (detection errors); Lung nodule detection rate (sensitivity) of [A]; Characteristics of missed nodules [A]; Accuracy of malignancy risk prediction [A] [D]; Comparison of malignancy risk prediction accuracy: [A] vs [D].

Reference, country	Study design	Aim	Population	Index test	Comparator	Reported outcomes
Martini 2021,[38] Switzerland	MRMC study: Software-assisted readers vs unaided readers.	To evaluate if vessel suppressed-CT increases nodule detection rate, improves interreader agreement, and decreases reading time in chest CT of oncologic patients.	100 consecutive oncologic patients who were referred for a clinically indicated contrast-enhanced CT between January 2014 and December 2017 at 2 institutions in Zurich.	ClearRead-CT (Riverain Technologies) Vessel suppression function; [C] Concurrent AI: 6 radiologists (2 with 1-2 years, 2 with 5 years and 2 with 8-9 years of experience). Nodule detection on vessel-suppressed CT images (with access to standard CT images).	[D] Unaided reader: 6 radiologists (2 with 1-2 years, 2 with 5 years and 2 with 8-9 years of experience). Nodule detection on standard CT images.	Nodule detection rate [C] [D]; Inter-reader agreement [C] [D]; FP rate in vessel-suppressed CT images [C]; FN rate in vessel-suppressed CT images [C]; Reading time [C] [D].
Park 2021,[39] Korea	Retrospective test accuracy study: Performance of stand-alone AI.	To assess the effect of CT section thickness on the performance of CAD for detecting SSNs and to investigate whether DL-based super-resolution algorithms for reducing CT section thickness can improve performance.	Electronic medical records of a tertiary referral institution (Asan Medical Center) from March 2018 to December 2018; 308 patients with SSN (6-30 mm) who underwent curative resection of lung adenocarcinoma.	VUNO Med-LungCT AI version 1.0.0 (VUNO); Detection of SSNs on CT images of each section thickness (1 mm, 3 mm, 5 mm; super-resolution algorithm for CT section reduction applied to the 3- and 5-mm CT images to convert them into 1-mm CT images; [A] Stand-alone AI.	None	Performance of [A] to detect SSNs at all 3 section thicknesses and at converted images (per-lesion and per-patient).
Wagner 2018,[41] Germany	Retrospective test accuracy study:	To evaluate the accuracy of a CAD application for pulmonary nodular lesions in CT scans.	100 consecutive patients with 106 biopsied nodules (50 confirmed as bronchial cancer; 11 metastatic disease of various	ClearReadCT (Riverain Technologies), versions V1 and V2, release candidates (V2 now commercially available);	None	Segmentation failure rate [A]; Agreement in segmentations between V1 and V2; Volume-diameter correlation;

Reference, country	Study design	Aim	Population	Index test	Comparator	Reported outcomes
	Stand-alone AI versions V1 vs V2; Effect of contrast enhancement, reconstruction kernel and slice thickness.		origin; 39 with a benign lesion) from the University Hospital Jena between 6/2007 and 2/2016.	Nodule detection and measurement (volume, maximal diameter). [A] Stand-alone AI.		Sensitivity and FP rate for V1 and V2 (non-enhanced vs enhanced, lung vs soft tissue kernel, 0.75 mm vs 1.5 mm vs 3.0 mm); [A] oversights (FN); Additional verified nodules detected by [A].
Wang 2019,[44] China	MRMC study: Stand-alone AI vs unaided readers.	To achieve imaging report standardization and improve the quality and efficiency of the intra-interdisciplinary clinical workflow, we proposed an intelligent imaging layout system (IILS) for a clinical decision support system-based ubiquitous healthcare service, which is a lung nodule management system using medicalimages.	Independent test set: 1,965 LDCT with or without contrast selected from retrospective cohorts of adult patients from Nanjing Drum Tower Hospital, Northern Jiangsu People's Hospital, Ningbo No.2 Hospital, and NanJing GaoChun People's Hospital between October 2016 and November 2018.	Intelligent imaging layout system (IILS) – Company confirmed that this is "same product" as InferRead CT Lung (Infervision); Nodule detection and classifying benign or malignant; [A] Stand-alone AI.	[D] Unaided readers: 6 radiologists; twice by six experts with a 1-month time interval.	Nodule detection: agreement with gold standard for [A] and 6 unaided readers [D] for all nodules and by nodule size.

Reference, country	Study design	Aim	Population	Index test	Comparator	Reported outcomes
Yacoub 2021,[42] USA	Retrospective test accuracy study: Stand-alone AI vs original radiology reports.	To assess the performance of an AI platform compared against clinical radiology reports on non-contrast chest CT scans.	100 consecutive patients who had previously undergone non-contrast chest CT between October to November 2019 were retrospectively identified. 24% Evaluation for metastasis; 11% Lung cancer; 5% Evaluation for primary tumour.	AI-Rad Companion (Siemens Healthineers) Detection of pulmonary lesions (nodule or mass ≥5 mm in size). [A] Stand-alone AI.	[E] Original radiology reports.	Sensitivity; Specificity; Positive predictive value; Negative predictive value; AUC.

AI, Artificial intelligence; AIS, Adenocarcinoma in situ; AUC, Area under the receiver-operating curve; CAD, Computer-aided detection; CT, Computed tomography; DL, Deep learning; FN, False negative; FP, False positive; GGN, Ground glass nodule; HD, High definition; IA, Invasive adenocarcinoma; LDCT, Low-dose computed tomography; MIA, Minimally invasive adenocarcinoma; MRMC, Multi-reader multiple-case study; SDCT, Stand-dose computed tomography; SSN, Sub-solid nodule; TP, True positive.

Table 67. Characteristics of ongoing and/or unpublished studies (7 studies)

Reference, country	Title and study design	Population	Index test	Comparator	Outcomes	Completion date
[redacted] 86	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted] 87	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
NCT04119960 (2019),[88] USA	Clinical Validation of InferRead Lung CT.AI / MRMC study	N=250 50-75 years, lung cancer screening eligible patients. Probability sample. University of Maryland Medical Center (Baltimore, Maryland, USA)	InferRead Lung CT.AI (Infervision): [C] Radiologists with software	[D] Radiologists without software	Detection accuracy (AUC, sensitivity, specificity, PPV, NPV) ([C] vs [D])	30 October 2019
NCT02871856 (2021),[89, 90] Australia, Canada,	International Lung Screen Trial (ILST) / Randomised controlled sub-study to evaluate the utility	N=4,500 Ever smokers between 55-80 years.	Veolity 1.2 (MeVis): [B] Radiologist-read first then CAD-verified;	Radiologist review using a standardised reporting protocol will	Sub-study on CAD: Radiologist reporting time ([B] vs [C]);	Anticipated December 2023

Reference, country	Title and study design	Population	Index test	Comparator	Outcomes	Completion date
Hongkong, Spain	of CAD to improve radiologist reporting time and accuracy.	9 recruitment sites in Australia, Canada, Hongkong, Spain. Sub-study on CAD will be performed at some sites only.	[C] CAD-verified first then radiologist-read.	remain the gold standard	Diagnostic accuracy ([B] vs [C]).	
NCT04792632 (2021),[91] USA	Clinical Performance Evaluation of Veye Lung Nodules (CPEVLN) / MRMC study	N=350 18 years and older; US population that received a chest CT scan either as part of a lung cancer screening programme or during routine practice (Intrinsic Imaging, Boston, Massachusetts)	Veye Lung Nodules (Aidence): [A] Stand-alone software; [C] AI-assisted radiologists	[D] Radiologists without software	Detection accuracy ([C] vs [D]; Segmentation accuracy ([A] vs expert radiologists); Growth assessment accuracy ([A] vs expert radiologists); Composition classification accuracy ([C] vs [D]).	Estimated July 2021
[redacted][92]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Reference, country	Title and study design	Population	Index test	Comparator	Outcomes	Completion date
KCT0005065 (2020),[93] Korea	A multi-center, retrospective pivotal trial to evaluate the efficacy of artificial intelligence-based pulmonary nodule detection software 'VUNO Med – Lung CAD' in thoracic CT / Retrospective test accuracy study	N=855 1) Adults at the age 19 or above who had a thoracic CT scan within the period from Jan 2012 to Jun 2018; 2) Patients whose thoracic CT scan showed no or 1 to 5 pulmonary nodules with the long-axis diameter from 4 mm to 30 mm	VUNO Med – Lung CAD (VUNO): [A] Stand-alone software	None	Per lesion sensitivity; Per patient sensitivity; Per patient specificity; Per patient false positive; Per patient false negative; Per lesion false negative	19 November 2019

[A] Stand-alone software; [B] Assisted 2nd-read software; [C] Concurrent software use; [D] Unaided reading.

AUC, Area under the receiver operating curve; CAD, Computer-aided detection; FN, False negative; FP, False positive; LDCT, Low-dose computed tomography; MRMC, Multi-reader, multi-case study; NPV, Negative predictive value; NR, Not reported; PPV, Positive predictive value; TN, True negative; TP, True positive; VLN, Veye Lung Nodules (Aidence).

13.3 Appendix 3: Data extraction tables

EVIDENCE ID	STUDY NAME (Author Year)	EXTRACTOR	CHECKER										
PATIENT SAMPLING ITEMS	PATIENT SAMPLING	**PATIENT CHARACTERISTICS AND SETTING ITEMS**	PATIENT CHARACTERISTICS AND SETTING	**INDEX TEST ITEMS**	INDEX TEST (software-based nodule detection and analysis)	**COMPARATOR ITEMS**	COMPARATOR (no software for nodule detection or analysis)	**REFERENCE STANDARD ITEMS**	REFERENCE STANDARD	**FLOW AND TIMING ITEMS**	FLOW AND TIMING	**NOTES items**	NOTES
A1 Review question relevance *- Q1: Test accuracy and other intermediate outcomes* *- Q2: Clinical effectiveness* *- Q3: Cost effectiveness*		*B1 Setting*		*C1 Index test mode, e.g.* *[A] Stand-alone AI* *[B] 2nd read CAD* *[C] Concurrent CAD*		*D1 Reader details (number, general or thoracic radiologist or other, experience) (continue labelling with [D], [E]... as appropriate)*		*E1 Reference standard - General approach*		*F1 What was the time interval between index and reference tests?*		*G1: Funding*	
A2 Relevant outcomes for DAR		*B2 Location (include name of institution if available)*		*C2 AI name and version/date (label different AI-based index tests with [A], [B], [C]...)*		*D2 Reading conditions (reader study, clinical practice, other details)*		*E2 Reference standard for nodule detection*		*F2 Did all patients receive the same reference standard*		*G2: Publication status*	
A3 Study design *(and description of groups labelled [1] [2] ...)*		*B3 Dates*		*C3 Manufacturer and country*		*D3 Method of nodule detection*		*E3 Reference standard for malignant nodules*		*F3 Was the reference standard chosen based on only one of the index/comparator tests?*		*G3: Source (pre-print or Journal name)*	
A4 Aim of the study		*B4 Indication for CT scan* *- Symptomatic* *- Incidental (with reason)* *- Screening* *- CT surveillance*		*C4 Commercially available / CE mark*		*D4 Method of nodule composition/type*		*E4 Reference standard for benign nodules*		*F4 Missing data*		*G4: Author COI (including any manufacturer affiliations)*	
A5 Study type *1) Stand-alone software compared to nothing* *2) Stand-alone software compared to human* *3) Software-assisted reader compared to unassisted reader* *4) Software-assisted reader compared to nothing* *5) Software use in pathway*		*B5 Patient characteristics* *- Age* *- Gender* *- Ethnicity* *- Smoking*		*C5 AI algorithm details*		*D4 Method of nodule size measurement (segmentation, volume, diameter)*		*E5 Reference standard for nodule composition/type*		*F5 Uninterpretable results*		*G5 Comment*	
A6 Comparative study design: *1) Fully Paired* *2) Randomized* *3) Partially paired with random subset* *4) Partially paired with nonrandom subset* *5) Unpaired nonrandomized* *6) Other (please describe)*		*B6 Nodule characteristics* *- Number of nodules* *- Nodule size* *- Nodule type* *- Nodule shape*		*C6 AI training and tuning details*		*D5 Method of nodule growth rate*		*E6 Reference standard for nodule segmentation and size*		*F6 Indeterminate results*			
A7 Method of participant / CT image selection *- Source* *- Consecutive, random, selected (e.g. enriched), unclear*		*B7 CT image acquisition* *- CT scanner* *- Full or partial chest* *- With or without contrast* *- Acquisition parameters (e.g. dose)* *- Image reconstruction* *- Slice thickness*		*C7 Software functionality:* *- Nodule detection* *- Nodule composition* *- Nodule segmentation/ measurement* *- Growth rate*		*D6 Blinded to reference standard*		*E7 Reference standard for nodule growth rate*		*F7 Statistical analysis*			
A8 Were cases recruited prospectively or retrospectively?		*B8 Comments*		*C8 AI software settings (e.g. threshold)*		*D7 Blinded to the results of any other index tests/comparator tests*		*E8 Was it blind to index test/comparator test*		*F8 Comment*			
A9 Sample size				*C9 Reader details (number, general or thoracic radiologist or other, experience, location)*		*D8 Threshold pre-specified*		*E9 Did it incorporate index test/comparator test*					
A10 Inclusion criteria				*C10 Reading conditions where human readers are part of the test (reader study, clinical practice, other details)*		*D9 Other information available to unassisted reader (e.g. prior CT scans, family history)*		E10 Comments					
A11 Exclusion criteria				*C11 Method for nodule detection*		*D10 Description of a whole read (up to clinical decision on discharge, CT surveillance or further diagnostic investigation)*							
A12 Study flow *- Screened for eligibility* *- Eligible* *- Not eligible (with reasons)* *- Included in study/test set* *- Excluded from study/test set (with reasons)* *- Included in analysis* *- Excluded from analysis (with reasons)*				*C12 Method for nodule composition/type*		*D11 Comments*							
A13 Comment				*C13 Method for segmentation and nodule size measurement (volume, diameter)*									
				C14 Method for nodule growth determination									
				C15 Other information made available to AI system or AI assisted reader (e.g. prior CT scans, family history)									
				C16 Blinded to reference standard									
				C17 Blinded to the results of any index/comparator tests									
				C18 Threshold predefined									
				C19 Description of a whole read (up to clinical decision on discharge, CT surveillance or further diagnostic investigation)									
				C20 Comment									

Table 68. Study level description of the 27 included studies for key question 1

Study, country, design & 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) (3 studies)							
Abadia 2021,[45] USA, Retrospective test accuracy and MRMC study; VA10A prototype	Mixed population: Lung cancer screening, abnormal x-rays, suspicious nodule follow-up, abnormal lung-function tests, respiratory symptoms, or history of lung diseases[1]. Selected 143 patients with least 1 lung condition[1] 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; 1 expert chest radiologist: [C] With concurrent AI (MRMC study); [D] Without AI (MRMC study); [E] Original radiology reports (1 of 5 experienced chest radiologists without AI).	Per-nodule assessment / per-subject assessment: [1] [D] + AI-RAD (2nd read AI): [2] [E] + AI-RAD (2nd read AI)[2] AI-RAD versus radiology reports: [1] [E] + AI-RAD (2nd read AI)[1]	Nodule detection accuracy; Nodule size measurement ([A] vs [D]); Characteristics of nodules (FN, FP); Reading times; Confidence in lung nodule detection.	N/A
Chamberlin 2021,[46] USA, Retrospective test accuracy study,	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 (2 readers)	Nodule detection accuracy; Characteristics of detected nodules.	Quantification of coronary artery calcium volume; prediction of major cardiopulmonary outcomes; false positive analysis

[1] Interstitial lung disease, chronic obstructive lung disease, respiratory bronchiolitis, pulmonary oedema, or pulmonary embolism.
[2] If AI-RAD found additional nodules, the expert radiologist verified if they were TP or FP.

Study, country, design & 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)
VA10A prototype							
Rueckel 2021,[47] 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]). 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
AVIEW LCS+ (Coreline Soft) (4 studies)							
Hwang 2021,[49] South Korea, Before-and-after study, A-view Lungscreen	Screening population: 6,487 consecutive participants (1,821 pre-AI implementation; 4,666 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 2nd read AI [B]; 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;	Nodule size measured on transverse planes vs any maximum plane or maximum orthogonal plane

Study, country, design & 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)
						Technical failure rate.	
Hwang 2021,[48] South 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] 2nd 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 2021,[50] South 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] Assisted 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;

Study, country, design & 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)
							Disagreement between the institutional reading and central review.
Lancaster 2022,[30] 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 (3 experienced chest radiologists); [D] Unaided reader: 2 experienced chest radiologist using other semi-automated volumetric software.	Nodule categorisation: Consensus expert reading (3 radiologists with >10 years of experience and 1 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.	NA
ClearRead CT (Riverain Technologies) (6 studies)							
Singh 2021,[54] USA, MRMC study, ClearRead CT with vessel suppression	Screening population: enriched sample of 123 patients (100 with sub-solid 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); [C.1] Concurrent AI – 2 experienced radiologists reading	Nodule detection: Consensus expert reading (2 readers)	Nodule detection accuracy; Characteristics of detected nodules; Size measurement accuracy;	N/A

Study, country, design & 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)
and nodule detection				AI-VS images (with vessel suppression without automatic nodule detection feature); [C.2] Concurrent AI – 2 experienced radiologists reading AI-AD images (with vessel suppression and autodetection of pulmonary nodules) [D] 2 experienced radiologists reading standard CT images.		Inter-observer agreement to detect the dominant nodule; Technical failure rate; Impact on clinical decision making (change in Lung-RADS category).	
Lo 2018,[52] 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 2 hospitals.	Low dose, contrast and slice thickness unclear	Solid, part solid, ground glass; 5-44 mm	[A] Stand-alone AI; 12 experienced general radiologists: [C] With concurrent AI; [D] Without AI.	Nodule detection: Consensus expert reading (3 readers) assisted by corresponding NLST or source documentations containing radiologic, pathologic, and follow-up reports.	Accuracy of nodule detection; Radiologist reading time.	N/A

Study, country, design & 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 2018, Switzerland,[53] 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 to 366 mm^3	[C] Nodule measurement on vessel suppressed CT images (1 general radiologist with 3 years of experience, 1 resident radiologist) using semi-automatic segmentation software (MM Oncology, Siemens Healthcare) [D] Nodule measurement on standard CT images (1 general radiologist with 3 years of experience, 1 resident radiologist) using semi-automatic 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 2021,[51] Taiwan, MRMC study, ClearReadCT with vessel	Mixed population: 93 clinical routine; 57 screening population.	Low dose (n=57), standard dose (n=93),	Any type; ≤10 mm	[A] Stand-alone AI; 6 chest radiologists - 3 less experienced and 3 experienced: [B] With 2nd read AI;	Nodule detection: Consensus expert reading (2 readers)	Nodule detection accuracy; Radiologist reading time.	N/A

Study, country, design & 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)
suppression and nodule detection	Outcomes for screening population reported seaparately. 150 consecutive cases with lung nodules ≤1 cm or no nodules.	no contrast, 2.5 mm		[C] With concurrent AI; [D] Without software.			
Takaishi 2021,[55] Japan, MRMC study, ClearRead CT for vessel suppression	Mixed population[3]: Unclear how selected, 61 thoracic or thoracic-abdominal CT images conducted at 1 Japanese hospital in September 2019.	Standard dose, no contrast, 5 mm	Solid, ground glass; 4-54 mm diameter	3 general radiologists with 2-8 years of experience: [C] With concurrent AI; [D] Without software.	Nodule detection: Consensus expert reading (2 readers)	Nodule detection accuracy; Radiologist reading time.	N/A
Wan 2020,[56] 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.[4]	Low dose, Unclear contrast	Solid, part-solid, ground glass; ≤2 cm	[A] Stand-alone AI; [D] Consensus of 2 radiologists with 25-38 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	

[3] 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).

[4] 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.

Study, country, design & 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)
						stand-alone AI and unaided reader.	
Contextflow SEARCH Lung CT (contextflow) (1 study)							
Röhrich 2022,[29] Austria, MRMC study, prototype version	Mixed population[5] (Follow-up of a known lung disease, suspected lung disease, incidental): 100 with confirmed diffuse parenchymal lung disease, 8 with inconspicuous chest CT scans from 1 hospital in Austria.	Unclear dose, with or without contrast	Any type	4 radiology residents (2.1 ± 0.7 years of experience) and 4 general radiologists (12 ± 1.8 years of experience) [C] With concurrent AI; [D] Without AI.	Lung nodule detection: 1 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) (3 studies)							
Kozuka 2020,[57] Japan, MRMC study, Version NR	Symptomatic population (suspected cancer): Random 120 chest CT images from 1 hospital in Japan.	Standard dose; no contrast; 1 mm.	Solid, Part-solid, Calcified, Ground glass	[A] Stand-alone AI; 2 less experienced radiologists: [C] With concurrent AI; [D] Without AI.	Nodule detection: Consensus expert reading (3 readers)	Nodule detection accuracy; Radiologist reading time; Characteristics of detected nodules.	N/A
Liu 2019,[58] China	Mixed population: screening and inpatient, convenience sample,	Standard dose or low dose;	Solid, subsolid,	Evaluation 1: [A] Stand-alone AI;	Nodule detection:	Nodule detection accuracy,	AI performance by patient age (evaluation 2)

[5] 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.

Study, country, design & 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)
MRMC study, Software name and version NR	1,129 CT scans from >10 hospitals in China. Evaluation 1: N=1,129; Evaluation 4: N=123 (Batch 1); N=148 (Batch 2).	unclear regarding contrast; 0.8-2.0 mm	calcified, pleural	[D.1] 2 experienced general radiologists without AI. Evaluation 4: 2 experienced general radiologists: [C] With concurrent AI, [D.2] Without AI.	Consensus expert reading (3 readers)	Comparison of AI performance by radiation dose, Radiologist reading time.	and CT manufacturer (evaluation 3)
Zhang 2021,[59] China Retrospective test accuracy and MRMC study, Software version NR	Screening population: 860 consecutive patients from 1 hospital in China (part of NELCIN-B3 project)	Low dose; no contrast; 0.625-1.0 mm	Solid, part-solid, ground glass	1 radiology resident with supervision of 1 experienced radiologist: [C] With concurrent AI (MRMC study: 1 radiology resident and 1 experienced radiologist); [E] Without AI (clinical practice: 14 different radiology residents and 15 different experienced radiologists).	Nodule detection: Consensus expert reading (2 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.							

Study, country, design & 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)
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) (4 studies)							
Cohen 2017,[60] South 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 1 Korean hospital.	Standard dose; no contrast; 0.625 mm	Sub-solid nodules	2 radiologists with 4-5 years of experience: [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 versus MBIR.	N/A
Kim 2018,[61] South Korea, MRMC study, Version 1.2	Surveillance population with applicability concerns: 89 consecutive patients with preoperative CT scan for subsolid nodules	Standard dose; no contrast; 0.625 mm	Sub-solid nodules	2 experienced radiologists: [C] With concurrent AI; [D] Without AI.	No reference standard for nodule size measurement	Diameter measurement: Concordance between readers with and without software;	Diagnostic performance using binary logistic regression analysis for

Study, country, design & 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)
	and subsequent surgical resection at 1 Korean hospital.					Inter-observer variability; Repeatability / reproducibility; Technical failure rate Nodule classification by size of solid portion: Inter-observer variability; Repeatability / reproducibility.	invasive adeno-carcinoma
Hall 2022,[25] 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). [E] Without AI: 1 of 5 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 1 (if needed 2) radiologists for consensus.	Nodule detection accuracy; Lung cancer detection accuracy; Impact on decision making; Radiologist reading time; Software acceptability & experience; Proportion of scans referred for CT surveillance;	N/A

Study, country, design & 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)
						Proportion of scans referred to MDT; Characteristics of missed nodules; % of detected nodules being malignant.	
Jacobs 2021,[62] USA, Denmark, 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	3 experienced radiologists and 4 radiology residents from Denmark and the Netherlands: [C] With concurrent AI; [D] Without AI.	No reference standard	Lung-RADS categorisation: Inter-observer variability; Repeatability / reproducibility. Radiologist reading time; Technical failure rate; Impact on decision-making.	N/A
Veye Lung Nodules (Aidence) (5 studies)							
Blazis 2021,[63] 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	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 (3 readers)	Nodule detection accuracy	N/A

Study, country, design & 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)
	patients included in analyses) from 1 Dutch hospital.						
Hempel 2022,[32] Netherlands; MRMC study; Veye Chest v2.15.3	Mixed indication: 50 patients with incidentally detected nodules or no nodules from 1 Dutch hospital: 5 no nodules, 45 with ≤5 nodules (10 no prior CT, 35 with prior CT). 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	1 experienced chest radiologist and 1 experienced general radiologist: [C] With concurrent AI; [D] Without AI	Risk categorisation based on 2015 BTS grades: 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 2021,[64] Netherlands, Retrospective test accuracy study, Veye Chest versions (25-05-2018), and (18-03-2019)	Mixed indications (ruling out metastasis, follow-up of nodules or other pulmonary abnormalities, other miscellaneous indications): 145 randomly selected CT images performed at 1 Dutch teaching hospital.	Unclear dose; with or without contrast; 1 mm or 3 mm	Solid, sub-solid; 4-30 mm	[A] Stand-alone AI	Nodule detection, composition and measurement: Consensus expert reading (3 readers)	Nodule detection accuracy; Nodule type accuracy (Solid, Sub-solid); Size measurement accuracy; Characteristics of detected (TP, FP) and missed (FN) nodules; Technical failure rate;	N/A

Study, country, design & 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)
						Software acceptability and experience.	
Murchison 2022,[31] UK, MRMC study, Veye Chest version 2.0	Mixed indications (clinical routine mimicking a screening population in age and smoking history[6]): 337 CT scans of 314 subjects from 1 hospital in Edinburgh. [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 sub-solid nodules in original report (n=18).	Standard dose, with (n=22) or without contrast (n=315); 1.0-2.5 mm	Any type; 3-30 mm, 5-30 mm	[A] Stand-alone AI 2 experienced chest radiologists: [C] With concurrent AI, [D] Without AI.	Nodule detection and composition: Majority expert reading (2 index test readers with discrepancies adjudicated by a 3rd experienced chest radiologist); Nodule measurement and growth rate: 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. Technical failure rate. Growth rate: Nodule registration accuracy; Inter-observer variability; Concordance between stand-alone software	N/A

[6] Current smokers, a smoking history and/or radiological evidence of pulmonary emphysema.

Study, country, design & 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)
						and readers without software.	
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
VUNO Med-LungCT AI (VUNO) (1 study)							
Park 2022,[65] USA, Korea, MRMC study, v.1.0.1	Screening population: 200 cases randomly selected from an nodule- and cancer-enriched subset of the NLST database.	Low dose, No contrast	Solid, part-solid, non-solid	[A] Stand-alone AI; 1 resident radiologist and 4 radiologists with 1-20 years of experience: [C] With concurrent AI; [D] Without AI.	Lung cancer detection: NR (same-year positive cancer diagnosis)	Nodule detection and Lung-RADS categorisation: 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.

[a] Where the software evaluated in the study had a different name from that was listed in the NICE final scope, but the company confirmed its relevance.

AI, Articial intelligence; BTS, British Thoracic Society; CT, Computed tomography; FP, False positive; FN, False negative; K-LUCAS, Korean lung cancer screening project; LIDC-IDRI, Lung Image Database Consortium image collection; LSUT, Lung Screen Uptake Trial; Lung-RADS, Lung Imaging 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.

13.4 Appendix 4: Quality assessment

QUADAS-2+QUADAS-C tailored for AI technologies

First author surname and year of publication:

Name of first reviewer: *Name of second reviewer:*

Phase 1: State the review question:

Question 1) What is the accuracy of CT image analysis assisted by software for automated detection and analysis of lung nodules in people undergoing CT scans?

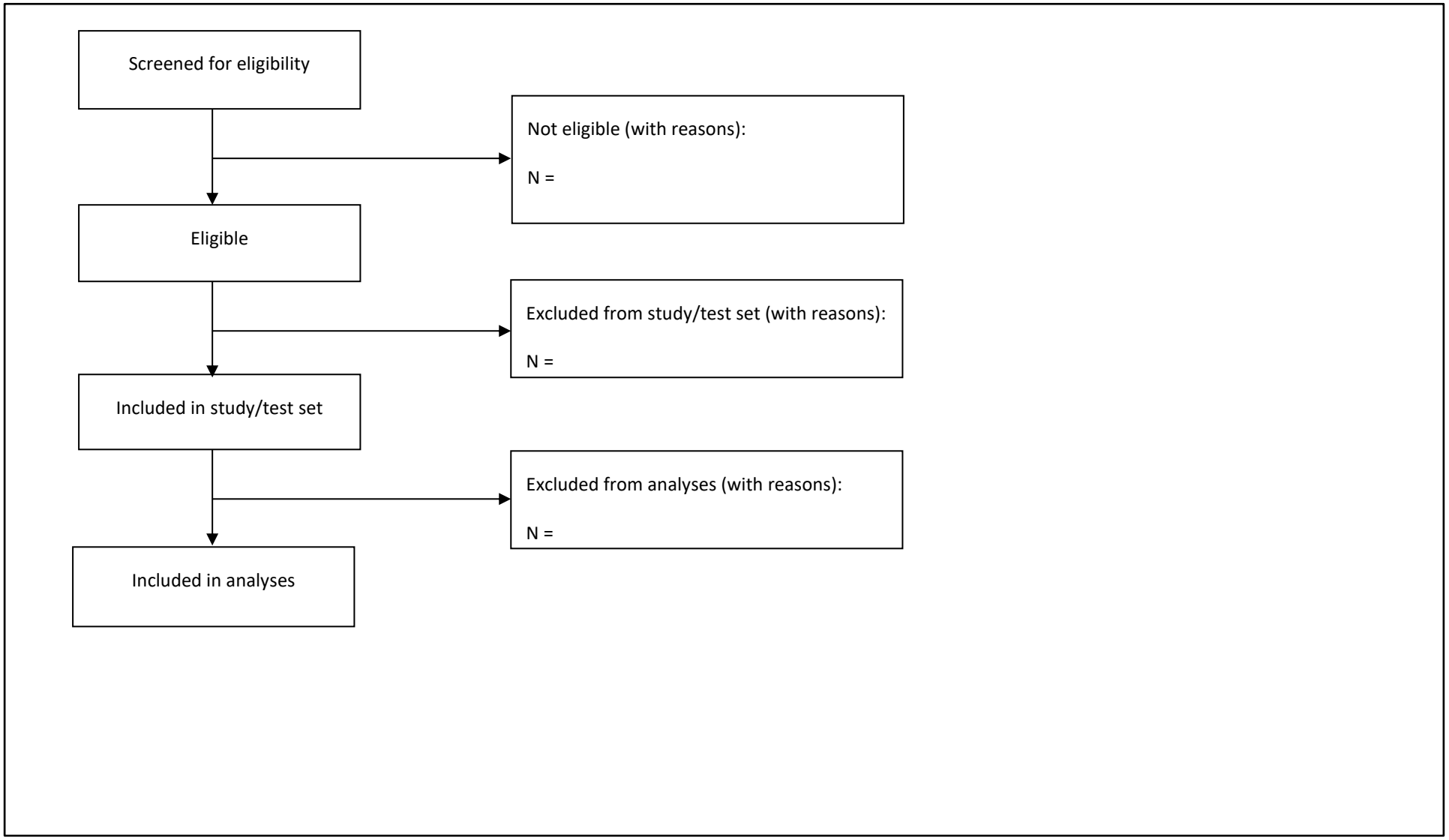
Patients (setting, intended use of index test, presentation, prior testing):
People who have no confirmed lung nodules or lung cancer and who are not having staging investigations or follow-up imaging for primary cancer elsewhere in the body, who have a CT scan that includes the chest: • *for reasons unrelated to suspicion of lung cancer (incidental population);* • *because of signs or symptoms suggestive of lung cancer (symptomatic population);* • *as part of lung cancer screening (screening population);* *People having CT surveillance for a previously identified lung nodule (surveillance population).*
Index test(s) (including human comparators):

• *CT scan review by*
  ○ *Index test [A]: any of the specified software <u>alone</u>;*
  ○ *Index test [B]: a radiologist or another healthcare professional using any of the specified software as <u>2nd reader</u>;*
  ○ *Index test [C]: a radiologist or another healthcare professional with <u>concurrent use</u> of any of the specified software;*
  ○ *Index test [D]: a radiologist or another healthcare professional <u>without</u> software assistance.*

Reference standard and target condition:		
• *Target condition: Lung cancer (or lung nodules)* • *Reference standard for nodule detection and nodule type: Experienced chest radiologist reading (single reader or consensus/majority reading of more than one reader).* • *Reference standard for nodule size measurement and nodule growth assessment: Experienced radiologist reading (single reader or consensus/mean size or mean growth rate) or measurement of nodules after excision.* • *Reference standard for malignant/benign nodules:* *Malignant: Histological analysis of lung biopsy or health record review;* *Benign: CT surveillance (imaging follow-up) without significant growth, follow-up without diagnosis of lung cancer.*		
Comparative review question (only fill this part for comparative diagnostic accuracy studies with at least 2 index tests, add more rows for index tests if needed)		
Patients:		
Index test [A] (stand-alone software)		
Index test [B] (second-read CAD)		
Index test [C] (concurrent CAD)		
Index test [D] (human reader without software)		
Reference standard and target condition:		
Comparative study design		
Which of the following study designs does the primary study most strongly resemble? *#1 Fully Paired* *#2 Randomized* *#3 Partially paired with random subset* *#4 Partially paired with nonrandom subset* *#5 Unpaired nonrandomized* *Other (please describe the study design):*		*#1 If participants receiving index test [A] and index test [B] are identical (all participants receive all index test).* *#2 If each participant is randomized to receive either one index test or the other.* *#3 If participants are randomly selected either to receive one index test or to undergo both index tests.* *#4 If a nonrandom mechanism is used to decide whether participants receive one or both index tests.*

		#5 If participants receive only one of the index tests without randomization. *Other (please describe study design)*

Phase 2: Draw a flow diagram for the primary study *(adapt template below or copy from paper)*



Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

DOMAIN 1: PATIENT SELECTION **A. Risk of Bias**					
Describe methods of patient selection:					
Single test accuracy (QUADAS-2)	**Answers for Index test [A]**	**Answers for Index test [B]**	**Answers for Index test [C]**	**Answers for Index test [D]**	**Guidance**
1.1 Was a consecutive or random sample of patients enrolled?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Consecutive (e.g. ALL patients in a certain time period) or random sampling – yes. If not stated – unclear. Other studies (selected or enriched sample) – no.
1.2 Was a case-control design avoided?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Studies with single set of inclusion criteria for study admission (1-gate); can be prospective or retrospective sampling – yes. If not stated – unclear. Studies with separate sampling schemes for diseased (cases) and non-diseased individuals (controls) (2-gate), e.g. if the samples are selected according to knowing whether people do or do not have lung nodules or lung cancer – no.
1.3 Did the study avoid inappropriate exclusions?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Use this to flag up that groups of people / CT images were systematically excluded who should not have been as their exclusion narrows the spectrum of diseased or non-diseased (e.g. exclusion of 'easy to diagnose' or 'difficult to diagnose' patients). Systematic exclusion of CT images that could not be processed by the software (e.g. segmentation failures), even if reported in the paper as 'Exclusions from the study', should be ignored in this domain but scored in the 'Flow & timing' domain. If nothing is said and consecutive or random sampling – yes.

					If non-consecutive sampling issue and nothing said – unclear. Exclusions by nodule number per image or unjustified (not based on management guidelines or minimal software manufacturer threshold) exclusion of certain nodule sizes) – no. Systematic exclusion of patients with other non-nodule related lung pathology that can mimic or mask lung nodules ('difficult to read' CT images; e.g. severe pulmonary fibrosis, diffuse bronchiectasis, extensive inflammatory consolidation, pneumothorax, and massive pleural effusion) – no. Systematic exclusion of 'easy to read' CT images (e.g. patients without other, non-nodule related lung conditions). – no.
1.4 Were the people/CT images included in the study independent of those used to train the AI algorithm?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	<u>For test set studies</u>, this translates as "Has the test set been clearly described as an external (geographically) validation set?" Any internal validation (e.g. split sample, cross-validation) or temporal validation – no. No details stated about the training set and tuning set - unclear. External geographical validation (Test set was sample from a different centre; can be in another country or the same country) – yes. For index test [D] without AI software involvement – NA. <u>For prospective applied studies in a clinical context</u>: If the study is located at different centre(s) to those who provided CT images used to train and tune the AI algorithm – yes. If not stated – unclear. If there is any overlap in patients or CT images – no. For index test [D] without AI software involvement – NA.
1.5 Could the selection of patients have introduced bias? **(Score HIGH if 'no' to any question.)**	**RISK:** **LOW** **HIGH** **UNCLEAR** **NA**	**RISK:** **LOW** **HIGH** **UNCLEAR** **NA**	**RISK:** **LOW** **HIGH** **UNCLEAR** **NA**	**RISK:** **LOW** **HIGH** **UNCLEAR** **NA**	All signalling questions answered with 'yes' – LOW. At least one signalling question answered with 'no' – HIGH. Only 'yes' and 'unclear' answers – UNCLEAR.
Comparative accuracy (QUADAS-C)	**Answers for the test comparison**		**Guidance**		

C1.1 Was the risk of bias for each index test judged 'low' for this domain?	Yes No	'yes' if the risk of bias judgment for single test accuracy (question 1.5 in QUADAS-2) was 'low' for each index test.
C1.2 Was a fully paired or randomized design used?	Yes No Unclear	'yes' if one of the following methods was used for allocating patients to index tests: (1) each patient receiving all of the index tests (fully paired design) or (2) random allocation of patients to one of the index tests (randomized design).
C1.3 Was the allocation sequence random?	Yes No Unclear NA	Only applicable to randomized designs 'yes' if the study generated a truly random allocation sequence, for example, computer-generated random numbers and random number tables.
C1.4 Was the allocation sequence concealed until patients were enrolled and assigned to index tests?	Yes No Unclear NA	Only applicable to randomized designs 'yes' if the study used appropriate methods to conceal allocation, such as central randomization schemes and opaque sealed envelopes.
C1.5 Could the selection of patients have introduced bias in the comparison?	**RISK:** **LOW** **HIGH** **UNCLEAR**	Risk of bias can be judged 'low' if questions C1.1 to C1.4 were answered 'yes' (questions C1.3 and C1.4 are only applicable to randomized designs). If at least one question was answered 'no', users should consider a 'high risk of bias' judgment if the bias associated with the design feature is of such concern that the entire domain is deemed problematic. If C1.2 was answered 'no', strongly consider 'high risk of bias'.
B. Concerns regarding applicability		
Describe included patients (prior testing, presentation, intended use of index test and setting)**:**		

Single test accuracy (QUADAS-2)	**Answers for Index test [A]**	**Answers for Index test [B]**	**Answers for Index test [C]**	**Answers for Index test [D]**	**Guidance**
Please fill in one of the following four rows based on the assessed population (Incidental, Symptomatic, Screening, Surveillance)					

Is there concern that the included patients *(INCIDENTAL)* do not match the review question?	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	High concerns if: - Not a consecutive or random sample of patients / CT images; - Enriched sample (e.g. in-/exclusion by nodule number, nodule type and nodule size, respectively); - Inclusion/Exclusion by age; - Patients not representative of UK target population (study not performed in UK or other North-Western European country); - >10% of included people have a different indication for the CT scan than the target population; - CT image acquisition details different to UK practice for this target population (UK practice: standard dose; slice thickness ≤2.0mm), with or without contrast).
Is there concern that the included patients *(SYMPTOMATIC)* do not match the review question?	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	High concerns if: - Not a consecutive or random sample of patients / CT images; - Enriched sample (e.g. in-/exclusion by nodule number, nodule type and nodule size, respectively); - Inclusion/Exclusion by age; - Patients not representative of UK target population (study not performed in UK or other North-Western European country); - >10% of included people have a different indication for the CT scan than the target population; - CT image acquisition details different to UK practice for this target population (UK practice: slice thickness ≤2.0mm; standard dose; with or without use of contrast).
Is there concern that the included patients *(SCREENING)* do not match the review question?	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	**CONCERN:** **LOW** **HIGH** **UNCLEAR** **NA**	High concerns if: - Not a consecutive or random sample of patients / CT images; - Enriched sample (e.g. in-/exclusion by nodule number, nodule type and nodule size, respectively); - Age not between 55-75 years; - Not at high risk for lung cancer (e.g. current or former smokers, identified by questionnaire or other risk prediction model); - Patients not representative of UK target population (study not performed in UK or other North-Western European country); - >10% of included people have a different indication for the CT scan than the

					target population; - CT image acquisition details different to UK practice for this target population (UK practice: slice thickness ≤2.0mm, low dose [<2 less mSV per scan], no contrast).
Is there concern that the included patients *(SURVEILLANCE)* do not match the review question?	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	High concerns if: - Not a consecutive or random sample of patients / CT images; - Enriched sample (e.g. in-/exclusion by nodule number, nodule type and nodule size, respectively); - Inclusion/Exclusion by age; - Nodule size: <5mm or >30mm maximal diameter; <80mm^3; - Patients not representative of UK target population (study not performed in UK or other North-Western European country); - >10% of included people have a different indication for the CT scan than the target population; - CT image acquisition details different to UK practice for this target population (UK practice: low radiation dose CT, slice thickness ≤2.0mm, with or without contrast).

DOMAIN 2: INDEX TEST(S)

If more than one index test (e.g. different functions of the software) or a human comparator was used, please complete for each test.

A. Risk of Bias

Describe the index test and how it was conducted and interpreted:

Single test accuracy (QUADAS-2)	**Answers for Index test [A]**	**Answers for Index test [B]**	**Answers for Index test [C]**	**Answers for Index test [D]**	**Guidance**

2.1 Were the index test results interpreted without knowledge of the results of the reference standard? (Requires 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)	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	[A] For index tests where AI software is used standalone (<u>without any human element</u>): - AI system has not previously been trained on these CT images or learned from these CT images or other CT images from the same patients – yes. - If data from the same dataset was used for training/tuning the software – no. - If repeat use of the same CT images or other CT images from the same patients within the same or previous studies – no (unless explicit that the AI algorithm was pre-set and did not change upon repeat use, and the study did not select one of several AI systems based on use with the same cases). - If nothing is said about training/tuning – unclear. - If not explicit that there has been no repeat use within the same or previous studies – unclear. [B] [C] [D] For index tests <u>where a human is involved</u> (either unassisted human read comparator, software-assisted human readers e.g. second-read CAD or concurrent CAD): - Requires clear statement of blinding, or clear temporal relationships where the human read occurred before the reference standard – yes. - If nothing is said and no clear temporal relationship – unclear. - If clearly unblinded – no.
2.2 If a threshold was used, was it pre-specified?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	[A] If the AI software threshold was pre-set by company or clearly pre-specified in methods (e.g. sensitivity and/or FP rate threshold or nodule size threshold) – yes. If AI software threshold clearly not pre-set by company or pre-specified in methods – no. Using sensitivity / specificity of the unaided reader as benchmark using the same dataset – no. Reporting AI software performance at various threshold settings or in a ROC curve – no. If nothing is said – unclear. No threshold used – NA. [B] [C] [D] Unaided or software-assisted human readers detecting nodules:

					Use of a pre-specified nodule size or volume threshold – yes. If a threshold is used but it is unclear if it was pre-specified – unclear. Nodule size or volume threshold not pre-specified – no. No threshold used – NA.
2.3 Where human readers are part of the test, were their decisions made in a clinical practice context? (i.e. avoidance of the laboratory effect)	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	This question has been added. [A] NA [B] [C] [D] If the readers made decisions in the clinical context, and those decisions were used to decide whether to discharge or recall patients (either prospectively as part of a trial or test accuracy study or retrospective studies using the original decision) – yes. If readers examined a test set (of any prevalence) outside clinical practice, or any other context likely to result in the laboratory effect (that their reading result is not influencing a patient's diagnosis) – no.
2.4 Could the conduct or interpretation of the index test have introduced bias?	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	All signalling questions answered with 'yes' – LOW. At least one signalling question answered with 'no' – HIGH Only 'yes' and 'unclear' answers – UNCLEAR.
Comparative accuracy (QUADAS-C)	**Answers for the test comparison**				**Guidance**
C2.1 Was the risk of bias for each index test judged 'low' for this domain?	Yes No				'yes' if the risk of bias judgment for single test accuracy (question 2.5 in QUADAS-2) was 'low' for each index test.
C2.2 Were the index test results interpreted without knowledge of the results of the other index test(s)?	Yes No Unclear NA				Only applicable if patients received multiple index tests (fully or partially paired designs) 'yes' if index test [A] was interpreted blind to the results of index test [B] and vice versa. Blinding is not necessary if none of the index tests involve subjective interpretation.
C2.3 Is undergoing one index test unlikely to	Yes No				Only applicable if patients received multiple index tests (fully or partially paired designs) 'yes' if one index test cannot influence or interfere with the results of subsequently performed index

affect the performance of the other index test(s)?	Unclear NA			test(s). Examples of such influence or interference include distortion of sampling area (biopsies) and patient fatigue (questionnaires).	
C2.4 Were the index tests conducted and interpreted without advantaging one of the tests?	Yes No Unclear			'yes' if there were no differences in the conduct and interpretation between the index tests that may unfairly benefit one of the tests. An example of such a difference is when index test A was performed by an expert and index test B by a nonexpert. Differences between tests that reflect clinical practice may be acceptable, in which case 'yes' is appropriate.	
C2.5 Could the conduct or interpretation of the index tests have introduced bias in the comparison? (Score HIGH if 'no' to any question.)	**RISK: LOW HIGH UNCLEAR**			Risk of bias can be judged 'low' if signaling questions C2.1 to C2.4 were answered 'yes' (C2.2 and C2.3 are only applicable to fully or partially paired designs). If at least one question was answered 'no', users should consider a 'high risk of bias' judgment if the bias associated with the design feature is of such concern that the entire domain is deemed problematic.	
B. Concerns regarding applicability					
Single test accuracy (QUADAS-2)	**Answers for Index test [A]**	**Answers for Index test [B]**	**Answers for Index test [C]**	**Answers for Index test [D]**	**Guidance**
Is there concern that the index test(s) or comparator, its conduct, or interpretation differ from the review question?	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	High concerns if: <u>For all functionalities:</u> - [A] [B] [C] Any prototype versions that did not later become the commercially available version (e.g. applicability not confirmed by the company). - Integration of software into pathway not applicable to UK (e.g. standalone AI performance [A] instead of concurrent [C] or second-read [B] CAD; for [B] and [C] – more than 1 human reader involved per read); - Human comparator [D] not applicable to the UK (e.g. human double reading instead of single human reader); - Human reader's experience and/or specialty not representative of UK clinical practice (The training for radiologists is 5 years. After that time they are considered "fully trained".) for this target population; - [B] [C] [D] Reader had no access to maximum intensity projections (MIP) and/or multiplanar reformations (MPR). <u>Nodule detection:</u>

					- Study did not use a pre-specified nodule size threshold similar to the UK 2015 BTS guidelines (e.g. ≥5mm maximum axial diameter or ≥80mm^3); - [A] CAD false positive rate set to >2 per case. <u>Nodule type determination:</u> - Other nodule types used than in the BTS guidelines (nodule type should be classified as solid, part-solid or pure ground glass nodules). <u>Nodule size measurement (volume/diameter):</u> - Nodules should be measured using semi-automated volumetry. Where volumetry segmentation is not possible or judged to be inaccurate, maximal axial manual diameter measurements should be recorded, excluding any spiculation. Manual adjustment of volumetric analysis should be avoided as this may introduce unquantified variability.

DOMAIN 3: REFERENCE STANDARD **A. Risk of Bias**					
Describe the reference standard and how it was conducted and interpreted:					
Single test accuracy (QUADAS-2)	**Answers for Index test [A]**	**Answers for Index test [B]**	**Answers for Index test [C]**	**Answers for Index test [D]**	**Guidance**
3.1 Is the reference standard likely to correctly classify the target condition?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Lung cancer: Histopathology after biopsy/excision – yes. Medical records – no. Benign nodules: Histopathology after biopsy/excision; <u>For solid nodules:</u> CT surveillance for at least 2 years with stable diameter or stable (or VDT>600 days) after 1 year on volumetry;

					<u>For subsolid nodules:</u> resolved at CT scan after 3 months or CT surveillance for at least 4 years without growth or altered morphology; At least 2 year follow-up without lung cancer diagnosis – yes. Nodule detection / nodule type / nodule pairs; No reference standard in in vivo studies: will accept majority or consensus reading of (at least) 3 experienced thoracic – yes. Less than 3 experienced thoracic radiologist – no. Nodule size: Measurement of nodule size after nodule excision or consensus/average size measurement of (at least) 3 experienced thoracic radiologists – yes.
3.2 Were the reference standard results interpreted without knowledge of the results of the index test?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Malignant / benign nodules: For retrospective studies if the original human reader is used as comparator test – no. For prospective studies if the investigators did not blind the clinicians undertaking the follow up tests to which index test examined the CT images - no. For retrospective studies where readers read CT scans prospectively (reader study) – yes. Nodule detection / nodule type / nodule pairs / nodule size: If the reference standard reader(s) performed their read prior to the index test(s) – yes. If reference standard reader(s) are blinded to AI and human reader results – yes. If reference standard reader(s) are part of the index test(s) or not blinded to index test markings / decisions – no.
3.3 Could the reference standard, its conduct, or its interpretation have introduced bias?	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	All signalling questions answered with 'yes' – LOW. At least one signalling question answered with 'no' – HIGH. Only 'yes' and 'unclear' answers – UNCLEAR.

Comparative accuracy (QUADAS-C)	Answers for the test comparison	Guidance
C3.1 Was the risk of bias for each index test judged 'low' for this domain?	Yes No	'yes' if the risk of bias judgment for single test accuracy (question 3.3 in QUADAS-2) was 'low' for each index test.
C3.2 Did the reference standard avoid incorporating any of the index tests?	Yes No Unclear	'Incorporation' means that an index test is part of the reference standard. This question is not about whether the reference standard results were interpreted without knowledge of the index test results. 'yes' if none of the index tests were part of the reference standard. Note that this issue is different from blinding (signaling question 3.2 in QUADAS-2).
C3.3 Could the reference standard, its conduct, or its interpretation have introduced bias in the comparison?	**RISK: LOW HIGH UNCLEAR**	Risk of bias can be judged 'low' if signaling questions C3.1 and C3.2 were answered 'yes'. If at least one question was answered 'no', users should consider a 'high risk of bias' judgment if the bias associated with the design feature is of such concern that the entire domain is deemed problematic.

B. Concerns regarding applicability

Single test accuracy (QUADAS-2)	Answers for Index test [A]	Answers for Index test [B]	Answers for Index test [C]	Answers for Index test [D]	Guidance
Is there concern that the target condition as defined by the reference standard does not match the review question?	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	**CONCERN: LOW HIGH UNCLEAR NA**	High concerns if: <u>Malignant/benign nodules:</u> - Different length of CT surveillance (e.g. solid nodules: <2 years with diameter measurements or <1 year with volume measurements; non-resolved sub-solid nodules <4 years); - Diagnosis of cancer not by pathology of biopsied/resected nodules; - No follow-up for at least two years for patients with nodules who are not receiving CT surveillance or biopsy/excision. <u>"Actionable" nodule present/absent:</u> - Different nodule size to BTS 2015 guideline definition ("actionable nodule" is ≥5 mm maximum axial diameter or ≥80 mm^3). <u>Nodule type:</u> - Other types used than in the BTS 2015 guidelines (nodule type should be classified as solid, part-solid or pure ground glass nodules). <u>Nodule size measurement (volume/diameter):</u> - Nodule size should be measured as volume or, if volumetry segmentation is not

					possible, as maximum axial diameter. <u>Nodule pairs:</u> - NA

DOMAIN 4: FLOW AND TIMING **Risk of Bias**					
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):					
Describe the time interval and any intervention between index tests(s) and reference standard:					
Single test accuracy (QUADAS-2)	**Answers for Index test [A]**	**Answers for Index test [B]**	**Answers for Index test [C]**	**Answers for Index test [D]**	**Guidance**
4.1 Did all patients receive a reference standard?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	<u>Malignant / benign nodules:</u> If any patients who should have received a biopsy/resection, other follow-up tests and/or CT surveillance after index test positive results did not receive one or results were unavailable – no. If index test negative patients were not followed up for at least one year (pragmatic threshold) to confirm absence of lung cancer – no. <u>Nodule detection / nodule type / detection of nodule pairs:</u> If ALL CT images are assessed by expert reading as reference standard - yes.
4.2 Did all patients receive the same reference standard?	Yes No Unclear	Yes No Unclear	Yes No Unclear	Yes No Unclear	Need to give separate answers for detection of lung cancer, nodule detection, nodule composition or detection of nodule pairs.

	NA	NA	NA	NA	<u>For nodule detection, nodule composition, detection of nodule pairs:</u> If all CT images received the SAME reference standard (e.g. consensus expert reading) - yes. <u>Malignant / benign nodules:</u> Usually NO – all studies will necessarily have differential verification, because not all patients can or should be biopsied.
4.3 Were all patients included in the analysis?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	If there were significant exclusions (>10%; cut-off determined pragmatically) after the point of selecting the cohort, for example indeterminate results (e.g. segmentation failures) or losses to follow up – no. If the number of excluded CT images after the point of selecting the test set / study sample is not reported – unclear. If there were <10% of CT images excluded from the analyses – yes.
4.4 If there were exclusions from the analysis, has it been reported how many were due to software processing failures?	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	Yes No Unclear NA	This signalling question was added. If the number of CT images excluded due to software processing failures (e.g. segmentation failures) has been reported – yes. If it is unclear if there were any exclusions from the analysis – unclear. If the number of CT images excluded due to software processing failures (e.g. segmentation failures) has not been reported – no. Unaided readers [D] or no exclusions from the analysis – NA.
4.5 Could the patient flow have introduced bias?	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	**RISK: LOW HIGH UNCLEAR NA**	All signalling questions answered with 'yes' – LOW. At least one signalling question answered with 'no' – HIGH. Only 'yes' and 'unclear' answers – UNCLEAR.
Comparative accuracy (QUADAS-C)	**Answers for the test comparison**		**Guidance**		
C4.1 Was the risk of bias for each index test judged 'low' for this domain?	Yes No		'yes' if the risk of bias judgment for single test accuracy (question 4.5 in QUADAS-2) was 'low' for each index test.		
C4.2 Was there an appropriate interval between the index tests?	Yes No		For many index tests, 'appropriate' would constitute performing the tests at the same time after patient enrolment. This excludes the possibility of disease progression or change in patient management. Some index		

	Unclear	tests have different 'diagnostic windows' and are ideally performed at different timepoints; subject-matter expertise is required to determine this.
C4.3 Was the same reference standard used for all index tests?	Yes No Unclear	'yes' if either (1) a single reference standard was used in all patients or (2) multiple reference standards were used (e.g., either surgery or follow-up) and these reference standards were the same for patients receiving index test [A] and patients receiving index test [B].
C4.4 Are the proportions and reasons for missing data similar across index tests?	Yes No Unclear	Missing data occurs if test results are unavailable, invalid, inconclusive, or if patients are excluded from the analysis. 'yes' if there is no missing data, or if the proportion and reasons for missing data are similar for index test [A] and index test [B].
C4.5 Could the patient flow have introduced bias in the comparison?	**RISK:** **LOW** **HIGH** **UNCLEAR**	Risk of bias can be judged 'low' if signaling questions C4.1 to C4.4 were answered 'yes'. If at least one question was answered 'no', users should consider a 'high risk of bias' judgment if the bias associated with the design feature is of such concern that the entire domain is deemed problematic.

COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument – Part B[24]

Standards for studies on reliability and/or studies on measurement error

	Design requirements	very good	adequate	doubtful	inadequate	NA
1	Were patients stable in the time between the repeated measurements on the construct to be measured?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	Na
2	Was the time interval between the repeated measurements appropriate?	Yes		Doubtful, OR time interval not stated	No	Na
3	Were the measurement conditions similar for the repeated measurements – except for the condition being evaluated as a source of variation?	Yes (evidence provided)	Reasons to assume standard was met, OR change was unavoidable	Unclear	No (evidence provided)	Na
4	Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
5	Did the professional(s) assign scores or determine values without knowledge of the scores or values of other repeated measurement(s) in the same patients?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
6	Were there any other important flaws in the design or statistical methods of the study?	No		Minor methodological flaws	Yes	

Reproduced from Mokkink et al. BMC Med Res Methodol 2020;20:293[24] under Creative Commons Attribution 4.0 International License.

Statistical methods – Reliability studies

Statistical methods	very good	adequate	doubtful	inadequate
7 For continuous scores: was an intraclass correlation coefficient (ICC) calculated?	ICC calculated; the model or formula was described, and matches study design and the data	ICC calculated but model or formula was not described or does not optimally match the study design OR Pearson or Spearman correlation coefficient calculated WITH evidence provided that no systematic difference between measurements has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic difference between measurements has occurred OR WITH evidence provided that systematic difference between measurements has occurred	
8 For ordinal scores: was a (weighted) kappa calculated?	Kappa calculated; the weighting scheme was described, and matches the study design and the data	Kappa calculated, but weighting scheme not described or does not optimally match the study design		
9 For dichotomous/nominal scores: was Kappa calculated for each category against the other categories combined?	Kappa calculated for each category against the other categories combined			

Reproduced from Mokkink et al. BMC Med Res Methodol 2020;20:293[24] under Creative Commons Attribution 4.0 International License.

Statistical methods – Studies on measurement error

Statistical methods	**very good**	**adequate**	**doubtful**	**inadequate**
7 For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated?	SEM, SDC, LoA or CV calculated; the model or formula for the SEM/SDC is described; it matches the reviewer constructed research question and the data	SEM, SDC, LoA or CV calculated, but the model or formula is not described or does not optimally match the reviewer constructed research question* and evidence provided that no systematic difference has occurred	$SEM_{consistency}$ $SDC_{consistency}$ or LoA or CV calculated, without knowledge about systematic difference or with evidence provided that systematic difference has occurred	SEM calculated based on Cronbach's alpha, or using SD from another population
8 For dichotomous/nominal/ordinal scores: Was the percentage specific (e.g. positive and negative) agreement calculated?	% specific agreement calculated	% agreement calculated		

Reproduced from Mokkink et al. BMC Med Res Methodol 2020;20:293[24] under Creative Commons Attribution 4.0 International License.

13.5 Appendix 5: Additional evidence on test accuracy of stand-alone AI and other evidence from non-comparative studies

13.5.1 Accuracy for detecting any nodules

13.5.1.1 Stand-alone AI vs unassisted reader (4 studies)

- **Symptomatic population (1 study)**

Kozuka 2020,[57] Japan - InferRead CT Lung (Infervision)

Kozuka et al.[57] randomly selected 120 chest CT images (117 cases included in analysis) from cases with lung cancer suspicion. 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 89.9-98.5%) for stand-alone AI and 68.0% (95% CI 61.4-74.1%) for the pooled unaided readers. Per-patient specificity was 83.3% (95% CI 35.9-99.6%) for stand-alone AI and 91.7% (95% CI 61.5-99.8%) for the pooled unaided readers. Per-nodule sensitivity was 70.3% (95% CI 66.8-73.5%) for stand-alone AI and 20.9% (95% CI 18.8-23.0%) for the pooled unaided readers. Stand-alone AI had a PPV of 57.9% (95% CI 54.6-61.1%), and the pooled unaided readers' PPV was 70.5% (95% CI 66.0-74.7%).

- **Incidental population (1 study)**

Rueckel 2021,[47] Germany - AI-Rad Companion (Siemens Healthineers)

Rueckel et al.[47] 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 FP 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 PPV of stand-alone AI was 20.0% (13/65).

- **Mixed population (2 studies)**

Abadia 2021,[45] USA - AI-Rad Companion (Siemens Healthineers)

Abadia et al.[45] performed a retrospective test accuracy and MRMC study using a case control dataset (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 FP detections per image (38/103). The unaided expert reader correctly detected 90.8% (366/403) nodules with no FP detections as per definition of the reference standard.

Liu 2019,[58] China – InferRead CT Lung (Infervision)

Liu et al.[58] included 1,129 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% (4,481/6,363) for stand-alone AI and 48.6% (6,179/12,726) for the two pooled unaided readers. The false positive rate for stand-alone AI was 46.5% (3,894 FP/8,375 detected nodules) and an average 3.4 per scan (3,894 FP per 1,129 scans), respectively. Using a FROC curve, the performance of stand-alone AI was demonstrated: at an average of one FP detection per scan, the per-nodule sensitivity was 74%. Sensitivity reached a maximum of 86% with an average of eight FP detections per scan.

13.5.1.2 Non-comparative results (6 studies)

Six studies[28, 45, 49, 56, 63, 64] evaluated **accuracy for detecting any nodules by stand-alone AI** without a comparator (see **Figure 9**). Of these, one included a screening population,[49] and five included mixed

populations.[28, 45, 56, 63, 64] Key characteristics and findings of studies with non-comparative outcomes are shown in **Table 9**.

- **Screening population (1 study)**

Hwang 2021a,[49] South Korea - AVIEW LCS+ (Coreline Soft)

Hwang et al.[49] included 4,666 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% (2,147/4,280; 95% CI 48.7–51.7%) 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% (5,981/8,128) of software-detected nodules as false positives (average 1.51 FP detections per image).

- **Mixed population (5 studies)**

[redacted]

Wan 2020,[56] Taiwan - ClearRead CT (Riverain Technologies)

Wan et al.[56] 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 FP rate was not reported.

Abadia 2021,[45] USA - AI-Rad Companion (Siemens Healthineers)

Abadia et al.[45] performed a retrospective test accuracy and MRMC study using a case control dataset (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 2021,[63] Netherlands - Veye Lung Nodules (Aidence)

Blazis et al.[63] 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 FP 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 2021,[64] Netherlands - Veye Lung Nodules (Aidence)

Martins et al.[64] 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, follow-up of other pulmonary abnormalities, and 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 sub-solid nodules and 100.0% (2/2) of mixed (solid/sub-solid) nodules correctly detected. The false positive rate for the detection of all nodules was 38.5% (average 1.04 FP per scan).

13.5.2 Accuracy for detecting actionable nodules

13.5.2.1 Stand-alone AI vs unassisted reader (2 studies)

- **Symptomatic population (1 study)**

Kozuka 2020,[57] Japan - InferRead CT Lung (Infervision)

Kozuka et al.[57] 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 (one and five 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). False positive rate has not been reported.

- **Mixed population (1 study)**

Liu 2019,[58] China – InferRead CT Lung (Infervision)

Liu et al.[58] included 1,129 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 sub-solid nodules >5 mm combined was 84.1% (581/691) for stand-alone AI and 73.4% (1,015/1,382) for the pooled unassisted readers (no level of significance reported).

13.5.2.2 Non-comparative results (2 studies)

Two studies[28, 46] evaluated the **accuracy for detecting actionable nodules by stand-alone AI**. Of these, one included a screening population,[46] and one included a mixed population.[28]

- **Screening population (1 study)**

Chamberlin 2021,[46] USA - AI-Rad Companion (Siemens Healthineers)

Chamberlin et al.[46] 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 FP/scan) was observed.

- **Mixed population (1 study)**

[redacted]

13.5.3 Accuracy for detecting malignant nodules

13.5.3.1 Stand-alone AI vs unassisted reader (No study)

No data available.

13.5.3.2 Non-comparative results (3 studies)

Three studies[25, 49, 56] evaluated **accuracy for detecting malignant nodules by stand-alone AI[49, 56] or with concurrent software use**.[25] Of these, two included a screening population,[25, 49] and one included a mixed population.[56]

- **Screening population (2 studies)**

Hwang 2021a,[49] South Korea - AVIEW LCS+ (Coreline Soft)

Hwang et al.[49] included 4,666 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 70.4% (19/27; 95% CI 49.8-86.2%) confirmed cancer nodules.

Hall 2022,[25] UK - Veolity (MeVis)

Hall's study[25] included 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 (1 study)**

Wan 2020,[56] Taiwan - ClearRead CT (Riverain Technologies)

Wan et al.[56] performed a retrospective analysis of 75 pathology-proven nodules (≤2 cm; 28 benign, 47 malignant) in 50 patients from hospitals in Taiwan. The study reported a sensitivity of 93.6% (44/47; 95% CI 82.5–98.7%) for detecting of malignant nodules by stand-alone AI. The specificity was 39.3% (11/28; 95% CI 21.5–59.4%).

13.5.4 Nodule type determination

13.5.4.1 Accuracy for nodule type determination

a) Non-comparative results (1 study)

Two studies evaluated the accuracy of stand-alone AI-based software (Veye Chest, Aidence) to determine nodule type.[31, 64] The indication for the chest CT scan was mixed in both studies. The overall accuracy of the composition algorithm for discriminating sub-solid from solid nodules 94.2% - 95.0% (**Table 69**). Based on the data by Murchison et al.[31], the [redacted]

a) Non-comparative results (2 studies)

Mixed population – Veye Chest (Aidence) (2 studies)

Both studies used the software Veye Chest from Aidence in stand-alone mode and compared the findings to a reference standard of consensus reading of two radiologists, with discrepancies resolved by a third radiologist (majority consensus).

Murchison et al.[31] used two composition classes (solid or sub-solid) 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 (**Table 69**). Accordingly, the sensitivity and specificity to determine the composition of sub-solid nodules was 68.4% and 98.8%. The overall accuracy for determining the composition (solid or sub-solid) of a pulmonary nodule was 94.2% (360/382), and the kappa was 0.77.

Martins Jarnalo et al.[64] 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 sub-solid by the radiologists, whereas another two were determined solid by the software and mixed solid/sub-solid by the radiologists. Using three composition classes (solid, sub-solid, 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 sub-solid nodules was 84.6% and 100.0% (**Table 69**). The composition of the two mixed (solid and sub-solid) nodules could not be correctly detected by the software as its composition algorithm can only allocate one composition class (solid or sub-solid) to a nodule

[Back to section 3.3.2]

b) Non-comparative results (1 study)

Table 69. Accuracy of stand-alone software to determine nodule type (2 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 2021,[64] Netherlands	Mixed indication; 65 solid, 13 sub-solid, 2 mixture of solid and sub-solid, 4-30 mm	Consensus reading of 2 radiologists, with discrepancies resolved by a 3rd radiologist	Solid	100.0	73.3	65	4	0	11
			Sub-solid	84.6	100.0	11	0	2	67
			Mixture solid/sub-solid	0	100.0	0	0	2	78
Murchison 2022,[31] UK	Mixed indication; 325 solid, 57 sub-solid; 3-30 mm?	Consensus reading of 2 radiologists, with discrepancies resolved by a 3rd radiologist	Solid	98.8	68.4	321	18	4	39
			Sub-solid	68.4	98.8	39	4	18	321

FN, False negative; FP, False positive; TN, True negative; TP, True positive

13.5.5 Whole read

13.5.5.1 Accuracy for lung cancer detection based on whole read

c) Non-comparative results (1 study)

Screening population – AVIEW Lungscreen (1 study)

A second analysis of the K-LUCAS project by Hwang et al. included 10,424 consecutive participants who underwent baseline LDCT after the implementation of the AVIEW Lungscreen software.[48] The LDCT 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 one year from LDCT as well as any lung cancer diagnosis after LDCT. The reference standard was based on medical record review, with 52 participants being diagnosed with lung cancer within one year from LDCT and 6 participants after one year from LDCT. Using the average transverse diameter (2-D measurement), the sensitivity for lung cancer within one year was 96.2% (50/52) and the specificity was 91.7% (9,515/10,372; 95% CI 91.2 to 92.3). Using the effective nodule diameter (based on volumetric measurement), the sensitivity for lung cancer within one year was also 96.2% (50/52) and the specificity was slightly lower with 90.9% (9,433/10,372; 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% (9,512/10,366; 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% (9,430/10,366; 95% CI 90.4 to 91.5).

13.5.6 Nodule registration and growth assessment

13.5.6.1 Nodule registration

Non-comparative results (1 study)

Mixed population – Veye Chest (Aidence) (1 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). 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 study used the software Veye Chest (Aidence) in stand-alone mode for nodule registration and compared the findings to 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 FP-pairs was 0.0.[31]

[Back to Section 3.4.2.1]

13.5.6.2 Nodule growth assessment

Stand-alone AI vs unaided reader

Mixed population – Veye Chest (Aidence) (1 study)

The same study mentioned above (Murchison et al. 2022) compared nodule growth rate assessment (relative volume difference between a nodule visible on the baseline and follow-up CT scan) for 23 nodule pairs between stand-alone AI and two unaided radiologists . The mean growth percentage difference was similar between 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, due to a single incorrect segmentation of the stand-alone software, the upper end of its confidence interval is twice as high compared to that of readers, illustrating that visual verification of the nodule segmentation by human readers is still advised.

[Back to section 3.4.2.3]

13.5.7 Practical implications – Additional results

13.5.7.1 Other outcomes (not pre-specified in the protocol)

Radiologist workload reduction when using AI-based software as pre-screen (1 study)

Screening population – AVIEW LCS (Coreline Soft) (1 study)

One study 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.[30] 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 based on NELSONplus/EUPS protocol.[94, 95] 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) to 86.7% (upper limit) could be expected.

13.5.8 Impact on patient management - Additional results

13.5.8.1 Characteristics of detected nodules

b) Non-comparative results (3 studies)

Three studies reported characteristics of nodules detected by software-assisted readers[48, 50] and stand-alone software,[64] respectively, without comparator.

Mixed population – Veye Chest (Aidence) (1 study)

Martins Jarnalo et al. randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital.[64] The average size of all 130 (80 true positive and 50 false positive) nodules between 4-30 mm detected by stand-alone software (Veye Chest, Aidence) was 9.0 mm (SD ± 7.1 mm); 85% were solid, 14% were sub-solid and 1% were mixed solid/sub-solid (**Error! Reference source not found.**).

Screening population – AVIEW Lungscreen (Coreline Soft) (2 studies)

The two prospective studies by Hwang et al.[48, 50] 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 for nodule detection. 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 22**.

13.5.8.2 Characteristics of true positive nodules

c) Non-comparative results (4 studies)

Four studies reported on characteristics of true positive nodules detected by stand-alone software,[49, 64] by software-assisted readers,[54] and/or by the reference standard.[30, 54, 64]

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

Hwang et al. included 4,666 participants who had undergone lung cancer screening as part of the K-LUCAS project after the implementation of the software AVIEW Lungscreen (Coreline Soft).[49] Stand-alone software correctly detected 2,147 nodules, of which 96.6% (2,075/2,147) were solid, 1.6% (34/2,147) were part-solid and 1.8% (38/2,147) were ground glass nodules. The Lung-RADS categories of the correctly detected nodules are reported in **Table 70**.

Table 70. Characteristics of correctly detected and missed nodules of stand-alone software in a consecutive screening population in Korea[49]

Lung-RADS category	Stand-alone software	
	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% (175/2,147)	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)

Screening population – ClearRead CT (Riverain Technologies) (1 study)

Singh et al. included 150 patients who underwent LDCT of the chest as part of the NLST - the first 125 patients with sub-solid nodules (154 part-solid or 156 ground glass nodules between 6 and 30 mm) and the first 25 patients with no nodules detected.[54] 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) (1 study)

Martins Jarnalo et al. randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital.[64] Ninety-one nodules with sizes between 4-30 mm were detected by the reference standard (consensus reading of an experienced chest radiologist and a resident radiologist, with discrepancies were 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 sub-solid, and two (2%) were a mixture of both solid and sub-solid (**Table 27**). 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 sub-solid and 3% were a mixture of both.

Screening population – Reference standard only (1 study)

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.[30] According to the 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.

13.5.8.3 Characteristics of false negative (missed) nodules

d) Non-comparative results (5 studies)

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

Hwang et al. included 4,666 participants who had undergone lung cancer screening as part of the K-LUCAS project after the implementation of the software AVIEW Lungscreen.[49] Stand-alone software nodule detection results were available in 3,972 (85.1%) of participants. Out of 2,133 nodules missed by the stand-alone software, 91.7% (1,957/2,133) were solid, 1.7% (36/2,133) were part-solid and 6.6% (140/2,133) were ground glass nodules. The Lung-RADS categories of missed nodules are reported in **Table 22**. Around 0.4% (8/2,133) of missed nodules were confirmed cancer nodules.

Screening population – Veolity (MeVis) (1 study)

The study by Hall et al. was performed in London (UK) and is a sub-study of the LSUT trial.[25] It included 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) (1 study)

Singh et al. included 150 patients who underwent LDCT of the chest as part of the NLST - the first 125 patients with sub-solid nodules (154 part-solid or 156 ground glass nodules between 6 and 30 mm) and the first 25 patients with no nodules detected.[54] As part of a MRMC 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 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) (1 study)

Wan et al. included LDCT images from 50 Taiwanese patients with mixed indications who had subsequent excision of their nodule(s)[56] Of 75 nodules ≤2 cm, the stand-alone software (ClearRead CT, Riverain Technologies) missed 14 nodules: 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. Among 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 statistically significant difference (p = 0.001).

Mixed population – Veye Chest (Aidence) (1 study)

Martins Jarnalo et al. randomly selected 145 chest CT scans from 145 different patients that were performed for various indications at a single Dutch hospital.[64] The nodules missed by stand-alone

software (Veye Chest, Aidence) had 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 missed nodules were sub-solid (4 mm, 18 mm and 20 mm).

13.5.8.4 Number of people undergoing CT surveillance

e) Non-comparative results (3 studies)

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

The study by Hwang et al. included 3,353 consecutive CT images from the K-LUCAS lung cancer screening project in Korea.[50] Based on the original reading by single experienced thoracic radiologist with concurrent use of the AVIEW Lungscreen (Coreline Soft) software, 16.0% (535/3,353) were classed as Lung-RADS category 3 or 4A and 21.6% (723/3,353) were classed as 'intermediate' according to NELSON criteria, respectively.

Screening population – Veolity (MeVis) (1 study)

The study by Hall et al. included all 770 patients from the UK-based LSUT trial who received LDCT for lung cancer screening.[25] 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 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) were referred for CT surveillance of which eight people were later discounted after comparison with previous imaging, leaving 16.2% (125/770) receiving CT surveillance.

Symptomatic population – InferRead CT Lung (Infervision) (1 study)

Kozuka et al. randomly selected 120 chest CT images from cases of suspected lung cancer who underwent CT examination at a single hospital.[57] 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.

13.5.8.5 Number of people having biopsy or excision

f) Non-comparative results (3 studies)

Screening population – AVIEW Lungscreen (Coreline Soft) (1 study)

The study by Hwang et al. included 3,353 consecutive CT images from the K-LUCAS lung cancer screening project in Korea.[50] In the original reading by single experienced thoracic radiologist with concurrent use of the AVIEW Lungscreen (Coreline Soft) software, 4.1% (137/3,353) were positive on the narrow definition of Lung-RADS (i.e. Lung-RADS category 4B or 4X) and 1.6% (52/3,353) were positive according to NELSON criteria.

Screening population – Veolity (MeVis) (1 study)

The study by Hall et al. was performed in London (UK) and is a sub-study of the LSUT trial.[25] 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) (1 study)

Kozuka et al. randomly selected 120 chest CT images from cases of suspected lung cancer in patients aged 20 years or older who underwent CT examination at a single hospital in Japan between November and December 2018.[57] Of all 743 nodules ≥3 mm that were detected by the reference standard (majority reading of three experienced radiologists), 12 (1.6%) nodules were diagnosed as malignant and 44 (5.9%) nodules were followed up as nodules suspected lung cancer.

13.5.8.6 Other outcomes (not pre-specified in the protocol)

Positivity rate (Lung-RADS category 3 or higher) (3 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 or higher) of LDCT images taken and assessed in screening practice with and without the use of the

AVIEW Lungscreen software (Coreline Soft).[48-50] The only comparative study[49] 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.

g) Comparative results – Reader with and without software (1 study)

Screening population - AVIEW Lungscreen (Coreline Soft) (1 study)

In a before-after study, Hwang et al. included 6,487 consecutive participants of the K-LUCAS project: 1,821 participants were screened before the implementation of the AVIEW Lungscreen software, and 4,666 participants received screening after the implementation of the software.[49] The LDCT images were read by single experienced chest radiologists in clinical practice, and patients with Lung-RADS category 3 or higher 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 the implementation of the software (9.9% [180/1,821]) compared to the images interpreted after software implementation (11.0% [511/4,666]; $p = 0.211$). With software use, the per-participant positive rate was significantly increased though when nodules were measured on maximum orthogonal planes (14.1% [657/4,666]; $p < 0.001$) or any maximum planes (17.4% [812/4666]; $p < 0.001$) compared with measurement on transverse planes.

h) Non-comparative results (2 studies)

Screening population – AVIEW Lungscreen (Coreline Soft) (2 studies)

In 10,424 LDCT images that were interpreted with concurrent software use, 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.[48] 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 included 3,353 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% (672/3,353).

13.6 Appendix 6: Literature search strategies: searches to inform the economic model

13.6.1 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	01/12/21	549
Embase	01/12/21	970
NHS EED (CRD)	01/12/21	122
HTA database (CRD)	01/12/21	90
INAHTA HTA database	01/12/21	107
Cost-Effectiveness Analysis (CEA) registry (Tufts Med Centre)	01/12/21	33
EconPapers (Research Papers in Economics (RePEc)	02/12/21	69
ScHARRHUD	02/12/21	13
Total number of records retrieved: 1,953 **Duplicates removed (EndNote): 689** **Final number for screening: 1,264**		
Other sources		
Source	**Date searched**	**Documents retrieved**
National Institute for Health and Care Excellence (NICE) website	07/12/21	0
Canadian Agency for Drugs and Technologies in Health (CADTH) website	07/12/21	4
Google	07/12/21 08/12/21	15, plus 1 ongoing study
ISPOR conference presentations	09/12/21	7; plus 5 posters related to abstracts previously identified
HTAi annual meetings	09/12/21	2
iHEA congresses	09/12/21	0 (2 potentially relevant abstracts unavailable)
Total number sought for retrieval: 35 (+ 1 ongoing study) **Reports not retrieved/available: 2 (iHEA abstracts)** **Final number for screening: 33 (+ 1 ongoing study)**		

Search strategies used:

MEDLINE ALL
Date searched: 01/12/21

Ovid MEDLINE(R) ALL <1946 to November 30, 2021>

1 exp Lung Neoplasms/dg or Solitary Pulmonary Nodule/dg 26945
2 exp Lung Neoplasms/ or Solitary Pulmonary Nodule/ 253279
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. 271939
4 ((pulmonary or lung) adj2 lesion*).kf,tw. 14650
5 2 or 3 or 4 357079
6 Mass Screening/ 111107
7 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw. 4748
8 "Early Detection of Cancer"/ 31301
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/ 665323
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. 2037778
11 6 or 7 or 8 or 9 or 10 2374261
12 5 and 11 68258
13 1 or 12 [lung neoplasms; diagnostic imaging or screening] 74051
14 *economics/ 10766
15 exp *"costs and cost analysis"/ 76423
16 (economic adj2 model*).mp. 14167
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. 37484
18 (cost-effective* or pharmacoeconomic* or pharmaco-economic* or cost-benefit or costs).ti,kf,kw. 79637
19 (life year or life years or qaly* or cost-benefit analys?s or cost-effectiveness analys?s).ab,kf,kw. 34382
20 (cost or economic*).ti,kf,kw. and (costs or cost-effectiveness or markov or monte carlo or model or modeling or modelling).ab. 74307
21 or/14-20 [CADTH Narrow Economic Filter - OVID Medline, Embase https://www.cadth.ca/strings-attached-cadths-database-search-filters] 201764
22 13 and 21 481
23 Quality-Adjusted Life Years/ 14121
24 (quality adjusted or adjusted life year*).ti,ab,kf. 19799
25 (qaly* or qald* or qale* or qtime*).ti,ab,kf. 12541
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. 17499
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. 14119
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. 24278

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 http://dx.doi.org/10.1017/S0266462317000897] 81449
36 13 and 35 193
37 22 or 36 549

Embase
Date searched:L 01/12/21
Embase <1974 to 2021 November 30>

1 exp lung cancer/di or lung nodule/di 46425
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. 388958
3 ((pulmonary or lung) adj2 lesion*).kf,tw. 20844
4 1 or 2 or 3 416579
5 mass screening/ or cancer screening/ 141880
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/ 1351059
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. 2769230
10 5 or 6 or 7 or 8 or 9 3410416
11 4 and 10 113394
12 *economics/ 27332
13 exp *"costs and cost analysis"/ 84204
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. 57878
16 (cost-effective* or pharmacoeconomic* or pharmaco-economic* or cost-benefit or costs).ti,kf,kw. 117531
17 (life year or life years or qaly* or cost-benefit analys?s or cost-effectiveness analys?s).ab,kf,kw. 53133
18 (cost or economic*).ti,kf,kw. and (costs or cost-effectiveness or markov or monte carlo or model or modeling or modelling).ab. 112254
19 or/12-18 [CADTH Narrow Economic Filter - OVID Medline, Embase https://www.cadth.ca/strings-attached-cadths-database-search-filters] 286393
20 11 and 19 767
21 Quality-Adjusted Life Years/ 30198
22 (quality adjusted or adjusted life year*).ti,ab,kf. 28814
23 (qaly* or qald* or qale* or qtime*).ti,ab,kf. 23274
24 (illness state*1 or health state*1).ti,ab,kf. 12756
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. 27682
28 utilities.ti,ab,kf. 13218
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. 25481
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. 41638
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 http://dx.doi.org/10.1017/S0266462317000897] 133355
34 11 and 33 400
35 20 or 34 970

NHS EED and HTA database (CRD) https://www.crd.york.ac.uk/CRDWeb/HomePage.asp
date searched: 01/12/21

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: 01/12/21

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
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: 01/12/21
Basic search screen: Methods selected
Results of each search were copied and pasted into 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
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: 02/12/21
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 https://www.scharrhud.org/index.php?recordsN1&m=search

Date searched: 02/12/21

(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 https://www.nice.org.uk/

Date searched: 07/12/21

Browsed: NICE Guidance > Conditions and diseases > Cancer > Lung cancer:
https://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: https://www.nice.org.uk/guidance/published?sp=on
Filters: Diagnostics guidance, Technology appraisal guidance

lung cancer	48 results, of which 0 relevant
nodule	0 results
nodules	0 results

Browsed guidance In consultation: https://www.nice.org.uk/guidance/inconsultation
20 results, 0 relevant to lung cancer/pulmonary nodules

Searched guidance In development: https://www.nice.org.uk/guidance/indevelopment
Filters: Diagnostics guidance, Technology appraisal guidance

lung cancer	51 results, of which 0 relate to diagnostic imaging
nodule	1 result; 0 unique results
nodules	1 result; 0 unique results

Canadian Agency for Drugs and Technologies in Health (CADTH) website https://www.cadth.ca/
Date searched: 07/12/21
Search box on homepage, results limited to Reports tab.
Search terms:
lung cancer 76 results; 6 on imaging;of which 1 not a cost-effectiveness/economic evaluation; 1 already retrieved by database searches; **4 reports retrieved**
nodules7 results; 3 on imaging; all 3 already identified above

Google www.google.co.uk
Dates searched: 07-08/12/21
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 (Dept 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 4 (LeMense 2020, Edelman Saul 2020, Pyenson 2019, Gilbert 2018)
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 https://www.ispor.org/heor-resources/presentations-database/search

Date searched: 09/12/21
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: 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: 09/12/21

HTAi 2021 Virtual (Manchester). Full program available at:
https://htai.org/wp-content/uploads/2021/06/HTAi_AM21_Full-Program.pdf
Searched (Ctrl + F) for:
lung
pulmon
chest
thora *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); *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
https://www.healtheconomics.org/page/PastCongresses
Abstracts not available

Date searched: 09/12/21
Searched (Ctrl + F) for:

- lung
- pulmon
- chest
- thora

in all of the following:

Beijing 2009. Programme available at:
https://cdn.ymaws.com/www.healtheconomics.org/resource/resmgr/past_congresses/ihea-2009-beijing-programme.pdf

Toronto 2011. Programme available at:
https://cdn.ymaws.com/www.healtheconomics.org/resource/resmgr/past_congresses/ihea-2011-toronto-programme.pdf

Sydney 2013. Programme available at:
https://cdn.ymaws.com/www.healtheconomics.org/resource/resmgr/past_congresses/ihea-2013-sydney-programme.pdf

Dublin 2014. Programme available at:
https://cdn.ymaws.com/www.healtheconomics.org/resource/resmgr/past_congresses/ihea-2014-dublin-programme.pdf

Milan 2015. Programme available at:
https://cdn.ymaws.com/www.healtheconomics.org/resource/resmgr/past_congresses/ihea-2015-milan-programme.pdf

Boston 2017. Programme available at:
https://cdn.ymaws.com/www.healtheconomics.org/resource/resmgr/iHEA_Program_2017.pdf

Basel 2019. Programme available at:
https://cdn.ymaws.com/www.healtheconomics.org/resource/resmgr/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, United States
Author(s): Joshua T. Cohen, David van Klaveren, D6ora 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.

*13.6.2 **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	02/03/22	375
Embase	02/03/22	517
CISNET website: publications list	03/03/22	144
Total:		1,036
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, 03-09/03/22	10 potentially relevant documents retrieved (9 articles, 1 conference abstract) 0 potentially useful registries/websites identified 2 ongoing studies of potential interest identified (IDEAL, Watch the Spot)
Google Dataset Search, 29-30/03/22	1 potentially relevant dataset retrieved

Search strategies used:

MEDLINE via Ovid
Date searched: 02/03/22
Ovid MEDLINE(R) ALL <1946 to March 01, 2022>
1 (growth rate* or growth curve* or doubling time*).kf,tw. 95469
2 Lung Neoplasms/di, dg 50814
3 Solitary Pulmonary Nodule/ 4475
4 (lung nodule* or pulmonary nodule*).kf,tw. 11482
5 2 or 3 or 4 58727
6 1 and 5 375

Embase via Ovid
Date searched: 02/03/22
Embase Classic+Embase <1947 to 2022 March 01>
1 (growth rate* or growth curve* or doubling time*).mp. 139373
2 exp lung cancer/di [Diagnosis] 43090
3 lung nodule/ 24693
4 (lung nodule* or pulmonary nodule*).kf,tw. 19482
5 2 or 3 or 4 68270
6 1 and 5 517

CISNET: Cancer Intervention and Surveillance Modeling Network https://cisnet.cancer.gov/
Date searched: 03/03/22
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) 3/3/22

search terms: list patient registries *browsed 1st 30 results. Checked:*

https://www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries
> https://epi.grants.cancer.gov/cancer-registries/

>https://cancer.ca/en/

https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/encepp-resource-database-inventory-patient-registries_en.pdf

https://www.encepp.eu/encepp/search.htm searched:
Data source > 'lung cancer'
lung
nodule
cancer *nothing relevant found*

https://www.ncri.ie/ *has good survival statistics, but nothing on growth*

https://www.infodesk.com/life-sciences/types-of-patient-registries-and-where-to-find-them/
> CDC resources: browsed https://www.cdc.gov/cancer/lung/ *lung cancer stats are available (USCS)but not growth rates.*
CDC search box: pulmonary nodules *nothing relevant*
https://www.cdc.gov/cancer/npcr/meaningful_use.htm

> https://www.pcori.org/
browsed https://www.pcori.org/topics/cancer
search box: nodules :
> Watch the Spot:
https://www.pcori.org/research-results/pcori-literature/methods-watch-spot-trial-pragmatic-trial-more-vs-less-intensive-strategies-active-surveillance-small-pulmonary-nodules
https://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

https://www.eunethta.eu/parent/ *appears to be closed; links are dead*

https://www.ncra-usa.org/Advocacy/IMSWR/List-of-Medical-Registries

https://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) 7/3/22

search terms: pulmonary nodule growth dataset OR registry OR audit *browsed 1st 30 results. Checked:*

BTS guideline https://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 1st 10 results. Checked:*

https://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*

https://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:
https://www.theabr.org/
https://www.acr.org/ *2 'incidental findings' papers on adherence/real life follow up may be of interest*
https://www.ahra.org/Default.aspx
https://car.ca/
https://www.myesr.org/
https://www.myesti.org/
https://fleischner.memberclicks.net/
https://www.hkcr.org/
https://www.icimagingsociety.org.uk/
https://www.iria.in/
http://www.isradiology.org/
https://www.ranzcr.com/
https://www.radiology.ie/
https://www.rsna.org/
https://www.rssa.co.za/ *need membership to access most documents*
https://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) 9/3/22

search terms: pulmonary nodule natural history database OR registry OR audit *browsed 1st 50 results. Checked:*

https://clinicaltrials.gov/ct2/show/NCT01540552
> https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4405280/
potentially relevant study/article

https://www.frontiersin.org/articles/10.3389/fonc.2020.00318/full - *potentially relevant study/article*

https://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
>https://www.nationaljewish.org/directory/lung-nodule-registry-program
>https://doi.org/10.1016/j.jacr.2021.01.018
>https://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 1st 30 results. Checked:*

https://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: 9/3/22

https://data.gov.uk/
searched (topic: health):
lung cancer
lung nodules
pulmonary nodules
nodule

nothing relevant found

National Comprehensive Cancer Network https://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)
https://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) 29-30/03/22

pulmonary nodule growth rate	20 data sets found	*1 potentially relevant dataset downloaded*
pulmonary nodules doubling time	2 results; both already found above	
lung nodules doubling time	same 2 results retrieved	

*13.6.3 **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	30/06/22	228	20
Google	23/06/22	20	1, plus section of BTS guideline on prevalence (see below)
Reference checking from BTS guideline	23/06/22	32	8
Total:		*280*	***29***

Search strategies used:

MEDLINE via Ovid
Date searched: 30/06/22
Database: Ovid MEDLINE(R) ALL <1946 to June 29, 2022>

1 exp Lung Neoplasms/dg (27068)
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. (283697)
4 1 or 2 or 3 [lung cancer or SPNs] (293093)
5 Mass Screening/ (113855)
6 ((lung or lungs or pulmon*) adj3 (nodule* or cancer* or tumor* or tumour*) adj3 screen*).kf,tw. (5092)
7 5 or 6 [screening] (117282)
8 Tomography, X-Ray Computed/ or exp Tomography, Spiral Computed/ (424801)
9 (comput* adj2 tomograph*).kf,tw. (359889)
10 (CT or LDCT).kf,tw. (402762)
11 8 or 9 or 10 [CT] (782143)
12 Prevalence/ (332019)
13 "prevalen*".kf,tw. (895491)
14 12 or 13 [prevalence] (975826)
15 Incidental Findings/ (11566)
16 (incidental* adj2 (finding* or found or discover* or diagnos* or detect*)).kf,tw. (29485)
17 "incidentaloma*".kf,tw. (2592)
18 15 or 16 or 17 [incidental findings] (36802)
19 4 and 7 and 11 and 14 [lung ca/PN screening CT prevalence] (337)
20 (pulmonary nodule* or lung nodule*).kf,tw. (11812)
21 2 or 20 [PNs - not Ca] (12891)
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/ (385304)
26 (national health service* or nhs*).ab,in,ti. (247302)
27 (english not ((published or publication* or translat* or written or language* or speak* or literature or citation*) adj5 english)).ab,ti. (45087)

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. (2322787)
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. (1633647)
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. (65320)
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. (240883)
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. (31250)
33 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 (2915825)
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/) (3215213)
35 33 not 34 [UK search filter, Ayiku et al 2017 https://onlinelibrary.wiley.com/doi/10.1111/hir.12187] (2762901)
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. (48708)
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] (54250)
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. (66593)

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 (68596)
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 Information & Libraries Journal* 2017;**34**(3):200-16. http://dx.doi.org/https://doi.org/10.1111/hir.12187

Google (Chrome browser) 23/06/22

search terms: lung nodule prevalence UK *browsed 1st 20 results. 2 potentially relevant, one of which is the BTS guideline:*

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-ii54. http://dx.doi.org/10.1136/thoraxjnl-2015-207168
8 potentially relevant papers

13.7 Appendix 7: Growth model and its development process

Introduction

Assessing the impact of AI assistance during CT surveillance requires modelling of the pathways which people with lung nodules would take between repeated CT scans based on the findings of the earlier CT scan. During the time period between CT scans, the nodule may grow and this needs to be taken into account when assessing the impact of AI assistance at follow-up CT scans. Thus, we need know the natural history of lung cancer in the form of growth in malignant nodules and quantify it using a malignant nodule growth model. In order to facilitate this, we firstly identified studies that have included such models, 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 PA and HG. 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 6.

Results

We screened 750 titles and abstracts, of which 15 were potentially relevant and were read in full. From these, four studies (Gould et al., 2003, Sutton et al., 2020, Edelsberg et al., 2018; Treskova et al., 2017) that modelled disease progression based on tumour growth were considered useful and discussed below. Details of these studies can be found in Table 71.

The underlying growth model used by Edelsberg et al. and Sutton et al. was obtained from Gould et al. (Gould et al., 2003). 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-terms and costs 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 1973. 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 volume doubling time measured in 67 cases of people with asymptomatic nodules measuring less than 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 to CT surveillance alone in people with an indeterminate risk of lung cancer. Authors fitted an exponential model to the observed data from Steele and Buell 1973 to derive monthly transition probabilities. Sutton et al. 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 to CT surveillance alone. Authors used the same approach to derive monthly transition probabilities. We noted similarities and differences in the assumptions made with regards to the growth models: Gould et al. 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 assumed that after three CT scans and there was no evidence of the nodule doubling, the nodule was considered benign. Authors further assumed that malignant nodules that were not diagnosed at model entry, increased in size, and progressed during CT surveillance. Sutton et al., assumed that the transition probability of progressing from local to regional is the same as 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 (Steele and Buell 1973) to model our growth model. We considered that this study may not be generalisable to our sub-populations of interest as study participants were male and all had lung nodules less than 6 cm. Additionally, the study is dated, and the characteristics of patients are likely to be different compared to 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 instead of an exponential. Moreover, evidence of volume doubling time was collected using routine chest films in the original study but now this is 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 low-density computed tomography (LDCT) by using a microsimulation model. The model was populated with 10% of the German population aged 40 years and older. Data on smoking behaviour was obtained from the German Health Update (GEDA) survey (years 2009–2012), and the demographic structure of 2012 was obtained from the German statistical office. The growth model also uses the data from US NLST, and NELSON lung cancer screening trials. The NLST algorithm assessed the nodule diameter, and according to the sizes 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, they could be recommended to undergo further screening (people with negative results), a follow-up exam (people with indeterminate results), or an immediate diagnostic workup for the people with positive results.

Treskova and colleagues assumed that the threshold tumour volumes at the stages of nodal involvement, and distant disease and clinical diagnoses were randomly drawn from lognormal distributions. Lung cancer progression was described via tumour growth, lymph nodes involvement and metastases, and growth of malignant nodules is defined by a Gompertz function. The model included a natural history of a biological two-stage clonal expansion (TSCE) of the disease incorporating the nodule growth (in terms of the rate and time). The TSCE 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 which can result in efficiency. They adopted a model that traced the efficiency and effectiveness of the lung cancer screening program 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 the US NLST and the NELSON trial.

Treskova et. al used the volume doubling (VDT), an indicator used in the BTS guidelines for managing people with lung nodules. 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 for 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 two years of CT surveillance for solid nodules, and four years of CT surveillance for sub-solid 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 to track the 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 the malignant nodule growth over the person's lifetime. In the proposed growth model, the disease progression is characterised by the volume of the nodule, its location, and the metastatic probability of the nodule. They assumed that if a person's threshold volume exceeds from calculated maximum expected volume (V_{max}), the corresponding cancer stage will not be reached during the lifetime of this patient.

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 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 sub-solid nodules (only the growth function differs between solid and sub-solid nodules):

$$\boldsymbol{Sphere\ volume} = \frac{\boldsymbol{\pi}}{\boldsymbol{6}} * (\boldsymbol{Diameter})^{\boldsymbol{3}}$$

$$\boldsymbol{Sphere\ diameter} = \sqrt[3]{\frac{\boldsymbol{6} * (\boldsymbol{sphere\ volume})}{\boldsymbol{\pi}}}$$

$$\boldsymbol{Volume\ doubling\ time}\ (\boldsymbol{VDT}) = \ \boldsymbol{time} * \frac{\mathbf{log}(\mathbf{2})}{\mathbf{log}\left(\frac{\boldsymbol{Sphere\ volume\ at\ time}_{\boldsymbol{t=i+1}}}{\boldsymbol{Sphere\ volume\ at\ time}_{\boldsymbol{t=i}}}\right)}$$

Solid nodules

$$\boldsymbol{Gompertz\ growth\ function = Volume_{max} * \frac{Volume_{t=0}}{Volume_{max}}^{-time*alpha}}$$

Where $\boldsymbol{Volume_{max} = 14137.17}$ and $\boldsymbol{alpha \sim Normal\ distribution}(\boldsymbol{-7.765, 0.5504})$.

Sub-solid nodules

$$\boldsymbol{Linear\ growth\ function = Volume_{t=i} + 2 * \frac{time}{alpha}}$$

Where $\boldsymbol{alpha \sim \log\left(Normal\ distribution(3.6316, 1.5279)\right)}$.

Table 71. Characteristics of studies that included a growth model

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
Gould et al., 2003	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 Assumed that pulmonary nodules measured 2cm in diameter 12.5% of people with malignant nodules had regional lymph node involvement Monthly probabilities for disease progression depended on VDT,	Used in several economic analyses Doubling time by cell type (squamous cell, adenocarcinoma, bronchiolar, adenosquamous and undifferentiated)	Based on dated information that included males only Appears to be solitary nodules only Unclear about definitions used for lung nodules (TP, TN, FP, FN). Historical data in males with asymptomatic nodules measuring <6cm. Evidence of VDT is collected using routine chest films.

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				a measure of tumour growth Tumour starts from a single cell that measures 10 microns in diameter that doubles in volume at a constant rate Death occurs after 40 doublings for a tumour size 10cm Untreated lung cancer progresses from local → regional → distant → dead Transition probabilities for progressing from local → regional		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				→ distant disease are equal Growth would be detected when the nodule double once in volume		
Sutton et al., 2020	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*	Progression rates in people with undiagnosed malignant nodules were based on observed VDT obtained from Gould et al., which were originally obtained from Steele 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 months, 12 months, and 24 months. People with a negative test continued to undergo surveillance.	The model includes both detection and treatment phases Included a probability associated with growth of a benign nodule at the first month and subsequent probability of growth Transition probabilities reported for the natural history model	Unclear about stage shift Not revealed natural history for the growth rate 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). used.

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
		offered surveillance through repeat CT scanning.'		Probability is the same for progression from undiagnosed local to regional disease and from regional to distant disease 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. 100% compliance with CT surveillance		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
Edelsberg et al., 2018	Cost-effectiveness analysis	To assess the cost-effectiveness of autoantibody test compared to 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 to 30mm and have an estimated 5-60% risk of lung cancer. After three CT scans and there is no volume doubling, the nodule is assumed to be benign. Malignant nodules are diagnosed at biopsy. 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	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 data base), focused only on malignant nodules, natural history is based on the VDT, but not elaborated,

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				assumed to be diagnosed soon after CT scan following volume doubling. Patients who are benign that had tested positive would receive a biopsy that would confirm no malignancy.		
Chen et al., 2014	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 dataset	Several assumptions were made for the tumour growth and metastasis modelling: The primary tumour grows from a single cell, with an assumed volume of $1x10^{-9}$ cm^3. The growth rate λ, is related to	Provided tumour size frequency distribution for local, regional, and distant disease. Incorporating the smoking behaviour in the natural history 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 doesn't provide a clear lung nodules growth pattern over time.

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				the tumour doubling time and is determined when first detected and is assumed to remain constant over time. Growth rate follows a gamma distribution Metastases are defined as nodal or distant. Different rates for each type of metastases		The study seems more suitable for predicting the lung cancer probability due to smoking and then for non-smoker population probably not applicable.
Treskova M, et al. (2017)	A stochastic modular microsimulation model that simulated	The study aimed to investigate the effects of the eligibility criteria	The model was populated with 10% of the German	The module uses the age at the onset of the first malignant cell.	The natural history module contains a biological two stage clonal expansion (TSCE) model and a tumour	The model is only focused on the screening population

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
	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.	and nodule management on the benefits, harms and cost-effectiveness of lung screening with LDCT in a population-based setting.	population aged 40 years and older. Data on smoking behaviour was obtained from the German Health Update (GEDA) survey (years 2009–2012), 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	Threshold tumour volumes at the stages of nodal involvement, distant metastases and clinical diagnosis are randomly drawn from log-normal distributions. Threshold tumour volumes at the stages of nodal involvement, distant metastases and clinical diagnosis are randomly drawn from log-normal distributions. The clinical detection module determines the stage of lung	growth component and simulates a complete flow of events in the development of lung cancer. The model has space for smoking and its impacts The probabilities of overdiagnosis, by using data from both NLST, and NELSON. The survival probabilities based on the histological staging of lung cancer, size specific sensitivity of LDCT. Rate of cases at stage II as an earlier stage of lung cancer. The complication rates at workup by the diameter	No cost per QALYs analysis (only cost er LYG). The total cost of screening has not been 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 limitation in the dataset

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				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. 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. The parameters vary over the	of the malignant nodule and for benign nodule Developing a two steps calibration: for each lung cancer type mean and standard deviation 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, Secondly, 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”.	

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				histological classes and stages at the time of diagnosis. Two nodule management algorithms were designed based on those used in the NELSON and NLST trials. The tumour is staged according to TNM classification based on the volume and spread. Individuals with screen-detected lung cancer live at least if they would in the no	In order to obtain the costs for people with early-stage cancer in our model we applied ratio of costs between III and I stages is used to define a base case scenario. 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	

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				screening scenario. 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. The tumour growth rate is based Gompertz model.		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
Hofer F, et al. (2018)	A cost-effectiveness analysis from the public payer perspective for a high-risk population defined as heavy former and current smokers (≥20 cigarettes per day) between 55 and 75 years of age. The underlying model consisted of two Markov models (for capturing the diagnosis of patients at an early stage of disease a model requires simulating the period of disease progression before diagnosis (i.e., the natural	To evaluate the cost-effectiveness of a population-based lung cancer screening program from the perspective of a German payer.	Combination of data from the Federal Statistical Office and the nationwide Epidemiological Survey on Addiction equals the number of heavy current and former smokers (≥20 cigarettes per day) aged 55 to 75 within the German system of statutory health insurance. The underlying data contained no information about the duration of heavy smoking habits and may – although	Individuals aged between 55 and 75 to be eligible for the screening program. We chose a cycle length of three months and ran the model for 60 cycles (i.e., 15 years). Half cycle correction was applied. Costs and quality-adjusted life years were discounted by 3% per year. The natural history component of our model consisted of seven states, representing lung cancer stages I to	The model comprised two separate components to distinguish between the natural history of disease and treatment paths and aftercare depending on patients' lung cancer stage at diagnosis. The model has incorporated the parameters such as early recall rate. For lung cancer states in the natural history component of our model, the authors used mortality rates calibrated by the Metropolis Hastings algorithm using priors informed by results of a systematic review. The model used results from another systematic review to estimate mortality	The model is for screening. There is no explicit explanation on how to incorporate the VDTs as a parameter in the model, measurement, calibration, and treating it in the model. The sensitivity is more size-dependant rather lung nodule histological stage. No detailed information on parameters calibrations.

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
	history of disease) and, separately, the treatment of lung cancer and aftercare)		based on best empirical evidence – be an overestimation of the number of people eligible.	IV, a state of no apparent lung cancer, and a state for death. Lung cancer stages IIIa and IIIb were modelled separately because of different treatment regimens. After diagnosis, the simulated patients entered the second component of the model, in which the researcher estimated treatment and aftercare. Treatment paths were designed in accordance with German clinical	rates for the treatment and aftercare component. The model utility weights have been derived from a meta-analysis. The model has a detailed transition probabilities for each state in the natural history. The model has a good utility value for each singular procedure and combination of procedures. The sensitivity has been considered by different stages of the lung nodule progression. Both linear and probabilistic sensitivity analyses have been	No information about the overdiagnosis. No information about the cumulative exposure to the radiation. No details to determine the actual duration of heavy smoking behaviour, or the time passed since individuals changed their (heavy) smoking habits. This may affect the number of people eligible for screening. Diagnosis and treatment for

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				practice guidelines. We assumed that 90% of stage I patients were treated with surgical resection alone, 5% with a combination of surgical resection and chemotherapy, and 5% with a combination of surgical resection, chemotherapy, and radiotherapy. Patients were at risk of local recurrence or distant metastases have been assumed to by a combination of chemo- and radiotherapy or	performed as uncertainty analysis.	small cell lung cancer (SCLC) was not modelled separately.

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				received palliative care. In any of the states, all individuals were at risk of all-cause mortality. The probability of being diagnosed and thus entering the second component of our model differed between individuals who (a) took part in the annual LDCT-based lung cancer screening program or (b) were diagnosed through standard clinical care (i.e., when they became symptomatic).		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				Individuals in the cohort receiving standard clinical care could only be diagnosed when they developed symptoms such as cough, hemoptysis or fatigue that had been identified through a physician visit related to the symptoms. Since the model ran in three-month periods, one quarter of the participants in the screening cohort could also be diagnosed through annual CT screening in each period if they were adherent.		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				Annual screening was assumed to follow the same screening protocol as that in the German LUSI trial, which focused on nodule size and volume doubling time (VDT). Individuals in the screening cohort did not change their radiologist.		
Lin RS, et al. (2012)	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 datasets and validated on screening trials.	Growth of primary tumour volume grows exponentially. The tumour has a constant TVDT. The "*treatment cure threshold*" of cancer as the primary tumour volume at which	The model has been evaluated by using simulation of data from different databases. The model is not only for screening population, and it seems to be helpful for considering other route of diagnosis of lung cancer. The study has a good explained natural history-	The analysis was limited to Caucasians because it is the largest ethnic group of lung cancer patients. The analysis was limited to males because external validation dataset

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				the disease transitions from being curable to incurable, assuming standard of care following detection. The patient would never die from their specific disease if detected and treated at or before the treatment cure threshold. The lethal metastatic burden starts increasing at the treatment cure threshold, thereby we are implicitly excluding	based VDT and the parameters that have been defined and explained very well. 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.	from the Mayo Lung Project (described below) was limited to males only.

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				metastasis that may be eradicated or controlled by systemic treatment when treated before the onset of the lethal metastatic burden. 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. If the patient is not diagnosed and treated before the		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				treatment cure threshold, the lethal metastatic burden becomes the cause of death at the maximal lethal metastatic burden. Disease is symptomatically detected either due to the primary tumour or the lethal metastatic burden, dependent on which presents with symptoms first. Patients are clinically staged with advanced disease		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				if lethal metastatic burden is detected at symptomatic detection. 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; The treatment cure threshold *VC* is assumed to have a Weibull distribution with shape parameter *c1* and scale parameter		

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
				c2; and the ratio *BD/f* is assumed to have a Weibull distribution with shape parameter *b1* and scale parameter *b2*.		
Heuvelmans MA. et al. (2017)	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 = V1exp(t/_), with t time from baseline (days), V1 estimated baseline volume (mm3), and _ estimated time constant. The R2coefficient of determination was used to	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. The study has used the NELSON trial database. The study has some findings in terms of VDT (the number of cancers by VDT groups). Figure 5 reports the VDT in days for 46 lung cancers from the NELSON trial	The study assumptions have not been stated. The study natural history model has not been elaborated well. The VDTs have not been compared with different growth models (e.g., Gompertz, linear or log-linear). Growth patterns for slow-growing lung cancers were evaluated in this

Author	Type of study	Aim(s)/Objective(s)	Data underpinning growth model	Assumptions	Pros	Cons
	evaluate goodness of fit. Overall volume-doubling time for the individual lung cancer is given by _ * log(2).					study. Faster growing lung cancers did not receive at least three CT scans.
CT, computed tomography; TVDT; tumour volume doubling time; VDT, volume doubling time; NELSON, Nederlands–Leuvens Longkanker Screenings Onderzoek; SEER, The Surveillance, Epidemiology, and End Results; SCLC, small cell lung cancer;						

13.8 Appendix 8: Methods for 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,[11] 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 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 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 two sections below.

13.8.1 Simulation for nodule sizes at baseline (baseline measurement simulation)

We firstly generated a cohort of risk dominant nodules (the largest nodule or the one being most suspicious of being malignant) in people with at least one 'true' nodule (≥3mm 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[48] and served as the reference standard. We generated the values from a log-normal distribution which 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 One (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 Zero, 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 earlier **in Appendix 7 (section 13.7)**.

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 Two, R2), a radiologist with concurrent

AI (Reader Three, 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 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 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 sub-solid nodules).[11] Based on size-specific cancer risk estimated from the NELSON lung cancer screening trial,[3] we could then estimate the proportion of true malignant nodules that goes through individual nodule management pathways (e.g. discharge, surveillance, definitive management) and subsequently being detected or missed. These outputs could then be used as parameter inputs for our model to compare downstream impacts.

13.8.1.1 Reader One (R1): Consensus reading (reference standard)

Data from Hwang et al.[48] 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) for solid nodules, 11.9 mm (11.1) for part-solid nodules, and 5.8 mm (IQR 4.7) for non-solid nodules. The part-solid and non-solid nodules were combined at a ratio of 4:5 to create the simulated sub-solid 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.[57] were used as 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 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 sub-solid nodules based on the screening population:

The median nodule size was 3.6 mm for solid nodules and 8.5 mm for sub-solid nodules, a factor of 2.36.

This was applied to the 4.7 mm from reported by Kozuka et al.[57], resulting in an assumed median sub-solid 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.

13.8.1.2 Reader Zero (R0): the unobservable 'true' nodule size

Reader Zero was the assumed 'true' nodule size which was simulated using the values from R1. We expect 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 nodules had a probability of being malignant. These lung cancer probabilities were derived from Horeweg et al.[3] who used 9,681 non-calcified nodules detected by CT screening in 7,155 participants in the screening group of the NELSON trial. For solid nodules, this was estimated to be 0.009 for nodules between 5 to <6 mm, 0.011 for nodules between 6 to <8mm 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,[57] screening population[96] and incidental population.[97]

13.8.1.3 Reader Two (R2): Stand-alone AI

Although in current practice all CT scans would 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 this was the only data reported in some of the included studies, and it is generally recommended that size/volume measurements obtained by AI should not be manually adjusted

unless there are clear issues related to nodule segmentation in order to preserve consistency afforded by AI measurements.[17]

The base case simulation for R2 was based on the discrepancies between nodule size measurements by stand-alone AI and majority consensus of three radiologists as reported by Martins Jarnalo et al.[64]. This study was chosen as it was the only identified study that reported individual measurement discrepancies of stand-alone AI compared to a reference standard for each of the 77 nodules (**Table 72**). 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 sub-solid nodules (**Table 75**).

Table 72. Nodule size measurement discrepancies of stand-alone AI compared to the reference standard as reported by Martins Jarnalo et al. 2021[64]

Size discrepancy (mm)	# nodules (R2 base case)
-2	2
-1	2
0	54
1	16
2	2
4	1

Scenario analysis 1 also used data by Martins Jarnalo et al.[64] 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 R1, only SD was changed so that the agreement between R1 and R2 was approximately 67.5% (**Table 75**).

Scenario analysis 2 was based on a phantom study by Wu et al.,[98] where 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 sub-solid nodules (**Table 75**).

13.8.1.4 Reader Three (R3): Concurrent AI

The base case simulation for R3 was similar to that of R2, using the discrepancies reported by Martins Jarnalo et al.[64] 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 73**). Therefore, the mean size of R3 simulated nodules was R1+0.182 mm, with a SD of 0.639, for both solid and sub-solid nodules (**Table 75**).

Table 73. Discrepancies of concurrent AI diameter measurements, estimated from Martins Jarnalo et al.[64]

Size discrepancy (mm)	# nodules (R3 base case)	# 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)

As a scenario analysis (scenario analysis 3), we further assumed that the radiologist would manually correct the ±2mm discrepancies of stand-alone software measurement (**Table 73**). Thus, the mean size of R3 simulated nodules in scenario analysis 3 was R1+0.182 mm and a SD of 0.448 (**Table 75**).

13.8.1.5 Reader Four (R4): Unaided radiologist

Inputs for the accuracy of manual nodule size measurement using electronic calipers were based on the phantom study by Xie et al.[99] This study was chosen as base case as it observed an underestimation of nodule size, whereas the second identified study (Cohen et al. 2016)[35] reported an overestimation. This DAR observed that "The studies found similar[56, 61] or significantly larger[45] nodule diameters with semi-automatic measurements compared to manual measurements" (see **section 3.3.3.3**); we therefore rated the underestimation observed by Xie et al.[99] as more plausible and used it as base case. This study found that the overall underestimation of diameter for nodules of any density was 9.2±6.0%, and for solid nodules the underestimation was 10.1±6.9%.

In the simulation, the mean size of solid nodules was based on that of R1 minus 10.1%, and for sub-solid 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.,[99] 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), therefore the standard deviation was fixed at 1.5*SD of R3 (1.5*0.639) (**Table 75**).

As a scenario analysis (scenario analysis 4), inputs based on results from Cohen et al.[35] were used. This study observed that the manual measurements of the entire nodule were larger compared to the tumour size on pathology after resection, by a mean difference of +2.38 mm. For both solid and sub-solid nodules, mean nodule size (mm) was R1+2.38, with a standard deviation 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 SD consistent with scenario analysis 1 (**Table 75**).

A final scenario analysis was performed for both the screening and symptomatic populations, scenario analysis 5, where the following assumptions were made for the standard deviations of simulated nodule sizes; the mean for each reader was based on that of R1 (see **Table 74**):

- R1: SD kept the same.
- R2 (Stand-alone AI): we assumed that AI alone would perform worse compared to R3 and R4 (SD multiplied by 2).
- R3 (Concurrent AI): we assumed that this reader would measure more accurately compared to R2 and R4 (SD multiplied with 0.5).
- R4 (Unaided radiologist): we assumed that this reader would measure more accurately compared to R2 but worse compared to R3 (SD multiplied with 1.5).

Table 74. Inputs for scenario analysis 5

Reader	SD multiple	Screening population		Symptomatic population	
		SD (solid)	SD Sub-solid)	SD (solid)	SD (Sub-solid)
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.

13.8.1.6 Other assumptions

Nodule type distribution was different for the screening and symptomatic populations.

13.8.1.7 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 sub-solid.

3. We simulate Reader One'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. σ = the solution to rearranged quantile functions of the log-normal distribution populated using the reported IQRto calculate σ.

4. The measurements for the other three readers are simulated.
5. Summary statistics produced.

The simulation was carried out using R version 4.1.0.

Table 75. Mean nodule size simulation inputs

Population	**Screening**	**Symptomatic**	**Both**	**Both**	**Both**	**Screening**	**Symptomatic**	**Both**	**Both**	**Both**
Reader	**Reader 1**	**Reader 1**	**Reader 2**	**Reader 3**	**Reader 4**	**Reader 2**	**Reader 2**	**Reader 2**	**Reader 3**	**Reader 4**
Distribution	Log-normal	Log-normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal
Solid										
Mean	3.6*	4.7*	R1 + 0.234	R1 + 0.182	R1 - 10.1%	Reader 1	Reader 1	R1 + 0.969	R1 + 0.182	R1 + 2.38
SD	2.1*	1.3*	0.771	0.639	0.639*1.5	2.60	0.63	0.249	0.448	0.50
Sub-solid										
Mean	11.9*	11.1*	R1 + 0.234	R1 + 0.182	R1 - 9.2%	Reader 1	Reader 1	R1 + 0.884	R1 + 0.182	R1 + 2.38
SD	11.1*	1.3*	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
*Median/IQR										

13.8.2 Simulation for nodule growth monitoring

We used the nodules simulated using the base case assumptions for Reader 0 and applied the different growth curves (for both solid and sub-solid nodules) to calculate the true nodule growth at each subsequent timepoint (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 timepoint.

Using these 'true' diameter values at each timepoint, we applied the same transformations to Reader 0 that we applied to at baseline for readers 3 and 4, and calculated the respective estimated nodule volumes and volume doubling times.

For tracking the solid nodules' growth over time from the baseline through turning to cancerous nodules; we used the model which has been developed by Treskova et al. (2018).[69] 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 has used a spherical volume measurement for computing the volume of the nodule and provided the Volume Doubling Time (VDT) for four common histological lung cancer types including:

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 have been 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 based on the assessed volume (V), the screening-detected nodule is classified as a negative ($V<Vfup$), positive ($V\geq Vcut$) or indeterminate result ($Vfup\leq V<Vcut$). More details on Treskova et al. (2018) study can be found in Appendix 7 (section 13.7).

For sub-solid nodules, a linear growth over time was assumed, as reported by Kakinuma et al., 2016___

Using Treskova et al. (2018) in this simulation as follows:

For reader zero: 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 readers 1, 3, and 4, we calculated the probabilities for the model structure. The formulae were used for the calculation has been described in Appendix 7 (section 13.7).

13.8.3 R code for the simulation

Population: Screening

Simulation set-up

Clear workspace

```
rm(list = ls())
```

Import file

```
nodules_df <- read.csv("…\\Simulation 2_Screening.csv", header=TRUE)
```

Importing file of simulated nodules for all 5 readers.

Functions

The following are the various functions used in this script.

The Gompertz function is the growth function used to calculate the volume growth of Reader 0's nodules over the 2 years of follow-up at 3/12/24 months.

The other functions are related to calculating volume from nodule diameter, back-calculating diameter from volume, and calculating volume doubling time.

Gompertz function

```
gompvol <- function (v0,alpha,vmax,t){
  vmax*(v0/vmax)^(exp(-t*alpha))
}
```

Sphere volume

```
spherevol <- function(D){
  (pi/6)*(D^3)
}
```

Sphere diameter

```
spherediam <- function(v){
  (6*v/pi)^(1/3)}
```

Growth to VDT

```
vdt <- function(v0,v1,t){
  t*log(2)/(log(v1/v0))}
```

Paramaters

```
vmax.cancer <- spherevol(30)
lmean.nsclc <- -7.765
lsd.nsclc   <- 0.5504
N           <- length(nodules_df$nodule_type)
nodule_n    <- length(nodules_df$nodule_type)
x           <- 0
```

Reader 0 - True growth

```
# Starting values
nodules_df$r0_dia_1 <- nodules_df$reader_0
nodules_df$r0_vol_1 <- spherevol(nodules_df$r0_dia_1*runif(N,1-x,1+x))
lnormvalues         <- rlnorm(N, lmean.nsclc, lsd.nsclc)

# Volume
nodules_df$r0_vol_2 <- gompvol(nodules_df$r0_vol_1, lnormvalues, vmax.canc
er, 90)
nodules_df$r0_vol_3 <- gompvol(nodules_df$r0_vol_1, lnormvalues, vmax.canc
er, 365.25)
nodules_df$r0_vol_4 <- gompvol(nodules_df$r0_vol_1, lnormvalues, vmax.canc
er, 730.50)
nodules_df$r0_vol_5 <- gompvol(nodules_df$r0_vol_1, lnormvalues, vmax.canc
er, 1461.0)

# Diameter
nodules_df$r0_dia_2 <- spherediam(nodules_df$r0_vol_2)*runif(N,1-x,1+x)
nodules_df$r0_dia_3 <- spherediam(nodules_df$r0_vol_3)*runif(N,1-x,1+x)
nodules_df$r0_dia_4 <- spherediam(nodules_df$r0_vol_4)*runif(N,1-x,1+x)
nodules_df$r0_dia_5 <- spherediam(nodules_df$r0_vol_5)*runif(N,1-x,1+x)

# Volume doubling time
nodules_df$r0_vdt_2 <- vdt(spherevol(nodules_df$r0_dia_1), spherevol(nodul
es_df$r0_dia_2), 90)
nodules_df$r0_vdt_3 <- vdt(spherevol(nodules_df$r0_dia_2), spherevol(nodul
es_df$r0_dia_3), 365.25 - 90)
nodules_df$r0_vdt_4 <- vdt(spherevol(nodules_df$r0_dia_3), spherevol(nodul
es_df$r0_dia_4), 730.50 - 365.25)
nodules_df$r0_vdt_5 <- vdt(spherevol(nodules_df$r0_dia_4), spherevol(nodul
es_df$r0_dia_5), 1461.0 - 730.50)
```

As Reader 0 is the assumed true size of the simulated nodule, nodule growth and diameter at each subsequent timepoint will be based on this Reader. Therefore:

- *1: Calculate the volume of each nodule at 3/12/24 months*
- *2: Calculate the diameter of each nodule at 3/12/24 months using the new volumes*
- *3: Calculate volume doubling time*

Reader 1

```
# Starting values
nodules_df$r1_dia_1 <- nodules_df$reader_1
nodules_df$r1_vol_1 <- spherevol(nodules_df$r1_dia_1*runif(N,1-x,1+x))
lnormvalues         <- rlnorm(N, lmean.nsclc, lsd.nsclc)
```

```
# Diameter
nodules_df$r1_dia_2 <- rnorm(nodule_n, nodules_df$r0_dia_2, sd= 0.100)
nodules_df$r1_dia_3 <- rnorm(nodule_n, nodules_df$r0_dia_3, sd= 0.100)
nodules_df$r1_dia_4 <- rnorm(nodule_n, nodules_df$r0_dia_4, sd= 0.100)
nodules_df$r1_dia_5 <- rnorm(nodule_n, nodules_df$r0_dia_5, sd= 0.100)

# Volume
nodules_df$r1_vol_2 <- spherevol(nodules_df$r1_dia_2)
nodules_df$r1_vol_3 <- spherevol(nodules_df$r1_dia_3)
nodules_df$r1_vol_4 <- spherevol(nodules_df$r1_dia_4)
nodules_df$r1_vol_5 <- spherevol(nodules_df$r1_dia_5)

# Volume doubling time
nodules_df$r1_vdt_2 <- vdt(spherevol(nodules_df$r1_dia_1), spherevol(nodul
es_df$r1_dia_2), 90)
nodules_df$r1_vdt_3 <- vdt(spherevol(nodules_df$r1_dia_2), spherevol(nodul
es_df$r1_dia_3), 365.25 - 90)
nodules_df$r1_vdt_4 <- vdt(spherevol(nodules_df$r1_dia_3), spherevol(nodul
es_df$r1_dia_4), 730.50 - 365.25)
nodules_df$r1_vdt_5 <- vdt(spherevol(nodules_df$r1_dia_4), spherevol(nodul
es_df$r1_dia_5), 1461.0 - 730.50)
```

For the remaining readers:

- *1: Calculate nodule diameter using the same transformations applied when simulating the baseline nodule size, this time onto Reader 0's simulated nodule diameters at 3/12/24 months*
- *Calculate volume and VDT.*

Reader 2 - AI alone

```
# Starting values
nodules_df$r2_dia_1 <- nodules_df$reader_2
nodules_df$r2_vol_1 <- spherevol(nodules_df$r2_dia_1*runif(N,1-x,1+x))
lnormvalues         <- rlnorm(N, lmean.nsclc, lsd.nsclc)

# Diameter
nodules_df$r2_dia_2 <- rnorm(nodule_n, nodules_df$r0_dia_2 + 0.234, sd= 0.
771)
nodules_df$r2_dia_3 <- rnorm(nodule_n, nodules_df$r0_dia_3 + 0.234, sd= 0.
771)
nodules_df$r2_dia_4 <- rnorm(nodule_n, nodules_df$r0_dia_4 + 0.234, sd= 0.
771)
nodules_df$r2_dia_5 <- rnorm(nodule_n, nodules_df$r0_dia_5 + 0.234, sd= 0.
771)

# Volume
nodules_df$r2_vol_2 <- spherevol(nodules_df$r2_dia_2)
nodules_df$r2_vol_3 <- spherevol(nodules_df$r2_dia_3)
nodules_df$r2_vol_4 <- spherevol(nodules_df$r2_dia_4)
nodules_df$r2_vol_5 <- spherevol(nodules_df$r2_dia_5)

# Volume doubling time
nodules_df$r2_vdt_2 <- vdt(spherevol(nodules_df$r2_dia_1), spherevol(nodul
```

```
es_df$r2_dia_2), 90)
nodules_df$r2_vdt_3 <- vdt(spherevol(nodules_df$r2_dia_2), spherevol(nodul
es_df$r2_dia_3), 365.25 - 90)
nodules_df$r2_vdt_4 <- vdt(spherevol(nodules_df$r2_dia_3), spherevol(nodul
es_df$r2_dia_4), 730.50 - 365.25)
nodules_df$r2_vdt_5 <- vdt(spherevol(nodules_df$r2_dia_4), spherevol(nodul
es_df$r2_dia_5), 1461.0 - 730.50)
```

Reader 3 - Concurrent CAD

```
# Starting values
nodules_df$r3_dia_1 <- nodules_df$reader_3
nodules_df$r3_vol_1 <- spherevol(nodules_df$r3_dia_1*runif(N,1-x,1+x))
lnormvalues         <- rlnorm(N, lmean.nsclc, lsd.nsclc)

# Diameter
nodules_df$r3_dia_2 <- rnorm(nodule_n, nodules_df$r0_dia_2 + 0.182, sd= 0.
639)
nodules_df$r3_dia_3 <- rnorm(nodule_n, nodules_df$r0_dia_3 + 0.182, sd= 0.
639)
nodules_df$r3_dia_4 <- rnorm(nodule_n, nodules_df$r0_dia_4 + 0.182, sd= 0.
639)
nodules_df$r3_dia_5 <- rnorm(nodule_n, nodules_df$r0_dia_5 + 0.182, sd= 0.
639)

# Volume
nodules_df$r3_vol_2 <- spherevol(nodules_df$r3_dia_2)
nodules_df$r3_vol_3 <- spherevol(nodules_df$r3_dia_3)
nodules_df$r3_vol_4 <- spherevol(nodules_df$r3_dia_4)
nodules_df$r3_vol_5 <- spherevol(nodules_df$r3_dia_5)

# Volume doubling time
nodules_df$r3_vdt_2 <- vdt(spherevol(nodules_df$r3_dia_1), spherevol(nodul
es_df$r3_dia_2), 90)
nodules_df$r3_vdt_3 <- vdt(spherevol(nodules_df$r3_dia_2), spherevol(nodul
es_df$r3_dia_3), 365.25 - 90)
nodules_df$r3_vdt_4 <- vdt(spherevol(nodules_df$r3_dia_3), spherevol(nodul
es_df$r3_dia_4), 730.50 - 365.25)
nodules_df$r3_vdt_5 <- vdt(spherevol(nodules_df$r3_dia_4), spherevol(nodul
es_df$r3_dia_5), 1461.0 - 730.50)
```

Reader 4 - Unaided reader

```
# Starting values
nodules_df$r4_dia_1 <- nodules_df$reader_4
nodules_df$r4_vol_1 <- spherevol(nodules_df$r4_dia_1*runif(N,1-x,1+x))
lnormvalues         <- rlnorm(N, lmean.nsclc, lsd.nsclc)

# Diameter
nodules_df$r4_dia_2 <- rnorm(nodule_n, nodules_df$r0_dia_2   - (0.101*nodu
les_df$r0_dia_2)   , sd= 0.639*1.5)
nodules_df$r4_dia_3 <- rnorm(nodule_n, nodules_df$r0_dia_3   - (0.101*nodu
les_df$r0_dia_3)   , sd= 0.639*1.5)
nodules_df$r4_dia_4 <- rnorm(nodule_n, nodules_df$r0_dia_4   - (0.101*nodu
les_df$r0_dia_4)   , sd= 0.639*1.5)
```

```r
nodules_df$r4_dia_5 <- rnorm(nodule_n, nodules_df$r0_dia_5   - (0.101*nodu
les_df$r0_dia_5)   , sd= 0.639*1.5)

# Volume
nodules_df$r4_vol_2 <- spherevol(nodules_df$r4_dia_2)
nodules_df$r4_vol_3 <- spherevol(nodules_df$r4_dia_3)
nodules_df$r4_vol_4 <- spherevol(nodules_df$r4_dia_4)
nodules_df$r4_vol_5 <- spherevol(nodules_df$r4_dia_5)

# Volume doubling time
nodules_df$r4_vdt_2 <- vdt(spherevol(nodules_df$r4_dia_1), spherevol(nodul
es_df$r4_dia_2), 90)

## Warning in log(v1/v0): NaNs produced

nodules_df$r4_vdt_3 <- vdt(spherevol(nodules_df$r4_dia_2), spherevol(nodul
es_df$r4_dia_3), 365.25 - 90)

## Warning in log(v1/v0): NaNs produced

nodules_df$r4_vdt_4 <- vdt(spherevol(nodules_df$r4_dia_3), spherevol(nodul
es_df$r4_dia_4), 730.50 - 365.25)

## Warning in log(v1/v0): NaNs produced

nodules_df$r4_vdt_5 <- vdt(spherevol(nodules_df$r4_dia_4), spherevol(nodul
es_df$r4_dia_5), 1461.0 - 730.50)

## Warning in log(v1/v0): NaNs produced
```

13.9 Appendix 9: Findings of probabilistic sensitivity analyses for the cost-effectiveness analyses from the full model

Symptomatic population

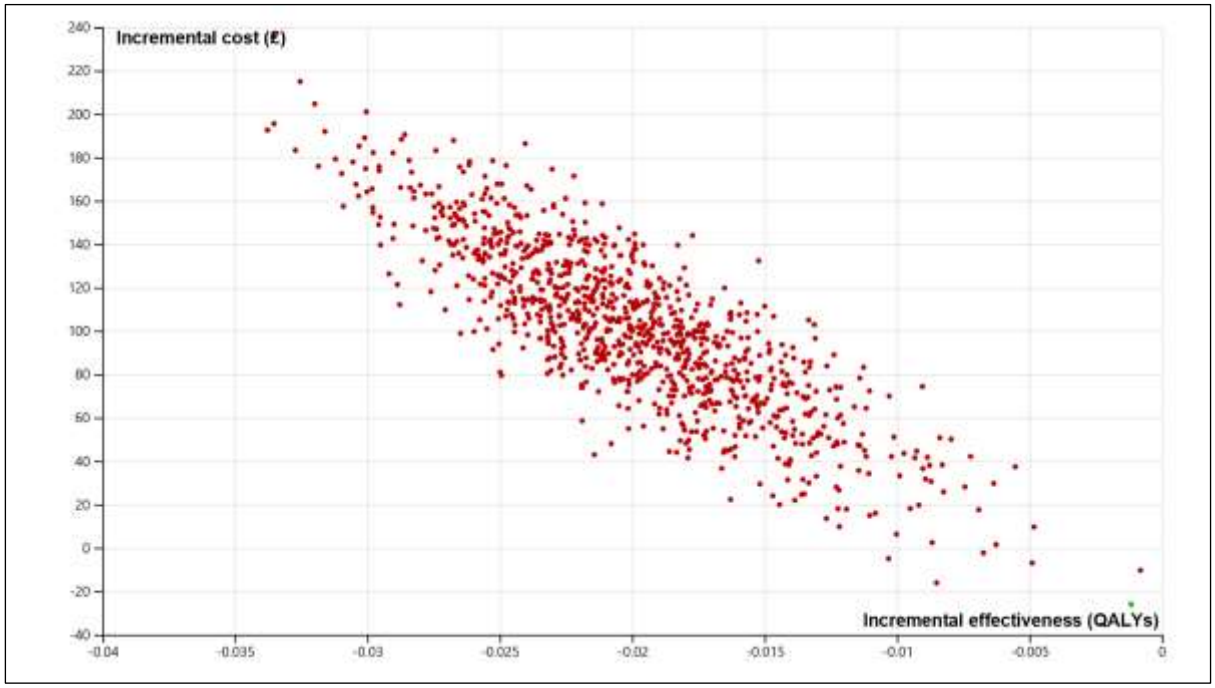


Figure 24. Incremental cost-effectiveness scatterplot for the comparison between AI-assisted radiologist reading versus unaided radiologist reading (symptomatic population)

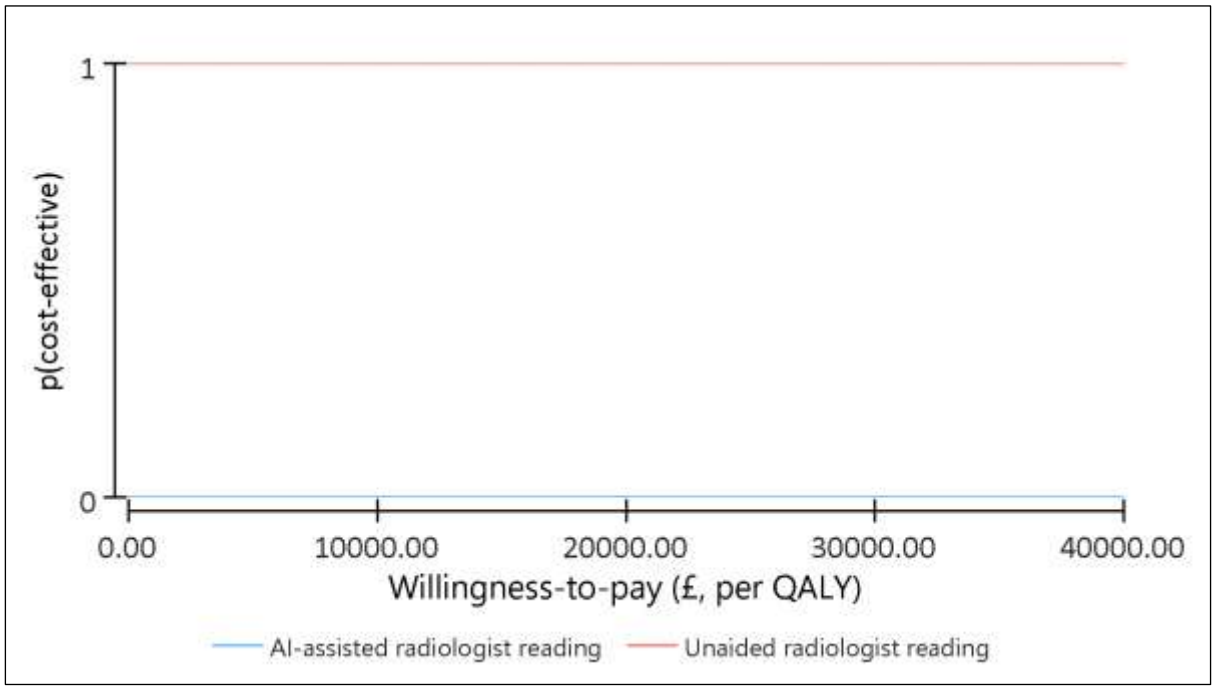


Figure 25. Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (symptomatic population)

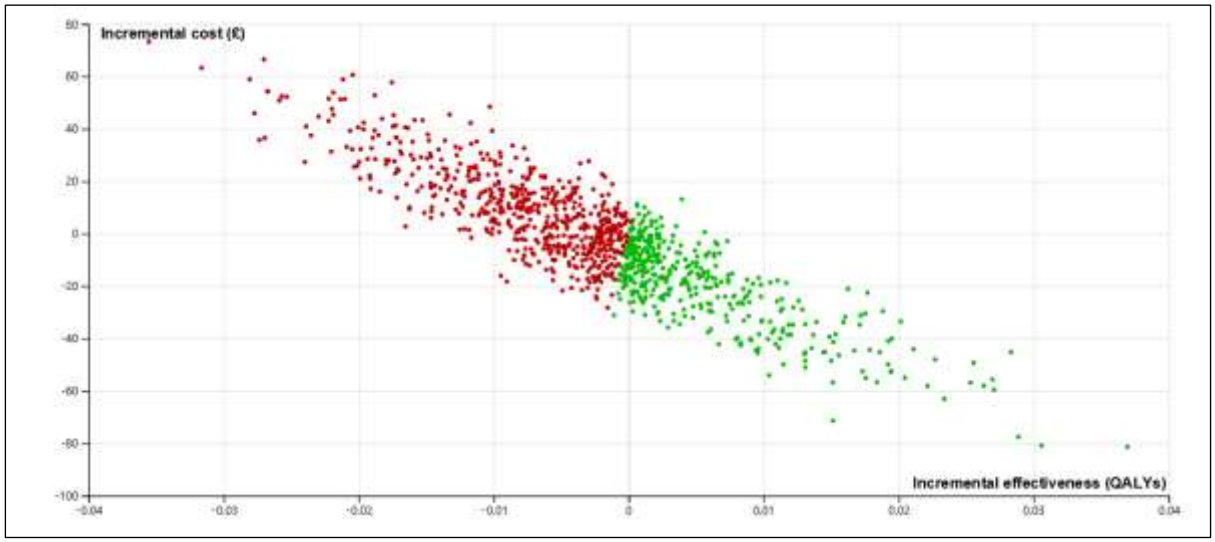


Figure 26. Incremental cost-effectiveness scatterplot for the comparison between AI-assisted radiologist reading versus unaided radiologist reading (incidental population)

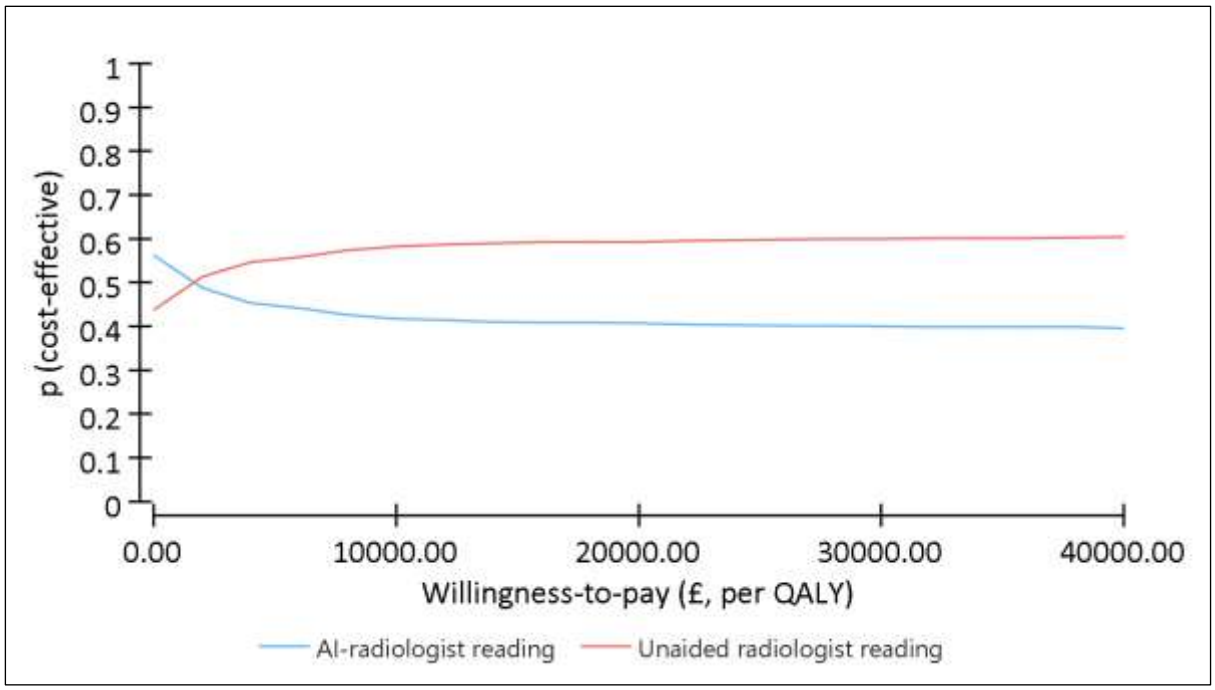


Figure 27. Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (incidental population)

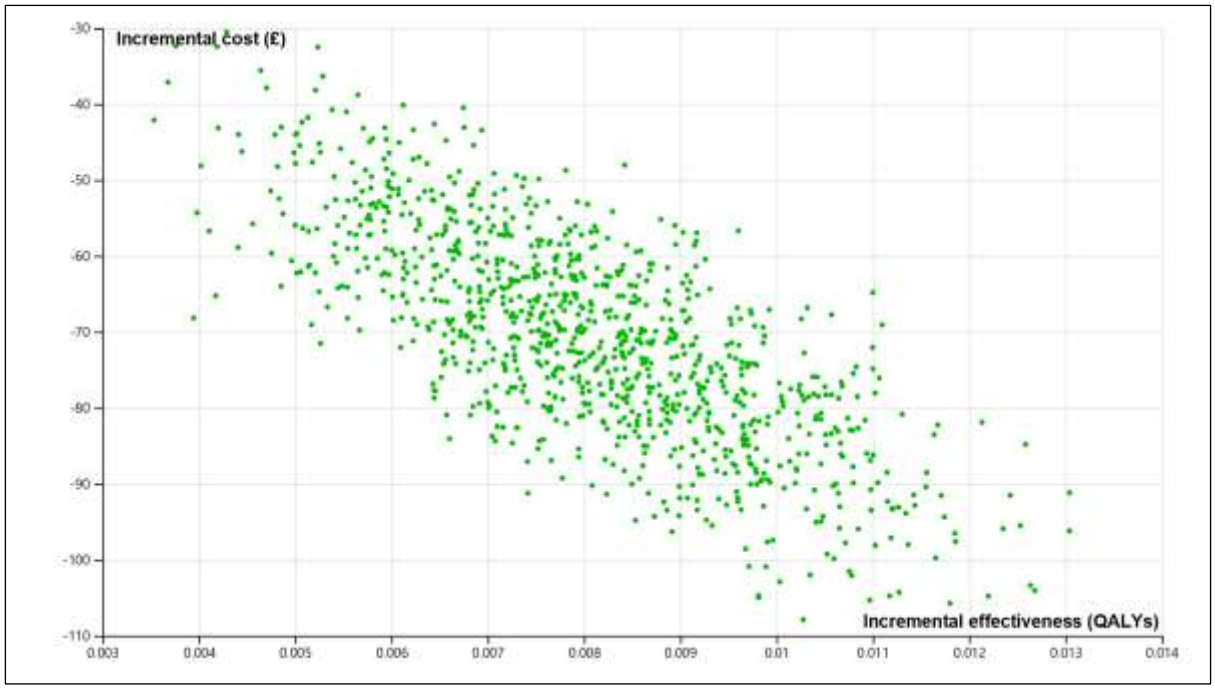


Figure 28. Incremental cost-effectiveness scatterplot for the comparison between AI-assisted radiologist reading versus unaided radiologist reading (screening population)

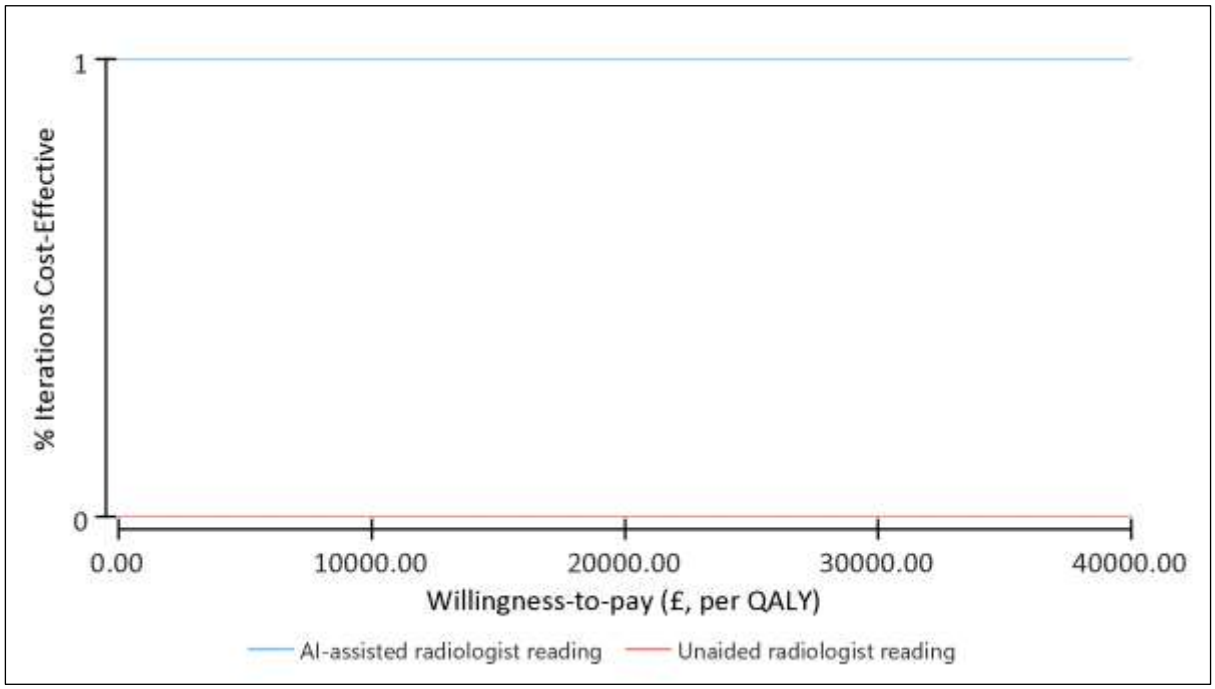


Figure 29. Cost-effectiveness acceptability curves for AI-assisted and unaided reading at different willingness-to-pay thresholds (screening population)

13.10 Appendix 10: Rationale for developing the Warwick Evidence (WE) model and comparison with the Exeter NAtural History-Based economic model of Lung cancer screening (ENaBL) model used by the National Screening Committee (NSC)

Components	ENaBL model	WE model	Justification for creating an alternative model
Purpose	To evaluate the cost-effectiveness of different screening strategies in terms of screening frequency and characteristics (age, predicted cancer risk) of target population for the screening programme .	To evaluate the cost-effectiveness of AI-assisted reading compared with unaided reading for detection and analysis of lung nodules in chest CT.	The UK NSC model evaluates the impact of different characteristics of populations being screened and different levels (frequency) of opportunities for detecting lung cancer. Although the cancer detection would also start from using CT scans to identify lung nodules, this was part of standard practice that was not evaluated in the NSC model. A de novo model was required for our assessment as we are evaluating the impact of AI assistance that operates within and across each opportunity for detecting and analysing lung nodules along the nodule management and lung cancer diagnosis pathways.
Population	Screening population with some eligibility criteria in terms of the age, smoking profile and lung cancer predicted risk according to the Liverpool Lung Project.	WE's model targeted four different populations which are from four routes: symptomatic, incidental, screening and surveillance.	The UK NSC model targets population at age 55-85 years who are currently or formerly smokers. The population who has a risk of lung cancer at 3%, 4% or 5% according to the version 2 of the Liverpool Lung Project lung cancer risk prediction model (LLPv2) are invited for screening and are included in the modelling. AI assisted chest CT image analysis can be potentially applied to wider populations in addition to the screening population.

Natural history	Searched three different natural history models developed outside of the UK as below: 1- USA The Lung Cancer Policy Model. 2- The Cancer Risk Management Model (renamed OncoSim) in Canada. 3- The Microsimulation Screening Analysis (MISCAN) in Canada The researchers developed a new natural history model called:" the Exeter NAtural History-Based economic model of Lung cancer screening (ENaBL)".	WE's model uses a previously developed natural history based on the measurement of the lung nodules growth over time. The natural history is based on Treskova et. al. (2017) The model included a natural history of a biological two-stage clonal expansion (TSCE) of the disease incorporating the nodule growth (in terms of the rate and time), which fits with the recommended lung nodule monitoring and management in the UK based on the BTS Guideline. For the purposes of this technology assessment, we needed a natural history model which enables us to track the trajectory of a lung nodule from when they are recognised as actionable nodules until they turn into malignant/cancerous phase.	UK NSC model (ENaBL) included a natural history model that does not explicitly model the growth of lung nodules. ENaBL models the patients who have a predicted risk of lung cancer at (≥ 3%, ≥ 4% and ≥ 5%) as calculated using the Liverpool Lung Project tool version 2. Therefore, the ENaBL assumed these high-risk patients/cases are in the pre-clinical phase, and undergo different rounds of LDCT screening. The simulated patients in ENaBL do not have clinical presentation of the lung cancer at the start but they can turn into cancerous phase. In WE's model, we identify people with lung nodules, then track their growth characterised by their VDT over the surveillance period as stipulated in the BTS guidelines.

Model Structure	The model includes two parts: people with no lung cancer are assumed to be at a pre-clinical state which is itself encompassing different health states according to seven considered lung cancer stages. There is a clinical state including the identical lung cancer stages. There is also a death state that can be moved into from the other states.	The model have two interrelated parts. The first part models the detection of lung nodules with classification of nodule type (solid vs sub-solid) and size categories both based on the BTS guideline. This is the same for all four interested population. The second part models the surveillance phase based on a decision tree, in which patients who have been detected with actionable lung nodules undergo scheduled monitoring by LDCT or definitive work-up and treatment. The surveillance part leads to lung cancer detection at a specific stage or being missed, and people without lung cancer being discharged after surveillance or had unnecessary investigation and treatment.	The UK NSC model structure was designed for cases who are at higher risk of lung cancer, who have already received a predicted lung cancer risk. The model structure does not take into account detection and management of lung nodules. The WE's model focuses on lung nodules detection and management rather than predefined risk of lung cancer. The need to follow the trajectory of the nodule according to their diameter and volume requires development of a model structure which can explicitly represent the nodule management pathway.

Model types	Individual Patients Simulation model with a Discrete Event Simulation (DES) framework.	Decision tree.	The UK NSC model follows the characteristics of people in terms of their smoking background, age and sex, and time dependency of the event (developing cancer). WE's model requires a structure that enables the researchers to capture the harms and benefits of AI assistance incorporated into current practice. Such benefits and harms are tied to features of different image analysis strategies such as measurement errors, discrepancies in nodule detection between human assisted by AI and human alone and the impact of using volumetric measurement provided by AI software on the surveillance process. Consequently WE's model focuses on mid-term and long-term benefits or harms of AI assistance such as reducing/increasing the number of false negative/positive, number of people requiring CT surveillance and earlier detection of cancer. We also aimed to follow the current BTS guideline. Given these, the decision tree structure was considered suitable for building the model and conducting the analysis.
BTS, British Thoracic Society, WE, Warwick Evidence, NSC, National Screening Committee; VDT, volume doubling time, biological two-stage clonal expansion (TSCE), ENaBL, the Exeter NAtural History-Based economic model of Lung cancer screening. DES, Discrete Event Simulation, AI, Artificial Intelligence, LLPV, the Liverpool Lung Project lung cancer risk prediction model.			