CADTH

CADTH RAPID RESPONSE REPORT:
SUMMARY WITH CRITICAL APPRAISAL

Artificial Intelligence for Classification of Lung Nodules: A Review of Clinical Utility, Diagnostic Accuracy, Cost-Effectiveness, and Guidelines

Service Line:	Rapid Response Service
Version:	1.0
Publication Date:	January 22, 2020
Report Length:	23 Pages

CADTH

Authors: Chantelle C. Lachance, Melissa Walter

Cite As: Artificial intelligence for classification of lung nodules: clinical utility, diagnostic accuracy, cost-effectiveness and guidelines. (CADTH rapid response report: summary with critical appraisal). Ottawa: CADTH; 2020 Jan.

ISSN: 1922-8147 (online)

About CADTH: CADTH is an independent, not-for-profit organization responsible for providing Canada's health care decision-makers with objective evidence to help make informed decisions about the optimal use of drugs, medical devices, diagnostics, and procedures in our health care system.

Funding: CADTH receives funding from Canada's federal, provincial, and territorial governments, with the exception of Quebec.

Questions or requests for information about this report can be directed to Requests@CADTH.ca

Abbreviations

ACR lung-RADS	American College of Radiology Lung CT Screening Reporting and Data System
AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
CADx	Computer-aided diagnosis
CI	Confidence interval
CNN	convolutional neural network
CRD	Centre for Reviews and Dissemination
CT	computer tomography
DLCST	Danish Lung Cancer Screening Trial
GGO	Ground-glass opacity
LIDC/IDRI	Lung Image Database Consortium and Image Database Resource Initiative
MeSH	Medical Subject Headings
PanCan	Pan-Canadian Early Detection of Lung Cancer Study
PICO	Population Intervention, Comparator, Outcome
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RF	Random forest
SD	Standard deviation
SE	Standard error
SEGvAC	an automated segmented approach
SMOTE	Synthetic minority oversampling technique
SSPN	Small solid pulmonary nodules
SVM	Support vector machine
SVM-LASSO	Support vector machine with a least absolute shrinkage and selection operator

Context and Policy Issues

A lung nodule is a small (< 30 millimetres), well defined lesion completely surrounded by pulmonary parenchyma (i.e., functional tissue of the lung).[1-5] Lung nodules are classified as solid or subsolid, and subsolid nodules are subdivided into pure ground-glass nodules (no solid component) and part-solid nodules (both ground glass and solid components).[1,5] A lesion that measures over 30 millimetres is considered a lung mass.[5] An important distinction for the patient and treatment plan is whether the presenting lung nodule(s) are benign or malignant. For lung nodules, this appropriate classification is crucial to prevent any unnecessary procedures as well as for appropriate treatment planning (e.g., biopsy, surgical resection). It has been found that the majority of lung nodules identified on computed tomography (CT) scans are benign with a prevalence of malignancy as low as one percent for Canadians with lung nodules.[5,6]

To discern whether a lung nodule is benign or malignant, the initial evaluation usually involves a radiologist using clinical and radiographic features (often from a CT scan) to determine the likelihood of malignancy; this likelihood assists in determining further management (e.g., CT surveillance, biopsy).[5] However, discerning malignancy from clinical and radiographic features can be challenging and novel methods are being considered, including artificial intelligence (AI).

AI is a branch of computer science concerned with the development of systems that can perform tasks that would usually require human intelligence, such as problem-solving, reasoning, and recognition.[7-11] AI is an umbrella term that includes a number of subfields and approaches.[7,8] AI algorithms for reading CT scans often include a machine learning system (e.g., support vector machine [SVM], artificial neural networks [deep learning, including convolutional neural network or CNN]).[8] Machine learning involves training an algorithm to perform tasks by learning from patterns in data rather than performing a task that it is explicitly programmed to do.[7,8] In order to train the machine learning program, data are divided into learning sets (i.e., human indicates if an outcome of interest is present or absent) and validation sets (i.e., system used what it learns to indicate if the outcome of interest is present or absent).[8,12]

CADTH has previously reviewed the evidence for the use of AI for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer via a Rapid Response Summary of Abstracts.[13] The aim of the current report is to summarize and critically appraise the evidence initially identified in the Summary of Abstracts, based on additional screening and review of the full text of these publications.

Research Questions

1. What is the clinical utility of artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer?
2. What is the diagnostic accuracy of artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer?
3. What is the cost-effectiveness of artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer?
4. What are the evidence-based guidelines regarding artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer?

Key Findings

Seven diagnostic case-control studies were identified regarding the diagnostic accuracy of artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer. No evidence regarding the cost-effectiveness, clinical utility or evidence-based guidelines regarding artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer were identified.

Results from the case-control studies were mixed. Two studies reported that artificial intelligence models are significantly more accurate at nodule classification when compared to radiologists classifying lung nodules using the American College of Radiologists Lung CT [computed tomography] Screening Reporting and Data System. Two studies descriptively reported that artificial intelligence models are more accurate at nodule classification compared to human observation. However, three studies reported that artificial intelligence models were comparable or had a reduced accuracy when versus human observers (statistical testing performed for one study, descriptive results provided for two studies).

Three studies descriptively reported on the sensitivity and specificity outcomes and found that artificial intelligence models had higher values for sensitivity and specificity outcomes than their respective comparators for the diagnosis of lung malignancy.

It may be premature to draw conclusions about artificial intelligence for lung nodule classification given the paucity of clinical utility, cost-effectiveness evidence and guidelines, and mixed results and inherent methodological flaws noted within the included diagnostic accuracy studies.

Methods

Literature Search Methods

A limited literature search was conducted by an information specialist on key resources including MEDLINE, the Cochrane Library, the University of York Centre for Reviews and Dissemination (CRD) databases, the websites of Canadian and major international health technology agencies, as well as a focused Internet search. The search strategy was comprised of both controlled vocabulary, such as the National Library of Medicine's MeSH (Medical Subject Headings), and keywords. The main search concepts were AI and lung nodules. No filters were applied to limit the retrieval by study type. Where possible, retrieval was limited to the human population. The search was also limited to English language documents published between January 1, 2014 and October 31, 2019.

Selection Criteria and Methods

One reviewer screened citations and selected studies. In the first level of screening, titles and abstracts were reviewed and potentially relevant articles were retrieved and assessed for inclusion. The final selection of full-text articles was based on the inclusion criteria presented in Table 1.

Table 1: Selection Criteria

Population	Patients with lung nodules (< 3 cm) suspected of having lung cancer - Patients identified during a screening program - Patients identified incidentally when having a scan for an unrelated reason - Patients with a known or suspected malignancy
Intervention	Artificial intelligence algorithms for reading computed tomography (or computerized axial tomography) scans to classify nodules, excluding detection alone (e.g., machine learning [supervised and unsupervised learning, support vector machines, random forests, black box learning], deep learning, artificial neural network [convolutional neural networks])
Comparator	Radiologist-read computed tomography (or computerized axial tomography) scan; clinical decision support tools or criteria (e.g., Vancouver Lung Cancer Risk Prediction Model)
Outcomes	Q1: Clinical utility (e.g., avoidance of biopsy, psychological distress, early or appropriate treatment, detection of cancer, survival, additional testing) Q2: Diagnostic accuracy (i.e., clinical validity: sensitivity, specificity, negative predictive value, positive predictive value, true positives, precision, accuracy, recall) Q3: Cost effectiveness (e.g., incremental cost effectiveness ratio, quality adjusted life years) Q4: Recommendations regarding the use of artificial intelligence for lung nodule classification
Study Designs	Q1&2: Health technology assessments, systematic reviews, randomized controlled trials, non-randomized studies Q3: Economic evaluations Q4: Evidence-based guidelines

Exclusion Criteria

Articles were excluded if they did not meet the selection criteria outlined in Table 1, were duplicate publications or were published prior to 2014. Guidelines with unclear methodology were also excluded.

Critical Appraisal of Individual Studies

The included non-randomized studies was critically appraised using the Downs and Black checklist[14] supplemented with QUADAS-2[15] for key questions related to diagnostic accuracy study design. Summary scores were not calculated for the included studies; rather, the strengths and limitations of each included study were described narratively.

Summary of Evidence

Quantity of Research Available

A total of 570 citations were identified in the literature search. Following screening of titles and abstracts, 556 citations were excluded and 14 potentially relevant reports from the electronic search were retrieved for full-text review. Two potentially relevant publications were retrieved from the grey literature search for full-text review. Of these 16 potentially relevant articles, nine publications were excluded for various reasons, and seven non-randomized studies met the inclusion criteria and were included in this report. Appendix 1 presents the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[16] flowchart of the study selection. Additional references of potential interest are provided in Appendix 5.

Summary of Study Characteristics

Additional details regarding the characteristics of included publications are provided in Appendix 2.

Study Design

All primary studies included in this report were diagnostic case-control studies.[17-23]

Country of Origin

The included studies originated from China (n = 3),[17,20,21] the United States (n = 3),[18,19,22] and the Netherlands (n = 1).[23]

Patient Population

The diagnostic test accuracy studies included patient data from a single institution,[17] cancer centres,[18,21] cohort data from clinical university archives,[22] large data sets, including the Lung Image Database Consortium and Image Database Resource Initiative (LIDC/IDRI) [19] and Danish Lung Cancer Screening Trial (DLCST),[23] and from a combination of open-source data sets and multi-centre data sets.[20] When considering all included primary studies, the number of lung nodules used to perform analyses ranged from 31 to 300.[17-23]

Interventions, Comparators, and Reference Standard

The interventions (index test) of interest for the diagnostic test accuracy studies were AI algorithms for reading CT scans in order to classify lung nodules as either benign or malignant and the method of AI used varied widely (see Appendix 2).[17-23] As an example,

one study included 11 different interventions by enabling 10 groups of participants to apply their own computerized classification system methods (e.g., SVM, random forest [RF], CNN) to a dataset to determine which is best for classifying benign and malignant lung nodules.[22]

The comparators of interest for the included studies comprised of radiologists,[18,20-23] sometimes in combination with other clinicians (radiology residents, thoracic surgeons, and respiratory doctors)[18,20,23] and two studies had radiologists classify nodules into categories according to the common American College of Radiologists Lung CT Screening Reporting and Data System (ACR lung-RADS).[17,19]

The reference standard for the included studies involved histopathological confirmation (confirmed pathology by biopsy, surgical resection or bronchoscopy for malignant and some benign nodules),[17-23] nodule stability (i.e., long-term follow-up for benign nodules),[17,19,22] laboratory and microbiological examination (for benign nodules)[20], nodule resolution (benign nodules)[22] or lesional progression or response (for limited malignant and benign nodules).[19]

Outcomes

The outcomes of interest for the diagnostic test accuracy studies were sensitivity,[17,19,20] specificity,[17,19,20] accuracy,[17,19-21] and area under receiver operating characteristic curve (AUC).[17-19,22,23] AUC is defined as a sensitivity versus specificity metric for measuring the performance of binary classifiers; the area under the curve is equal to the probability that a randomly chosen positive sample ranks above a randomly chosen negative one or is regarded to have a higher probability of being positive.[24] An AUC of 1 represents an algorithm or test with 100% of its classifications being correctly classified. An AUC of 0.5 represent an algorithm or test that performed no better than chance.

Summary of Critical Appraisal

Additional details regarding the strengths and limitations of included publications are provided in Appendix 3.

Diagnostic test accuracy studies

The quality of evidence from the included diagnostic test accuracy studies[17-23] were assessed using the Downs and Black Checklist[14] supplemented with QUADAS-2.[15] Overall, the quality of the included studies was variable and the strengths and weakness of each study can be found in Table [17-23] Most studies adequately described objectives, intervention (index test), comparator and reference standard;[17-23] conducted the index test results and reference standard independently;[18-23] described the nodule characteristics used for the experiment,[17,19-23] and declared actual or potential conflicts of interest[17-20,22,23] and sources of funding.[17-23] However, all studies used a case-control study design, and all but one study obtained the CT dataset retrospectively (Zhang and colleagues[20] prospectively collected CT images). By using a case-control design and only including patients with confirmed diagnoses, difficult to diagnose patients may have been missed from the analysis, possibly inflating test accuracy results. Three studies included clinicians other than board certified radiologists (radiology residents, thoracic surgeons, pulmonologists) as part of the comparator group;[18,20,23] it is unclear how common in clinical practice for these clinicians to classify lung nodules. No studies reported that the study was prospectively registered or followed a priori protocol,[17-23] and eight studies did not describe sample size calculations.[17-19,22,23] These limitations should be considered when interpreting the results.

Additional details regarding the strengths and limitations of included publications are provided in Appendix 3.

Summary of Findings

Appendix 4 presents a table of the main study findings and authors' conclusions.

Clinical Utility of AI for Nodule Classification

No relevant evidence regarding the clinical utility of AI for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer; therefore, no summary can be provided.

Diagnostic Accuracy of AI for Nodule Classification

Seven diagnostic test accuracy studies were identified regarding AI for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer.[17-23]

Sensitivity

Three studies reported on the sensitivity of their AI intervention.[17,19,20] Zhang and colleagues (2019) reported a sensitivity of 96.0% (95% confidence interval (CI), 88.3% to 100.0%) with a trained three-dimensional CNN model compared to a sensitivity of 81.3% (95% CI, 66.0% to 96.6%) for manual assessment; it is unclear if this difference is statistically signifcant.[20] Two studies compared their AI model to radiologists classifying nodules into categories according to ACR-lung RADS structured reporting system[17,19] and found their AI models had a higher sensitivity: 81% for a radiomic prediction model (versus 47.6% for ACR-lung RADS)[17] and a 87.2 ± 1.4% SVM-LASSO model (which is support vector machine with a least absolute shrinkage and selection operator; versus 80.5% for ACR-lung RADS).[19] It is unclear if this difference was statistically significant.

Specificity

The same three studies that reported on sensitivity also reported on their specificity of their AI intervention.[17,19,20] Zhang et al. (2019) reported a specificity of 88.0% (95% CI, 76.0% to 100.0%) with a trained CNN model compared to a specificity of 77.9% (95% CI, 61.6% to 94.1%) for manual assessment.[20] Two studies compared their AI model to radiologists classifying nodules using ACR-lung RADS[17,19] and found their models had a higher specificity: 92.2% for a radiomic prediction model (versus 84.4% for ACR lung-RADS)[17] and a 81.2 ± 3.2% SVM-LASSO model (versus 61.3% for ACR lung-RADS).[19] It is unclear if any of these differences were statistically significant.

Accuracy and AUC

All included studies reported on accuracy as a percentage of correct classification of all nodules evaluated,[21 20] as an AUC value[18,22,23] or both.[17,19] Four studies found AI achieved a higher accuracy measurement for nodule classification compared to human observation and three studies did not observe higher accuracy, described below.

Zhang et al. (2019) descriptively reported (i.e., no statistical testing) an accuracy of 92.0% with a trained CNN model compared to an accuracy of 79.6% for manual assessment.[20] Thus, the authors concluded that the CNN showed to perform better that the manual assessment.[20] Moreover, Gong and colleagues[21] descriptively reported an accuracy of 61.3% with their proposed computer-aided diagnosis (CADx) scheme method for classifying

ground-glass opacity (GGO) nodules compared to human observation (range: 53.1 to 56.3%); though, these findings were not statistically compared. The two studies that compared their AI model to radiologists using the ACR-lung RADS[17,19] found their models also had a higher accuracy. Mao et al. (2019)[17] found their radiomic prediction model was significantly more accurate than ACR lung-RADS (89.8% versus 76.5%, $P < 0.01$) and AUC values were also higher (0.97 versus 0.77). The authors concluded that a radiomic model can improve the accuracy in predicting malignancy of small solid pulmonary nodules. Likewise, Choi et al. (20)[19] reported and concluded the SVM-LASSO model was significantly more accurate than ACR lung-RADS (84.6% versus 72.2%, a 12% increase, $P = 0.026$). The AUC values reported were also higher in favour of their SVM-LASSO model: 0.89 ± 0.01 versus 0.77 for ACR lung-RADS.[19]

Three studies reported AI interventions were comparable or had a reduced accuracy relative to their comparators.[18,22,23] Alilou et al. (2017)[18] found that the AUC was lower for the fully automated segmentation-based classifier versus manually classified nodules (0.64 versus 0.72), which may be attributable to nodule segmentation errors on account of the model, but no statistical comparison was performed. For risk-assessment of nodules of all sizes, van Riel et al. (2017)[23] found no significant difference in performances for benign and malignant nodule classification between the computer model and human observers ($P = 0.184$), but human observers performed better when nodules were matched in size $P < 0.001$). Finally, one study reported on 11 different models and found all models performed worse at nodule classification when compared to the mean AUC value across six radiologists.[22]

Cost-Effectiveness of AI for Nodule Classification

No relevant evidence regarding the cost-effectiveness of AI for nodule classification was identified; therefore, no summary can be provided.

Guidelines

No relevant guidelines regarding AI for nodule classification was identified; therefore, no summary can be provided.

Limitations

A number of limitations were identified in the critical appraisal as shown in Appendix 3 (Table 3); however, additional limitations exist. The main limitations of this review are related to the heterogeneity of the AI interventions examined and the generalizability of findings. Many different types of AI interventions were explored, and it is challenging to narratively synthesize these findings together as they are different models with different features and construct. All diagnostic test accuracy studies were case-control studies with associated methodological limitations.[17-23] In addition, there was an overall lack of comparative statistical analyses between treatment groups which prevents making strong conclusions for or against AI models for the classification of lung nodules. The applicability of the evidence to the Canadian setting is unclear since all studies were conducted outside of Canada. Finally, no relevant clinical utility studies, cost-effectiveness studies or guidelines were identified suggesting a lack of research regarding these two areas on AI for nodule classification. These limitations warrant the use of caution when interpreting the findings of this report.

Conclusions and Implications for Decision or Policy Making

This report identified evidence about the diagnostic accuracy of AI for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer. Seven case-control studies were identified from the search. These studies explored sensitivity, specificity, accuracy, and AUC outcomes to discern whether AI models are superior to conventional methods (i.e., human observation). Results from the included case-control studies varied. Two studies reported that AI models were significantly more accurate at classifying lung nodules when compared to radiologists classifying lung nodules using the ACR lung-RADS. Two additional studies descriptively reported (i.e., no statistical testing provided) that AI models classified nodules more accurately compared to human observation. However, three studies reported that AI models were comparable or had a reduced accuracy versus human observers (statistical testing performed for one study, descriptive results provided for two studies). Three studies descriptively reported on the sensitivity and specificity outcomes for the classification of lung malignancy and found that these values were higher for AI models versus their respective comparators. It is possible that the heterogeneity of AI models evaluated across studies contributed to this variability in the findings.

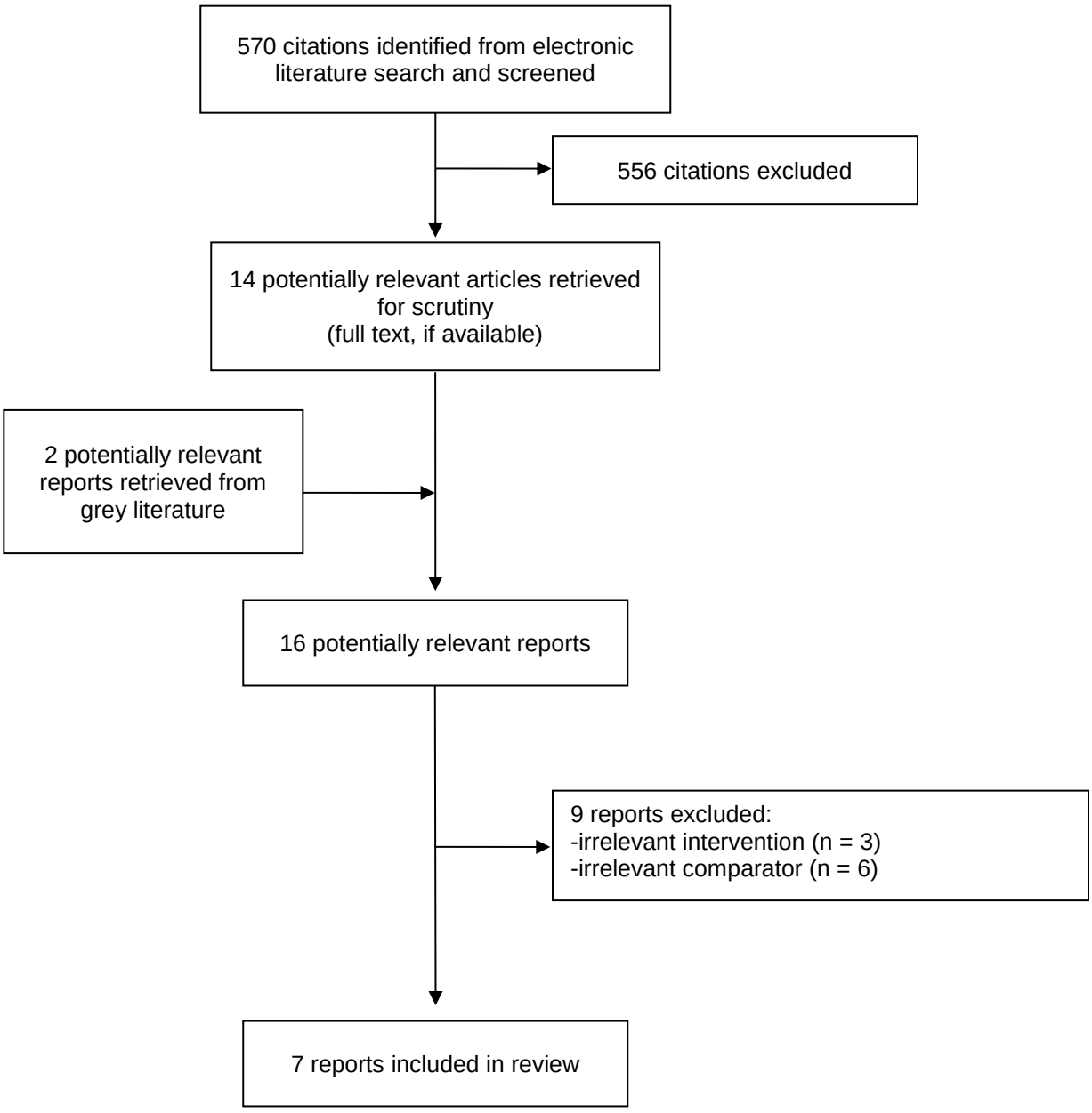
No evidence was identified regarding the clinical utility, cost-effectiveness or guidelines for the use of AI for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer.

It may be premature to draw conclusions about AI for lung nodule classification given the paucity of clinical utility and cost-effectiveness evidence and guidelines, and mixed results and inherent methodological flaws noted within the included diagnostic accuracy studies. Additional studies of high methodological quality may further aid in making definitive conclusions about AI for lung nodule classification.

References

1. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society *Radiol.* 2017;284(1):228-243.
2. Ost D, Fein AM, Feinsilver SH. Clinical practice. The solitary pulmonary nodule. *N Engl J Med.* 2003;348(25):2535-2542.
3. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest.* 2013;143(5 Suppl):e93S-e120S.
4. Tuddenham WJ. Glossary of terms for thoracic radiology: recommendations of the Nomenclature Committee of the Fleischner Society. *AJR: Am J Roentgenol.* 1984;143(3):509-517.
5. Weinberger SE, McDermott S. Diagnostic evaluation of the incidental pulmonary nodule. In: Post TW, ed. Waltham (MA): UpToDate; 2019 Jun: www.uptodate.com. Accessed 2020 Jan 22.
6. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med.* 2013;369(10):910-919.
7. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *Radiographics.* 2017;37(7):2113-2131.
8. Mason J, Morrison A, Visintini S. An overview of clinical applications of artificial intelligence.(*CADTH issues in emerging health technologies*). Ottawa (ON): CADTH; 2018 Sep: https://www.cadth.ca/sites/default/files/pdf/eh0070_overview_clinical_applications_of_AI.pdf. Accessed 2020 Jan 15.
9. Ogilvie K, Eggerton A. Challenge ahead : integrating robotics, artificial intelligence and 3D printing technologies into Canada's healthcare systems. *Report of the Standing Senate Committee on Social Affairs, Science and Technology*; 2017; http://publications.gc.ca/site/eng/9.846477/publication.html, Accessed 2020 Jan 15.
10. Fogel AL, Kvedar JC. Benefits and risks of machine learning decision support systems. *JAMA.* 2017;318(23):2356-2356.
11. Beam AL, Kohane IS. Translating artificial intelligence into clinical care. *JAMA.* 2016;316(22):2368-2369.
12. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health.* 2015;18(2):137-140.
13. Hill S, Walter M. Artificial intelligence for classification of lung nodules: clinical utility, diagnostic accuracy, cost-effectiveness, and guidelines. (*CADTH rapid response report: summary of abstracts)*. Ottawa (ON): CADTH; 2019 Nov: https://www.cadth.ca/sites/default/files/pdf/htis/2019/RB1411%20AI%20Lung%20Cancer%20Final.pdf. Accessed 2020 Jan 15.
14. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health.* 1998;52(6):377-384. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1756728/pdf/v052p00377.pdf. Accessed 2020 Jan 15.
15. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529-536.
16. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol.* 2009;62(10):e1-e34.
17. Mao L, Chen H, Liang M, et al. Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose CT screening. *Quant.* 2019;9(2):263-272.
18. Alilou M, Beig N, Orooji M, et al. An integrated segmentation and shape-based classification scheme for distinguishing adenocarcinomas from granulomas on lung CT. *Med Phys.* 2017;44(7):3556-3569.
19. Choi W, Oh JH, Riyahi S, et al. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. *Med Phys.* 2018;45(4):1537-1549.
20. Zhang C, Sun X, Dang K, et al. Toward an expert level of lung cancer detection and classification using a deep convolutional neural network. *Oncologist.* 2019;24(9):1159-1165.
21. Gong J, Liu J, Hao W, Nie S, Wang S, Peng W. Computer-aided diagnosis of ground-glass opacity pulmonary nodules using radiomic features analysis. *Phys Med Biol.* 2019;64(13):135015.
22. Armato SG, 3rd, Drukker K, Li F, et al. LUNGx Challenge for computerized lung nodule classification. *J Med Imaging (Bellingham).* 2016;3(4):044506.
23. van Riel SJ, Ciompi F, Winkler Wille MM, et al. Malignancy risk estimation of pulmonary nodules in screening CTs: comparison between a computer model and human observers. *PLoS One.* 2017;12(11):e0185032.
24. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18(8):500-510.

Appendix 1: Selection of Included Studies



Appendix 2: Characteristics of Included Publications

Table 2: Characteristics of Included Primary Clinical Studies

First Author, Publication Year, Country	Objective and Study Design	Population Characteristics	Intervention (Index Test), Comparator(s), and Reference Standard	Clinical Outcomes
Gong, 2019[21] China	**Objective:** to develop a CT based radiomic feature analysis approach for the diagnosis of ground-glass opacity pulmonary nodules, and to assess the performance of CADx in classifying benign and malignant nodules associated with histopathological subtypes **Design:** diagnostic case-control study	Relevant to this report, histopathology-confirmed ground-glass opacity nodules collected from a cancer center **n = 31 nodules (20 benign, 21 malignant) from 27 patients with stage I non-small cell lung cancer**	**Intervention:** CADx scheme by using radiomic features analysis. Specifically, A LOOCV method was used to build the model. The model was embedded with a Relied feature selection, SMOTE and a machine learning classifier (i.e., SVM) **Comparator:** 2 radiologists **Reference standard:** histopathological result of each nodule confirmed after surgical resection	• Overall classification accuracy
Mao, 2019[17] China	**Objective:** to assess the usefulness of a quantitative radiomic model for predicting malignancy in small solid pulmonary nodules **Design:** diagnostic case-control study	Malignant and benign small solid pulmonary nodules detected in baseline low-dose CT screening **n = 98 in the validation data set**	**Intervention:** Radiomic predictive model **Comparator**: 2 radiologists that classified nodules into four categories according to ACR lung-RADS structured reporting system **Reference standard:** pathological confirmation (malignant nodules), long term follow-up or pathological diagnosis (benign nodules)	• Sensitivity • Specificity • Accuracy (correct classification rate) • AUC
Zhang, 2019[20] China	**Objective**: to integrate a deep learning algorithm to detect and classify pulmonary nodules from CT images	Clinical CT data was obtained through open-source data sets and multi-center data sets **n = 50 thoracic CT images for validation (25 benign, 25 malignant)**	**Intervention:** 3D CNN **Comparator:** Manual assessments done by different ranks of doctors including licensed radiologists, thoracic surgeons, and respiratory doctors with more than 5	• Sensitivity • Specificity • Accuracy

First Author, Publication Year, Country	Objective and Study Design	Population Characteristics	Intervention (Index Test), Comparator(s), and Reference Standard	Clinical Outcomes
	Design: diagnostic case-control study		years of attending doctor work experience (n = 25) **Reference standard:** confirmed pathology for malignant nodules and laboratory and microbiological examination for benign nodules	
Choi, 2018[19] United States	**Objective:** to develop a radiomics prediction model to improve pulmonary nodule malignancy classification in low-dose CT; to compare the model with the ACR lung-RADS for early detection lung cancer **Design:** diagnostic case-control study	Pulmonary nodules from the LIDC/IDRI **n = 72 (31 benign, 41 malignant)**	**Intervention:** A prediction model was constructed by using an SVM-LASSO model **Comparator:** ACR Lung-RADS categorization based on the pulmonary nodule contour and annotations made by the 4 radiologists **Reference standard:** Confirmed pathology by biopsy or surgical resection (malignant nodules), stability at 2-year follow-up (benign nodules), or lesional progression or response (for limited malignant and benign nodules)	• Sensitivity • Specificity • Accuracy • AUC
Alilou, 2017[18] United States	**Objective:** to evaluate 3D shape features for discriminating benign from malignant nodules on lung CT images; to present an integrated framework for segmentation, feature characterization and classification of these nodules on CT **Design:** diagnostic case-control study	149 patients with lung nodules found on CT scan from 2 different institutions (82 and 67 patients) **n = 149 (69 benign, 80 malignant)**	**Intervention:** An automated nodule segmentation approach (SEGvAC): an SVM classifier was combined with a feature selection scheme **Comparator:** 2 expert radiologists and 1 pulmonologist **Reference standard:** histopathologic confirmation (either via surgical wedge resection, CT-guided biopsy or bronchoscopy)	• AUC

First Author, Publication Year, Country	Objective and Study Design	Population Characteristics	Intervention (Index Test), Comparator(s), and Reference Standard	Clinical Outcomes
van Riel, 2017[23] The Netherlands	**Objective**: to compare human observers to a mathematically derived computer model for differentiating between malignant and benign pulmonary nodules detected on baseline CT scans **Design:** diagnostic case-control study	300 participants from the complete screening data set (CT scans) under the following conditions: 60 participants with 1+ malignant nodule that had been found in the complete DLCST (group 1), 120 participants with at 1+ benign nodule randomly selected from the whole screening data set (group 2), and 120 participants also randomly selected from the whole screening data set but under the condition that they showed at 1+ benign nodule with a diameter in the range of 3 to 16 mm with a preference for lesions larger than 10 mm (group 3) **n = 300**	**Intervention:** A mathematically derived computer model called PanCan that assigned a malignancy probability score to each nodule **Comparator:** 11 clinicians (4 board certified radiologist, 7 radiology residents and pulmonologists) to assign a malignancy probability score to each nodule **Reference standard:** histopathologic confirmation	• AUC
Armato, 2016[22] United States	**Objective:** to describe and report the performance of the LUNGx Challenge for classifying benign and malignant lung nodules based on CT scans **Design:** diagnostic case-control study	Nodules selected on approximate size matching from 2 cohorts **n = 73 (37 benign, 36 malignant)**	**Intervention:** 10 groups that applied their own computerized classification system methods (e.g., SVM, RF, CNN, i.e., 11 computerized models) **Comparator:** 6 radiologists **Reference standard:** Confirmed pathology by pathological assessment (malignant, benign), nodule resolution (benign), or nodule stability for at least 2 years (benign)	• AUC

3D = three-dimensional; ACR-lung RADS = American College of Radiology Lung CT Screening Reporting and Data System; AUC = area under the receiver operating characteristic curve; CADx = computer-aided diagnosis; CNN = convolutional neural network; CT = computer tomography; DLCST = Danish Lung Cancer Screening Trial; LIDC/IDRI = lung image database consortium and image database resource initiative; LOOCV = leave-one-out cross-validation; PanCan = Pan-Canadian Early Detection of Lung Cancer Study; RF = random forest; SEGvAC = an automated segmented approach; SMOTE = synthetic minority oversampling technique; SVM = support vector machine; SVM-LASSO = support vector machine with a least absolute shrinkage and selection operator.

Appendix 3: Critical Appraisal of Included Publications

Table 3: Strengths and Limitations of Primary Clinical Studies using Downs and Black[14] supplemented with QUADAS-2[15]

Strengths	Limitations
Gong, 2019[21]	
- Objectives, intervention, and main outcomes of the study clearly described - Authors provided rationale for using selected sample size to build the model - Characteristics of the 31 nodules described in detail - The index test results and reference standard were conducted independently - The reference standard (histopathological result of each nodule confirmed after surgical resection) likely to correctly classify the target condition (i.e., benign versus malignant lung nodules) - Inclusion and exclusion criteria were included - Authors did report the source of funding for the study	- There was no information provided to suggest the study was prospectively registered - Used a case-control study design - The CT images which included 31 lung nodules from 27 patients were retrospectively collected (versus consecutive patients, along with their CT scans, being prospectively enrolled) - It is unclear if inappropriate exclusion criteria were avoided (e.g., not including difficult-to-diagnose patients) - It was unclear if the radiologists that participated were representative of the source population (e.g., level of experience classifying lung nodules) - It was unclear whether the radiologists were blinded to the patient's diagnosis when assessing the CT images - For findings relevant to this report, no statistical testing was performed and rationale for this choice not provided; therefore, no actual probability values or estimates of random provided - Authors did not declare any or potential conflict of interests
Mao, 2019[17]	
- Objectives, intervention, and main outcomes of the study clearly described - Authors used a computer-generated random numbers to assign cases - Appropriate statistical tests used to assess outcomes - Characteristics of the 98 nodules in the validation data set described in detail - The reference standard (stability during long term follow-up, pathological diagnosis) was likely to correctly classify the target condition (i.e., benign versus malignant lung nodules) - Though the index test and reference standard included the same two radiologists (i.e., not independently, they were blinded to the final diagnosis when they classified the nodules of the validation sets (i.e., classified nodules into four categories according to ACR Lung-RADS). - Actual probability values (*P* values) reported for accuracy outcome - Authors declared no competing interests. - Authors did report the source of funding for the study	- There was no information provided to suggest the study was prospectively registered - Used a case-control study design - CT dataset obtained retrospectively - Study investigators may have implemented inappropriate exclusion criteria (e.g., nodules were excluded with obscure border, which limited ability to perform robust segmentation). - Index test and reference standard included the same two radiologists (but blinded, as mentioned under strengths) - No sample size calculation for statistical power provided; the authors allude the study included a relatively small sample size - Estimates of the random variability not provided when applicable

Strengths	Limitations
Zhang, 2019[20]	
- Objectives, intervention, and main outcomes of the study clearly described - Authors reported sample size calculations conducted - Main findings of the study adequately described - Characteristics of the 50 test nodules well described - The reference standard (confirmed pathology for malignant nodules and laboratory and microbiological examination for benign nodules) was likely to correctly classify the target condition (i.e., benign versus malignant lung nodules) - The index test results and reference standard were conducted independently - Estimates of the random variability provided as standard deviation, as appropriate - Authors declared conflicts of interest, which included one author who provides consulting (AstraZeneca), receives funding (Roche); received honoraria (AstraZeneca, Roche, Eli Lilly, Pfizer, Sanofi) - Authors did report the source of funding for the study	- There was no information provided to suggest the study was prospectively registered - Used a case-control study design - It is unclear if inappropriate exclusion criteria were avoided (e.g., not including difficult-to-diagnose patients) - It was unclear if the radiologists that participated were representative of the source population (e.g., level of experience classifying lung nodules and how common it is for thoracic surgeons and respiratory doctors to classify nodules) - Did not perform statistical comparisons for outcomes applicable to this report.
Choi, 2018[19]	
- Objectives, intervention, and main outcomes of the study clearly described - Main findings of the study adequately described - The reference standard (confirmed pathology) was likely to correctly classify the target condition (i.e., benign versus malignant lung nodules) - The index test results and reference standard were conducted independently - Characteristics of the 72 nodules described - Actual probability values (*P* values) reported for main outcome (i.e. accuracy) - Estimates of the random variability provided as standard deviation, as appropriate - Authors declared they had no conflicts of interest - Authors did report the source of funding for the study	- There was no information provided to suggest the study was prospectively registered - Used a case-control study design - CT dataset obtained retrospectively - It is unclear if inappropriate exclusion criteria were avoided (e.g., not including difficult-to-diagnose patients) - No sample size calculation for statistical power provided

Strengths	Limitations
Alilou, 2017[18]	
- Objectives, intervention, and main outcomes of the study clearly described - The reference standard (confirmed pathology) was likely to correctly classify the target condition (i.e., benign versus malignant lung nodules) - The index test results and reference standard were conducted independently (i.e., the readers were blinded to the true histopathologic diagnosis for the 20 cases that were compared to automated method) - Main findings of the study adequately described - Number of patients included, and characteristics of the patients included in the study described - Authors declared no conflict of interests - Authors acknowledge the source of funding for the study	- There was no information provided to suggest the study was prospectively registered - Used a case-control study design - CT dataset obtained retrospectively - It was unclear whether the clinicians were representative of the source population (e.g., how common it is for pulmonologists to classify nodules) - No sample size calculation for statistical power provided; authors allude that the sample was small - It is unclear whether appropriate statistical tests used to assess outcomes as statistical methods not explicitly described - Characteristics of the nodules not described in detail
van Riel, 2017[23]	
- Objectives, intervention, and main outcomes of the study clearly described - The reference standard (confirmed pathology) was likely to correctly classify the target condition (i.e., benign versus malignant lung nodules) - The index test results and reference standard were conducted independently - The data set involved randomly selected nodules based on criteria explicitly described in the publication - Appropriate statistical tests used to assess outcomes - Main findings of the study adequately described - Actual probability values (*P* values) reported - Characteristics of the 300 nodules described in detail - Ranges provided for all findings that reported means or medians - Authors declared competing interests, which included five authors whom either hold grants (sources: Terry Fox Research Institute, Toshiba, MeVis Medical Solutions, Thirona), copyrights (Pan-Canadian lung nodule malignancy risk calculator for commercial users), licensing (non-exclusive license was issued to Phillips), pending patents (Riverain Technologies), and reports personal fees (companies: Bayer, Toshiba). - Authors did report the source of funding for the study	- There was no information provided to suggest the study was prospectively registered - Used a case-control study design - CT dataset obtained retrospectively - It was unclear if the radiologists that participated were representative of the source population (e.g., level of experience classifying lung nodules and how common it is for radiology residents and pulmonologists to classify nodules) - No sample size calculation for statistical power provided

Strengths	Limitations
Armato, 2016[22]	
- Objectives, intervention, and main outcomes of the study clearly described - The reference standard (confirmed pathology) was likely to correctly classify the target condition (i.e., benign versus malignant lung nodules) - The index test results and reference standard were conducted independently - Appropriate statistical tests used to assess outcomes - Challenge test set described in detail - Main findings of the study adequately described - Characteristics of the 73 nodules in the LUNGx - Actual probability values (*P* values) reported - Estimates of the random variability provided as standard deviation - Authors declared competing interests, which included four authors whom receive royalties and licensing fees through the University of Chicago related to computer-aided diagnosis. - Authors did report the source of funding for the study	- There was no information provided to suggest the study was prospectively registered - Used a case-control study design - CT dataset obtained retrospectively - It was unclear if the radiologists that participated were representative of the source population - No sample size calculation for statistical power provided; the authors allude that there was low statistical power for the Challenge

ACR lung-RADS = American College of Radiologists Lung CT Screening Reporting and Data System; CT = computer tomography.

Appendix 4: Main Study Findings and Authors' Conclusions

Table 4: Summary of Findings of Included Primary Clinical Studies

Main Study Findings	Authors' Conclusion
Gong, 2019[21]	
Accuracy (i.e., classification accuracies of GGO nodules) • Proposed CADx scheme method: 61.3% (AUC = 0.74 ± 0.05 [95% CI, 0.65 to 0.83]) • Radiologist 1 (5-year experience): 53.1% • Radiologist 2 (14-year experience): 56.3%	"In this study, we developed a CADx scheme to classify GGO nodules in CT images, and investigated the associations between performance changes with histopathological subtypes of GGO nodules. The experimental and data analysis results demonstrated that (1) comparing with radiologists' diagnosis scores, radiomic features analysis approach yielded higher performance in diagnosing GGO nodules, (2) a consistently positive trend between the CADx scheme performance and invasive grade of GGO nodules. Thus, this study provides new scientific evidence that radiomic features analysis based CADx scheme can improve the performance in discriminating different subtypes of GGO nodules. To build a high-performance classification scheme for GGO nodules, we should add the number of non-invasive and pre-invasive nodules into a large diverse training data set"(*p10*)
Mao, 2019[17]	
Sensitivity (true positive rate) • Radiomic prediction model: 81.0% • ACR-lung RADS: 47.6% **Specificity (true negative rate)** • Radiomic prediction model: 92.2% • ACR-lung RADS: 84.4% **Accuracy** • Radiomic prediction model (89.8%) significantly higher than ACR-lung RADS (76.5%, $P < 0.01$) • Related descriptive results: 6 cases were misdiagnosed in both approaches, but ACR lung-RADS misdiagnosed an additional 17 cases. **AUC** • Radiomic prediction model: 0.97 • ACR-lung RADS: 0.77	"A radiomic model based on baseline low-dose CT screening for lung cancer can improve the accuracy in predicting malignancy of SSPNs." (*p263*) "In conclusion, with this radiomic model, it is possible to predict malignant solid nodules 6–15 mm in diameter at baseline LDCT screening for lung cancer."(*p270*)
Zhang, 2019[20]	
Sensitivity • Trained model: 96.0% (95% CI, 88.3% to 100.0%) • Doctors: 81.3% (95% CI, 66.0% to 96.6%) **Specificity** • Trained model: 88.0% (95% CI, 76.0% to 100.0%) • Doctors: 77.9% (95% CI, 61.6% to 94.1%) **Accuracy**	"Under the companion diagnostics, the three-dimensional CNN with a deep learning algorithm may assist radiologists in the future by providing accurate and timely information for diagnosing pulmonary nodules in regular clinical practices."(*p1159*)

Main Study Findings	Authors' Conclusion
• Trained model: 92.0% • Doctors: 79.6%	
Choi, 2018[19]	
Sensitivity • SVM-LASSO model (two features: bounding box anterior-posterior dimension [BB_AP], standard deviation of inverse difference moment [SD_IDM]): 87.2 ± 1.4% • Lung-RADS (four features: size, type, calcification, spiculation): 80.5% **Specificity** • SVM-LASSO model (two features): 81.2 ± 3.2% • Lung-RADS (four features): 61.3% **Accuracy** • SVM-LASSO model (two features): 84.6% • Lung-RADS (four features): 72.2% • This results in a 12% increase in performance in favour of the intervention (*P* = 0.026) • Related, descriptive results**:** Lung-RADS misclassified 19 cases as it was mainly based on PN size; SVM-LASSO model correctly classified 10 of these cases by combining a size (BB_AP) feature and a texture (SD_IDM) feature. **AUC** • SVM-LASSO model (two features): 0.89 ± 0.01 • Lung-RADS (four features): 0.77 AUC	"We developed an SVM-LASSO model to predict malignancy of PNs with two CT radiomic features. We demonstrated that the model achieved an accuracy of 84.6%, which was 12.4% higher than Lung-RADS."(*p2*)
Alilou, 2017[18]	
AUC • Automated segmentation-based classifier: 0.64 • Manual classification: 0.72	"The major finding of this study was that certain shape features appear to differentially express between granulomas and adenocarcinomas and thus computer extracted shape cues could be used to distinguish these radiographically similar pathologies"(*p3*) "Major findings of our study were (a) both manual and automated segmentation approaches yielded a similar set of shape features for discriminating granulomas and adenocarcinomas, (b) our automated segmentation approach (SEGvAC) yielded very good concordance against manual segmentations. However, future work will be necessary to ensure that the automatic segmentation provides a nodule boundary that is more effective for classification and (c) the performance of the shape-based classifier on an independent validation for both automated and manual segmentation clearly seems to suggest that shape is an important attribute to consider for discriminating granulomas and adenocarcinomas."(*p13*)

Main Study Findings	Authors' Conclusion
van Riel, 2017[23]	

AUC
• For risk-assessment of nodules of all sizes, a non-significant difference between computer model (0.932) and all human observers (n = 11; mean: 0.910, range: 0.860 ± 0.950, *P* = 0.184)
  ○ Board certified radiologists (n = 4) AUC: 0.919, which would not change the result in favour of the computer model if statistically tested
• For differentiating malignant nodules from size-matched benign nodules, all human observers (n = 11; mean: 0.819, range 0.771 ± 0.881) performed significantly better than the computer model (0.706, *P* < 0.001)
  ○ Board certified radiologists (n = 4) AUC: 0.844, which would not change the result in favour of the computer model, if statistically tested
	"Computer model and human observers perform equivalent for differentiating malignant from randomly selected benign nodules, confirming the high potential of computer models for nodule risk estimation in population-based screening studies. However, computer models highly rely on size as discriminator. Incorporation of other morphological criteria used by human observers to superiorly discriminate size-matched malignant from benign nodules, will further improve computer performance."(*p2*)
	
	"In conclusion, the PanCan risk prediction model and human observers perform equally well for differentiating malignant from randomly selected benign screen-detected pulmonary nodules, underlining the large potential of computer-based risk estimation to trigger nodule management in population-based screening studies. Human observers, however, significantly outperform the PanCan model for differentiating malignant from size-matched screen detected benign nodules suggesting that integration of additional morphological characteristics, such as pleural retraction and perinodular lung parenchyma distortion, used by the human observers is very likely to lead to further improvement of computer-based risk prediction models" (*p12*)

Armato, 2016[22]	
AUC • Automated (range from 11 different methods): 0.50 to 0.68 (SE = 0.06 for the latter) ○ 3 of the methods performed better than random guessing, with p-values of 0.006, 0.008, and 0.048; these p-values do not remain statistically significant after the Holm–Bonferroni correction. • Manual: mean AUC value across 6 radiologists: 0.79 (SD = 0.06); range: 0.70 to 0.85	"The LUNGx Challenge was a successful scientific challenge for the computerized classification of lung nodules on CT scans ... Ten participating groups from academia and industry applied 11 computerized methods to the 73 lung nodules in the test set of scans; these methods ranged from fully automated to semiautomated with varying levels of radiologist input. Only three of these methods performed better than random guessing within the statistical limits of the Challenge. To place the performance of the computerized methods into a real-world context, an observer study was conducted with six attending radiologists manually performing the same Challenge task. Three of the radiologists performed better than the best-performing computer method. Challenges should be approached by both organizers and participants as a friendly competition within the research community, designed to foster interest in the designated task and encourage innovation in the field. The continued public availability of the Challenge cases will provide a valuable resource for the medical imaging research community into the future."(*p044506-7-8*)

CI = confidence interval; ACR lung-RADS = American College of Radiologists Lung CT Screening Reporting and Data System; AUC = area under the receiver operating characteristic curve; CADx = computer-aided diagnosis; CNN = convolutional neural network; CT = computer tomography; GGO = ground-glass opacity; LIDC/IDRI = lung image database consortium and image database resource initiative; LDCT = low-dose computed tomography; PanCan = Pan-Canadian Early Detection of Lung Cancer Study; SD = standard deviation; SE = standard error; SEGvAC = an automated segmented approach; SSPN = small solid pulmonary nodules; SVM-LASSO = support vector machine with a least absolute shrinkage and selection operator.

Appendix 5: Additional References of Potential Interest

Unclear or Inappropriate Comparator and/or Reference Standard for Diagnostic Test Accuracy Studies

Wu W, Hu H, Gong J, Li X, Huang G, Nie S. Malignant-benign classification of pulmonary nodules based on random forest aided by clustering analysis. *Phys Med Biol*. 2019;64(3):035017.

Causey JL, Zhang J, Ma S, et al. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci Rep*. 2018;8(1):9286.

Abbas Q. Nodular-Deep: Classification of pulmonary nodules using deep neural network. *Int J Med Res Health Sci*. 2017;6(8):111-118.

Protocol of a Potentially Relevant Systematic Review

Moon SJ, Kim JY, Ham T, Moon E, Hwang JS. A systematic review and meta-analysis on the accuracy of deep-learning algorithm in differentiating benign and malignant lung nodules on a computed tomography (CT) scan. (CRD42019122206). *PROSPERO: International Prospective Registrar of Systematic Reviews*. York (GB): University of York Centre for Reviews and Dissemination; 2019: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019122206 Accessed 2020 Jan 21