5-24-2022

# Artificial Intelligence Tool for Assessment of Indeterminate Pulmonary Nodules Detected with CT

Roger Y. Kim

Jason L. Oke

Lyndsey C. Pickup

Reginald F. Munden

Travis L. Dotson

*See next page for additional authors*

## Authors

Roger Y. Kim, Jason L. Oke, Lyndsey C. Pickup, Reginald F. Munden, Travis L. Dotson, Christina R. Bellinger, Avi Cohen, Michael J. Simoff, Pierre P. Massion, Claire Filippini, Fergus V. Gleeson, and Anil Vachani

Radiology

ORIGINAL RESEARCH • THORACIC IMAGING

# Artificial Intelligence Tool for Assessment of Indeterminate Pulmonary Nodules Detected with CT

*Roger Y. Kim, MD, MSCE • Jason L. Oke, MSc, DPhil • Lyndsey C. Pickup, DPhil • Reginald F. Munden, MD, DMD, MBA • Travis L. Dotson, MD • Christina R. Bellinger, MD • Avi Cohen, MD • Michael J. Simoff, MD • Pierre P. Massion, MD† • Claire Filippini, MBChB • Fergus V. Gleeson, PhD, FRCP, FRCR • Anil Vachani, MD, MSCE*

From the Division of Pulmonary, Allergy, and Critical Care, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Suite 216, Stemmler Hall, 3450 Hamilton Walk, Philadelphia, PA 19104 (R.Y.K., A.V.); Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom (J.L.O.); Optellum, Oxford, United Kingdom (L.C.P.); Department of Radiology and Radiological Science, Medical University of South Carolina, Charleston, SC (R.F.M.); Department of Pulmonary, Critical Care, Allergy and Immunologic Diseases, Wake Forest School of Medicine, Winston-Salem, NC (T.L.D., C.R.B.); Division of Pulmonary and Critical Care Medicine, Department of Medicine, Henry Ford Health System, Detroit, Mich (A.C., M.J.S.); Division of Allergy, Pulmonary and Critical Care Medicine, Vanderbilt Ingram Cancer Center, Nashville, Tenn (P.P.M.); and Department of Oncology, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom (C.F., F.V.G.). Received August 25, 2021; revision requested October 14; revision received March 2, 2022; accepted March 16. **Address correspondence to** A.V. (e-mail: *avachani@pennmedicine.upenn.edu*).

Supported by Optellum. R.Y.K. supported by a National Heart, Lung, and Blood Institute T32 Training Grant (HL-007891) and Siemens. J.L.O. supported in part by the NIHR Oxford Biomedical Research Centre and Oxford University Hospitals NHS Foundation Trust. A.V. supported in part by the National Institute of Environmental Health Sciences (P30-ES013508)

Conflicts of interest are listed at the end of this article.

See also the editorial by Yanagawa in this issue.

**Radiology 2022; 000:1–9 • https://doi.org/10.1148/radiol.212182 • Content codes: CH AI**

***Background:*** Limited data are available regarding whether computer-aided diagnosis (CAD) improves assessment of malignancy risk in indeterminate pulmonary nodules (IPNs).

***Purpose:*** To evaluate the effect of an artificial intelligence–based CAD tool on clinician IPN diagnostic performance and agreement for both malignancy risk categories and management recommendations.

***Materials and Methods:*** This was a retrospective multireader multicase study performed in June and July 2020 on chest CT studies of IPNs. Readers used only CT imaging data and provided an estimate of malignancy risk and a management recommendation for each case without and with CAD. The effect of CAD on average reader diagnostic performance was assessed using the Obuchowski-Rockette and Dorfman-Berbaum-Metz method to calculate estimates of area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. Multirater Fleiss κ statistics were used to measure interobserver agreement for malignancy risk and management recommendations.

***Results:*** A total of 300 chest CT scans of IPNs with maximal diameters of 5–30 mm (50.0% malignant) were reviewed by 12 readers (six radiologists, six pulmonologists) (patient median age, 65 years; IQR, 59–71 years; 164 [55%] men). Readers' average AUC improved from 0.82 to 0.89 with CAD ($P < .001$). At malignancy risk thresholds of 5% and 65%, use of CAD improved average sensitivity from 94.1% to 97.9% ($P = .01$) and from 52.6% to 63.1% ($P < .001$), respectively. Average reader specificity improved from 37.4% to 42.3% ($P = .03$) and from 87.3% to 89.9% ($P = .05$), respectively. Reader interobserver agreement improved with CAD for both the less than 5% (Fleiss κ, 0.50 vs 0.71; $P < .001$) and more than 65% (Fleiss κ, 0.54 vs 0.71; $P < .001$) malignancy risk categories. Overall reader interobserver agreement for management recommendation categories (no action, CT surveillance, diagnostic procedure) also improved with CAD (Fleiss κ, 0.44 vs 0.52; $P = .001$).

***Conclusion:*** Use of computer-aided diagnosis improved estimation of indeterminate pulmonary nodule malignancy risk on chest CT scans and improved interobserver agreement for both risk stratification and management recommendations.



*Online supplemental material is available for this article.*

Indeterminate pulmonary nodules (IPNs), rounded opacities 3 cm or less in diameter surrounded by aerated pulmonary parenchyma without clearly benign features, pose a diagnostic challenge for clinicians (1,2). IPNs are commonly identified on chest CT scans incidentally as part of routine clinical care (3), and a growing proportion of IPNs are expected to be detected with lung cancer screening (4–6).

As most IPNs are benign (2–5,7), it is critical for clinicians to accurately assess malignancy risk to both diagnose and treat malignant lesions in a timely fashion while avoiding unnecessary tests and procedures in patients with benign nodules (8). Guidelines from the American College of Chest Physicians recommend CT surveillance for lesions with very low risk (<5%) and functional imaging (ie, PET/CT) or nonsurgical biopsy for those with low or moderate risk (range, 5%–65%). Surgical biopsy can be considered in appropriately selected patients with high-risk (>65%) lesions (1). However, prior studies have demonstrated variable agreement among radiologists when risk stratifying IPNs (9,10) and inconsistent adherence to practice guidelines among pulmonologists (11). Moreover, although several clinical risk prediction models have been developed to estimate IPN malignancy risk (12–14), they have not consistently outperformed physician assessment of IPN risk (15–17).

### Abbreviations

AUC = area under the receiver operating characteristic curve, CAD = computer-aided diagnosis, IPN = indeterminate pulmonary nodule, LCP = lung cancer prediction, LCP-CNN = Lung Cancer Prediction Convolutional Neural Network


### Summary

An artificial intelligence–based computer-aided diagnosis tool improved radiologists' and pulmonologists' risk stratification of indeterminate pulmonary nodules on chest CT scans.

### Key Results

- In this retrospective multireader multicase study with 300 chest CT scans of pulmonary nodules and 12 readers (six pulmonologists, six radiologists), computer-aided diagnosis (CAD) improved readers' estimation of nodule malignancy risk (average area under the receiver operating characteristic curve increased from 0.82 to 0.89, $P < .001$), regardless of reader specialty.
- The average sensitivity and specificity of pulmonologists and radiologists improved with CAD at both the very low (5%) and high (65%) malignancy risk thresholds, suggesting that CAD may have a meaningful impact on pulmonary nodule management decisions.


Recently, radiomics tools using raw CT data have been developed to help clinicians classify IPNs as malignant or benign (18–20). The Lung Cancer Prediction Convolutional Neural Network (LCP-CNN) is an artificial intelligence–based computer-aided diagnosis (CAD) model that was derived and internally validated using data from the National Lung Screening Trial and externally validated in two cohorts of patients with incidentally detected IPNs (21,22). In these cohorts, malignancy risk estimates provided by the LCP-CNN model had greater discrimination than both the Brock and Mayo risk prediction models (12,13,21,22). However, for CAD tools such as the LCP-CNN to have practical clinical utility, they must provide additional diagnostic benefit to clinicians interpreting chest CT imaging results with IPNs.

Our main objective was to evaluate the effect of the LCP-CNN CAD tool on the performance of radiologists and pulmonologists in risk stratification of IPNs on chest CT scans. A prior pilot study suggested a significant improvement in clinician IPN discrimination with CAD (23). Here we report the results of a larger multicenter validation study assessing the impact of LCP-CNN CAD on clinicians' risk stratification and management of IPNs detected either incidentally or with lung cancer screening.

## Materials and Methods

In this retrospective multireader multicase study, readers evaluated CT scans with and without the LCP-CNN CAD tool. Deidentified imaging studies were collected from seven sources with local institutional review board approval: two institutions in the United States (Henry Ford Health System, Vanderbilt University), four institutions in the United Kingdom (Royal Berkshire Hospital, Leeds Teaching Hospital, Nottingham University Hospital, Oxford University Hospitals), and National Lung Screening Trial data obtained through the National Cancer Institute Cancer Data Access System. The use of deidentified imaging studies complied with Health Insurance Portability and Accountability Act guidelines, and the need for informed consent was waived. Evaluations by readers were conducted in June and July 2020, and data analysis was performed from August 2020 to December 2021.

Authors who are not employees of or consultants for Optellum (R.Y.K., J.L.O., A.V.) reviewed and analyzed the data. One investigator (L.C.P.), who is an employee of Optellum, helped curate the data and participated in study planning. One author (F.V.G.) is a shareholder in Optellum and reviewed CT scan data included in this study. The remaining authors have no financial relationships with Optellum.

### Calculation of the Lung Cancer Prediction Score

An artificial intelligence tool was used that evaluates raw CT imaging data of an IPN of interest and calculates an estimated probability of cancer (from 0 to 100). Methods for the development and validation of the LCP-CNN CAD software (Virtual Nodule Clinic, version 2.0.0; Optellum) have been previously described (21,22), and the product is commercially available. These risk estimates are then converted to a lung cancer prediction (LCP) score that categorizes the malignancy risk on a decile scale, with a score of 1 representing nodules at lowest risk and a score of 10 indicating nodules at highest risk. In a target population with a malignant nodule prevalence of 30%, approximately 10% of all nodules (malignant and benign) will be categorized within each decile of risk. The probabilities of a malignant nodule within each decile are summarized in Figure 1. For this study, the CAD software provided the LCP score and displayed Figure 1 to readers to allow the post-CAD risk determination.

### Case Selection

The cases included in this study comprised imaging data from both screening and diagnostic chest CT scans (Fig 2). No imaging studies in the training data sets used for development of the LCP-CNN CAD were included in this study. We selected the largest 5- to 30-mm IPN per chest CT scan, and IPNs were defined as malignant or benign based on histologic evaluation. For cases without a definitive histologic diagnosis, IPNs were determined to be benign if follow-up imaging demonstrated complete resolution of the nodule or 2-year stability by nodule diameter.

We used a random weighted sampling approach to select 300 cases from the 5023 cases meeting inclusion criteria for this study. The final data set was enriched to include a 50% prevalence of malignancy, equitable distribution of case difficulty, and source of data (United States vs United Kingdom). Representative axial images and corresponding LCP scores for a malignant nodule and a benign nodule are shown in Figure 3. A detailed description of the selection algorithm is included in Appendix E1 (online).

### Readers and Reading Procedures

The 12 readers had current medical licenses and specialty certifications in radiology or pulmonary medicine. Each reader was trained for 1 hour on the LCP-CNN CAD software and (21,22) and evaluated 17 example cases before assessing images included in the analysis. Readers were blinded to all patient clinical and demographic information and were informed that the set of 300

cases contained a higher prevalence of malignancy than in routine clinical practice to provide modest guidance for malignancy risk assessments.

Readers began the interpretation of each of the randomly ordered 300 cases by loading the scan into the software, which highlighted the IPN of interest, and scrolling through the entire set of images with axial views. Readers estimated malignancy risk on a 100-point scale and separately selected a management recommendation from the following six options: no action, long-term (≥6 months) CT follow-up, short-term (6 weeks to 6 months) CT follow-up, immediate imaging follow-up (eg, PET/CT), nonsurgical biopsy (eg, needle biopsy), or surgical resection. Immediately after this initial assessment, the LCP score was displayed, and readers were asked to provide an updated risk estimate and management recommendation (Fig E1 [online]). Readers were not able to change their initial malignancy risk estimate or management recommendation after seeing the LCP score. Readers evaluated all cases independently, and no reading time limit was imposed.

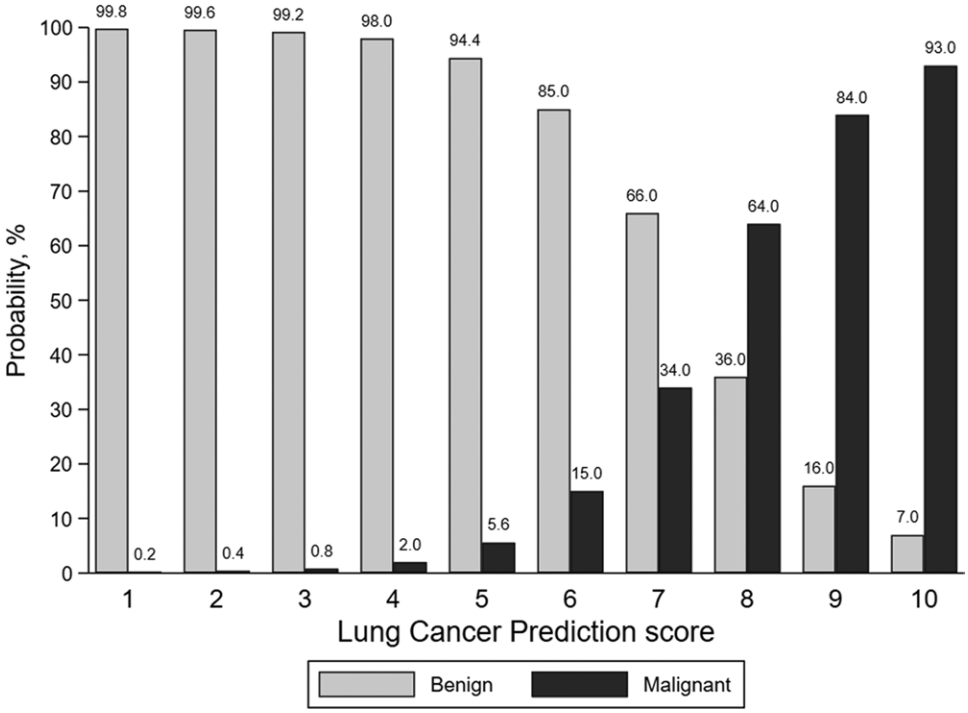


**Figure 1:** The lung cancer prediction score is generated by an artificial intelligence tool and categorizes pulmonary nodule malignancy risk on a decile scale, with a score of 1 representing nodules at lowest risk and a score of 10 indicating nodules at highest risk.

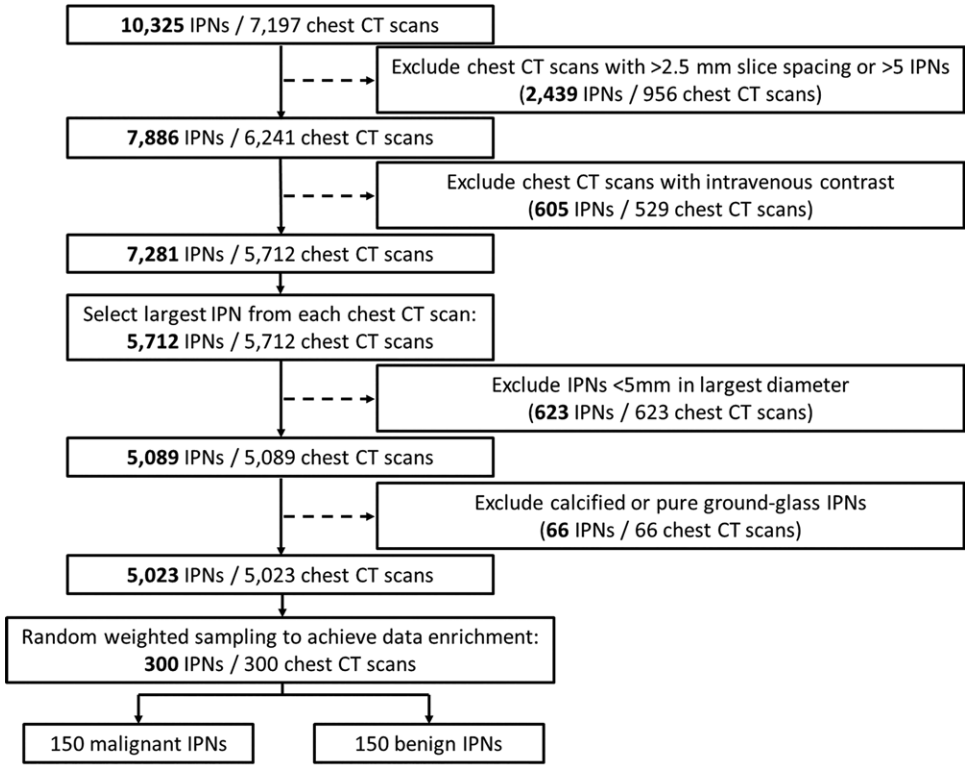


**Figure 2:** Flowchart shows inclusion and exclusion criteria for pulmonary nodules included in the study. IPNs = indeterminate pulmonary nodules.

## Statistical Analyses

Power calculations were performed using estimates of the relevant variances from the prior pilot study (23) and implemented in the R software package RJafroc for the Dorfman-Berbaum-Metz method. With 12 readers and 300 cases, the study had 92% power to identify a minimum difference in area under the receiver operating characteristic curve (AUC) of 0.04 with use of CAD.

The primary outcome was the change in readers' average AUC between case malignancy risk estimates with and without CAD. The Obuchowski-Rockette and Dorfman-Berbaum-Metz method for analyzing multireader multicase studies and the MRMCaov library (24) was used for all analyses of diagnostic performance in this study to test the null hypothesis that the readers' average AUC without CAD was equal to that with CAD (25). The Obuchowski-Rockette and

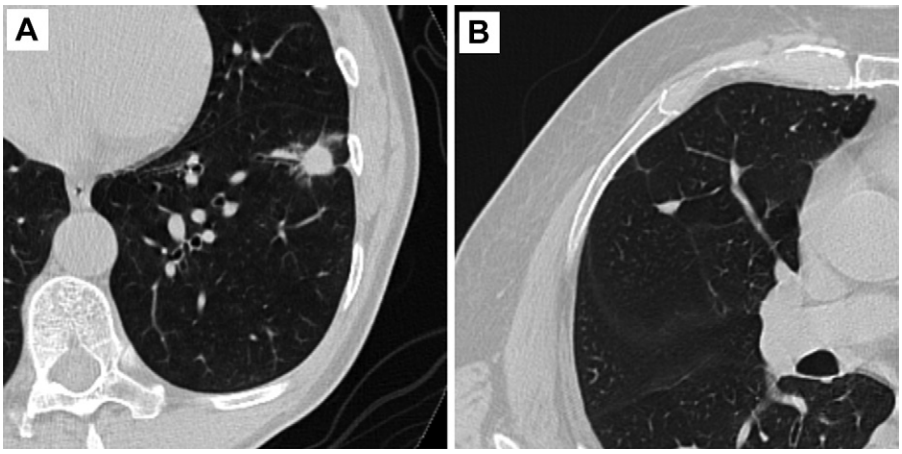


**Figure 3:** Representative axial CT images of pulmonary nodules included in the study. **(A)** Malignant nodule with a lung cancer prediction score of 10. **(B)** Benign nodule with a lung cancer prediction score of 2.

**Table 1: Characteristics of Patients, Pulmonary Nodules, and Chest CT by Diagnosis**

| Variable | Total ($n$ = 300) | Benign IPNs ($n$ = 150) | Malignant IPNs* ($n$ = 150) | $P$ Value |
|---|---|---|---|---|
| Age (y)† | 65 (59–71) | 62 (57–68) | 68 (62–73) | <.001 |
| Sex | | | | .003 |
| Female | 136 (45) | 55 (37) | 81 (54) | … |
| Male | 164 (55) | 95 (63) | 69 (46) | … |
| Nodule diameter | | | | <.001 |
| 5 to <10 mm | 120 (40) | 84 (56) | 36 (24) | … |
| 10–30 mm | 180 (60) | 66 (44) | 114 (76) | … |
| Nodule density | | | | .003 |
| Solid | 245 (82) | 122 (81) | 123 (82) | … |
| Mixed | 31 (10) | 22 (15) | 9 (6) | … |
| Part solid | 24 (8) | 6 (4) | 18 (12) | … |
| Nodule margins | | | | <.001 |
| Nonspiculated | 192 (64) | 122 (81) | 70 (47) | … |
| Spiculated | 108 (36) | 28 (19) | 80 (53) | … |
| Type of chest CT | | | | <.001 |
| Diagnostic | 176 (59) | 63 (42) | 113 (75) | … |
| Screening | 124 (41) | 87 (58) | 37 (25) | … |
| Reconstructed section thickness | | | | .15 |
| 0.5 to <1.5 mm | 194 (65) | 91 (61) | 103 (69) | … |
| 1.5–2.5 mm | 106 (35) | 59 (39) | 47 (31) | … |

Note.—Unless otherwise indicated, data are number of patients, and data in parentheses are percentages. Mixed nodules were defined as those with cystic airspaces or pseudocavitation or that were predominantly solid with a thin rim of ground-glass. Part-solid nodules had ground-glass and solid components. IPN = indeterminate pulmonary nodule.

*Malignant histologic diagnoses included adenocarcinoma (98 of 150 [65.3%]), squamous cell carcinoma (14 of 15 [93.3%]), other non–small cell lung cancer (27 of 150 [18.0%]), small cell carcinoma (three of 150 [2.0%]), and other neoplasms (eight of 150 [5.3%]).

† Data are median and data in parentheses are the interquartile range.

Dorfman-Berbaum-Metz method accounts for the fact that in a multireader multicase study, the same cases are evaluated by each reader. As such, error terms are assumed to be equi-covariant between readers and cases and are not assumed to be independent. We calculated 95% CIs using nonparametric bootstrap resampling ($n$ = 10 000) with the percentile method. We performed secondary analyses to compare the differences in average AUC stratified by reader specialty, nodule size, density, margins (spiculated vs nonspiculated), and type of CT imaging (screening vs diagnostic). We additionally calculated the average sensitivity and specificity across readers for IPN classification with and without CAD at the 5% and 65% malignancy risk thresholds. Multirater Fleiss κ statistics were used to measure interobserver agreement for IPN malignancy risk categories and management recommendations with and without CAD. The κ values were interpreted using Landis and Koch guidelines (26).

Statistical significance was defined with a two-sided $P < .05$, and no adjustments were made for multiplicity. Analyses were performed using R software, version 4.04 (R Foundation) and Stata/MP, version 17.0 (StataCorp), and the code is available at *https://github.com/jokeyjo/Pulmonary-Nodules*.

## Results

### Case and Reader Characteristics

Of the 300 chest CT scans each with a single 5–30-mm IPN of interest, an equal number ($n$ = 150 [50.0%]) were benign and malignant lesions (patient median age, 65 years; IQR, 59–71 years; 164 [55%] men, 136 [45%] women). Patient, pulmonary nodule, and imaging study characteristics are presented by IPN diagnosis in Table 1. Most IPNs (60.0% [180 of 300]) were 10 mm or greater in largest axial diameter, and 81.7% (245 of 300) and 64.0% (192 of 300) were solid and nonspiculated, respectively. Adenocarcinomas comprised 65.3% (98 of 150) of the malignant IPNs. Compared with benign nodules, malignant nodules were associated with older age (median age, 68 years; IQR, 62–73 years] vs 62 years; IQR, 57–68 years; $P < .001$), female sex (54% [$n$ = 81] vs 37% [$n$ = 55], $P$ = .003), part-solid density (12.0% [$n$ = 18] vs 4.0% [$n$ = 6], $P$ = .003), increased nodule diameter (median, 12.0 mm; IQR, 10.0–18.0 mm vs 8.5 mm; IQR, 6.0–13.0 mm; $P < .001$), and spiculation (53% [$n$ = 80] vs 19% [$n$ = 28], $P < .001$). Figure E2 (online) displays the distribution of LCP scores by diagnosis for the 300 cases. The 12 readers included six pulmonologists (two with expertise in tho-

racic oncology) and six radiologists (two with expertise in thoracic radiology) (Table E1 [online]). Tables E2 and E3 (online) summarize reader estimates of malignancy risk and management recommendations without and with CAD, respectively.

## Diagnostic Performance

The average AUC across all readers for estimating malignancy risk without CAD was 0.82 (95% CI: 0.77, 0.86) compared with 0.89 (95% CI: 0.85, 0.92) with CAD ($P < .001$, Fig 4). An improvement in AUC with CAD was observed for each of the 12 readers ($P \leq .001$, Fig 5) and across prespecified strata by reader specialty (pulmonology, $P = .001$; radiology, $P = .001$), nodule diameter (5 to <10 mm, $P < .001$; 10–30 mm, $P < .001$), nodule density (solid or mixed, $P < .001$; part solid, $P = .05$), nodule margins (nonspiculated, $P < .001$; spiculated, $P = .007$), and type of chest CT (diagnostic, $P < .001$; screening, $P < .001$; Table 2). We then evaluated the effect of CAD on classification of IPNs by comparing the average sensitivity and specificity of readers' risk estimates with and without CAD at thresholds of 5% and 65% (Table 3, Table E4 [online]). At the 5% threshold, the average sensitivity across readers was greater with CAD (1693 of 1800 [94.1%] vs 1762 of 1800 [97.9%], $P = .01$), as was the average specificity (674 of 1800 [37.4%] vs 761 of 1800 [42.3%], $P = .03$). At the 65% threshold, the average sensitivity improved with CAD (946 of 1800 [52.6%] vs 1136 of 1800 [63.1%], $P < .001$), as did the specificity (1572 of 1800 [87.3%] vs 1619 of 1800 [89.9%], $P = .05$). Figure 6 and Table E5 (online) summarize the effect of CAD on reclassification of all IPN cases.

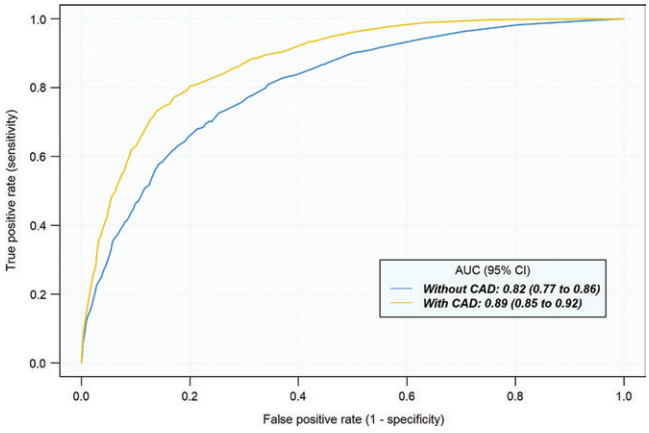


**Figure 4:** Average reader receiver operating characteristic curves for discrimination of indeterminate pulmonary nodules under two reading conditions: without computer-aided diagnosis (CAD) and with CAD. Average area under the receiver operating characteristic curve (AUC) was computed across 12 readers participating in the study using either the Obuchowski-Rockette and Dorfman-Berbaum-Metz method, which accounts for the multireader multicase study design.

## Interobserver Agreement

Overall agreement among readers for malignancy risk categories improved with CAD ($\kappa = 0.35$ vs $\kappa = 0.58$, $P < .001$; Table 4). For very low malignancy risk (<5%), reader agreement improved from moderate ($\kappa = 0.50$) to substantial ($\kappa = 0.71$) with CAD ($P < .001$). Agreement among readers for high malignancy risk (>65%) similarly improved from moderate ($\kappa = 0.54$) to substantial ($\kappa = 0.71$) with CAD ($P < .001$). Improvements in agreement were also observed in the 5%–30% ($\kappa = 0.21$ vs $\kappa = 0.45$, $P < .001$) and 31%–65% ($\kappa = 0.11$ vs $\kappa = 0.36$, $P < .001$) risk categories.

With CAD, there was improvement in agreement of management recommendations ($\kappa = 0.44$ vs $\kappa = 0.52$, $P = .001$). Reader agreement ($\kappa$) for management with diagnostic procedures improved from 0.60 to 0.68 with CAD ($P = .008$) and improved from fair ($\kappa = 0.36$) to moderate ($\kappa = 0.43$) for CT surveillance ($P = .02$).

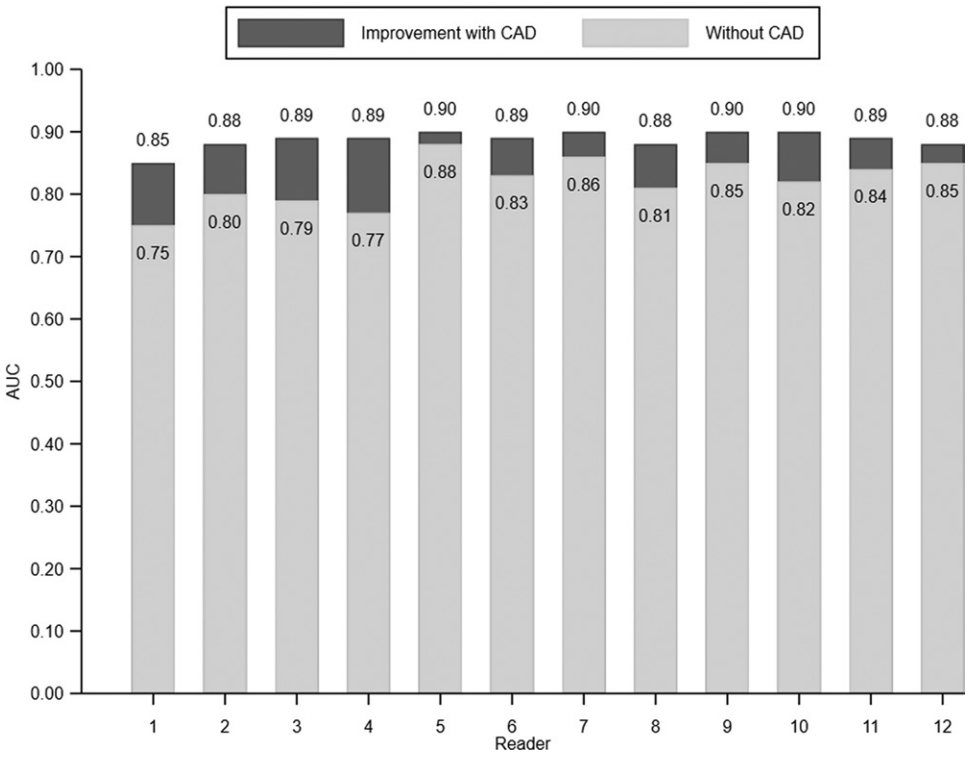


**Figure 5:** Individual reader discrimination under two reading conditions: without computer-aided diagnosis (CAD) and with CAD. There was a significant improvement in area under the receiver operating characteristic curve (AUC) for each reader ($P \leq .001$) with CAD.

## Discussion

Limited data exist on whether computer-aided diagnosis (CAD) improves clinicians' assessment of malignancy risk of indeterminate pulmonary

nodules (IPNs). In this multireader multicase study, the performance of radiologists and pulmonologists in estimation of IPN malignancy risk significantly improved with the assistance of an artificial intelligence–based CAD tool. The average area under the receiver operating characteristic curve improved from 0.82 to 0.89 ($P < .001$) with CAD, with an improvement observed for each of the 12 readers, regardless of clinician specialty. At the 5% and 65% malignancy risk thresholds, use of CAD improved reader sensitivity from 94.1% to 97.9% ($P = .01$) and from 52.6% to 63.1% ($P < .001$), respectively. Specificity improved from 37.4% to 42.3% ($P = .03$) and from 87.3% to 89.9% ($P = .05$) at these thresholds, respectively. Moreover, use of CAD improved the agreement among readers for risk assessment ($\kappa = 0.35$ vs $\kappa = 0.58$, $P < .001$) and management recommendations ($\kappa = 0.44$ vs $\kappa = 0.52$, $P = .001$). These results suggest that CAD may help clinicians more accurately risk stratify pulmonary nodules when interpreting chest CT imaging data.

Existing evidence on the management of IPNs suggests considerable misalignment between malignancy risk and subsequent management decisions, including a high rate of benign diagnoses identified among patients undergoing invasive diagnostic procedures (15,27–29). This misalignment exists at least in part because the two existing approaches to IPN risk estimation—clinician-estimated risk and clinical risk prediction models—provide acceptable but far from optimal discrimination in patients with IPNs (14–17,30). Especially in intermediate-risk IPNs, this lack of precision may contribute to instances in which malignant nodules are managed with CT surveillance and benign nodules are managed with biopsy, resulting in delayed lung cancer diagnoses and unnecessary procedural risks (27,31). The promise of radiomics-based CAD tools lies in the additional data invisible to the human eye (eg, shape, spatial complexity, textures, wavelet transformations) provided to clinicians beyond IPN size, spiculation, and density—with the goal of optimizing these challenging diagnostic management decisions (32,33). Our findings confirm that CAD can enhance clinician interpretation of risk based on imaging data alone. Moreover, as the average sensitivity and specificity of pulmonologists and radiologists improved with CAD at the very low (5%) and high (65%) malignancy risk thresholds, CAD may have a meaningful impact on pulmonary nodule management decisions.

**Table 2: Average AUC of Readers for Discrimination of Indeterminate Pulmonary Nodules with and without CAD by Prespecified Subgroups**

| | Average AUC | | |
|---|---|---|---|
| Variable | Without CAD | With CAD | *P* Value |
| Reader specialty | | | |
| Pulmonology | 0.82 (0.76, 0.87) | 0.88 (0.85, 0.92) | .001 |
| Radiology | 0.82 (0.77, 0.87) | 0.89 (0.86, 0.93) | .001 |
| Nodule diameter | | | |
| 5 to <10 mm | 0.80 (0.74, 0.87) | 0.91 (0.86, 0.96) | <.001 |
| ≥10–30 mm | 0.77 (0.70, 0.83) | 0.86 (0.80, 0.91) | <.001 |
| Nodule density | | | |
| Solid or mixed | 0.82 (0.78, 0.87) | 0.88 (0.85, 0.92) | <.001 |
| Part solid | 0.80 (0.63, 0.96) | 0.94 (0.86, 1.00) | .05 |
| Nodule margins | | | |
| Nonspiculated | 0.80 (0.74, 0.85) | 0.88 (0.84, 0.93) | <.001 |
| Spiculated | 0.74 (0.64, 0.83) | 0.81 (0.71, 0.90) | .007 |
| Type of chest CT | | | |
| Diagnostic | 0.77 (0.70, 0.83) | 0.84 (0.78, 0.90) | <.001 |
| Screening | 0.85 (0.79, 0.92) | 0.92 (0.87, 0.96) | <.001 |

Note.—Data in parentheses are 95% CIs. AUC = area under the receiver operating characteristic curve, CAD = computer-aided diagnosis.

**Table 3: Average Diagnostic Performance of Readers in Classification of Pulmonary Nodules with CAD**

| | Malignancy Risk Threshold | | | | | |
|---|---|---|---|---|---|---|
| | 5% | | | 65% | | |
| Classification Performance | Without CAD | With CAD | *P* Value | Without CAD | With CAD | *P* Value |
| True positive | 1693 | 1762 | … | 946 | 1136 | … |
| False positive | 1126 | 1039 | … | 228 | 181 | … |
| True negative | 674 | 761 | … | 1572 | 1619 | … |
| False negative | 107 | 38 | … | 854 | 664 | … |
| Sensitivity (%) | 94.1 (90.8, 97.4) | 97.9 (96.0, 99.7) | .01 | 52.6 (41.8, 63.3) | 63.1 (53.7, 72.5) | <.001 |
| Specificity (%) | 37.4 (27.2, 47.6) | 42.3 (31.3, 53.3) | .03 | 87.3 (81.0, 93.6) | 89.9 (83.3, 96.6) | .05 |

Note.—Unless otherwise indicated, data are number of findings. Data in parentheses are 95% CI. Calculations were made using the Obuchowski-Rockette and Dorfman-Berbaum-Metz method, which accounts for the multireader multicase study design. CAD = computed-aided diagnosis.

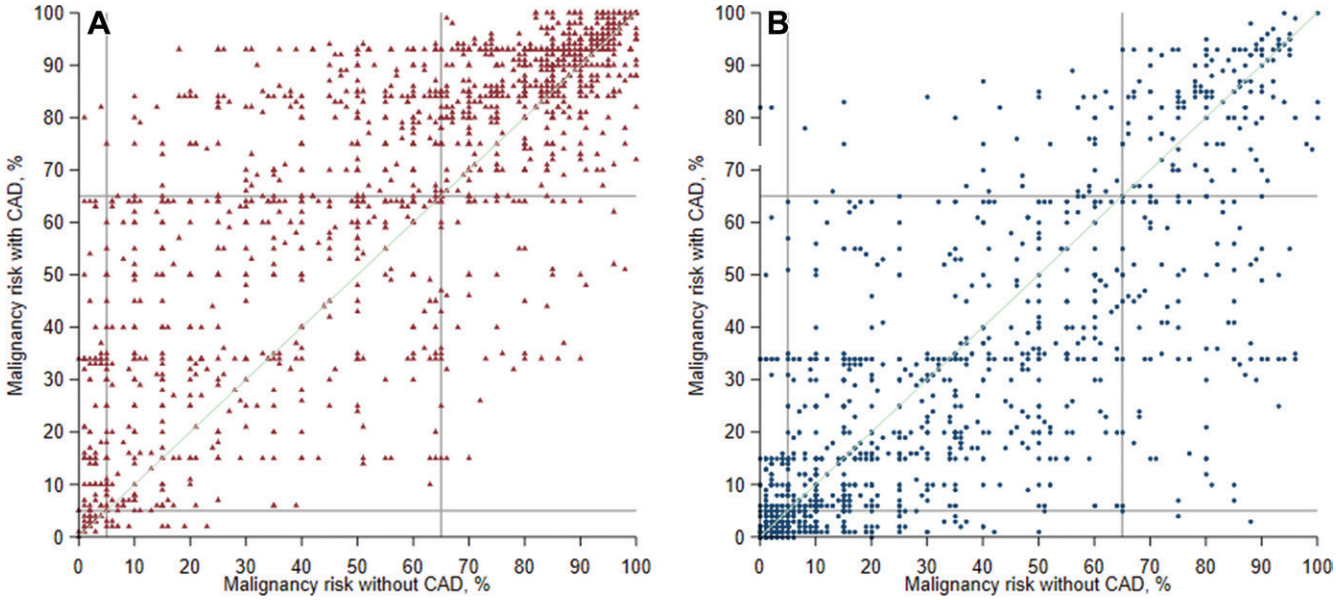


**Figure 6:** Reclassification plots with and without computer-aided diagnosis (CAD) for malignant and benign pulmonary nodules. Summary plots of all pairs of pre-CAD (x-axis) and post-CAD (y-axis) malignancy risk estimates for malignant ($n$ = 1800 [150 cases × 12 readers]) **(A)** and benign ($n$ = 1800 [150 cases × 12 readers]) **(B)** nodules. Malignancy risk decision thresholds of 5% and 65% are depicted as gray lines in each plot.

Prior studies have demonstrated considerable variability among clinicians when classifying IPNs, which might lead to differential management recommendations (9,10). Thus, we sought to determine if reader agreement changed with CAD. Agreement for both very low–risk and high-risk IPNs improved from moderate to substantial, suggesting that CAD may promote a more uniform approach to IPN assessment. Although reader agreement also increased in the intermediate-risk categories, it is unclear how this might affect downstream clinical care, as there is considerably more variation in management strategies for nodules with a malignancy risk of 5%–65%. That the overall reader agreement for management recommendations was only moderate in this study is consistent with prior studies demonstrating variable adherence to management guidelines (10,11,15,27). Moreover, agreement may have been further limited in this study because, by design, readers did not have access to patient demographic or clinical information when choosing management recommendations.

As our study included both screen-detected and incidentally detected IPNs, the results are generalizable to a broad range of IPNs. Although prior similar studies have included only experienced thoracic radiologists (19,20), the readers in our study included both radiologists and pulmonologists with a range of experience. Improvements in diagnostic performance with CAD did not differ by reader specialty, suggesting that CAD might benefit clinicians in a variety of clinical settings.

Our study had limitations. First, readers were not provided any clinical information when assessing IPN imaging data; thus, generalizability to a routine clinical setting is limited. Our intention was to avoid introducing bias and to exclude the uncertainty of whether variability in image interpretation was because of clinical context rather than nodule characteristics.

**Table 4: Interobserver Agreement for Malignancy Risk and Management Recommendation with and without CAD**

| Variable | Interobserver Agreement | | |
|---|---|---|---|
| | Without CAD | With CAD | *P* Value |
| Malignancy risk | | | |
| Very low (<5%) | 0.50 (0.45, 0.55) | 0.71 (0.67, 0.75) | <.001 |
| Low-moderate (5%–30%) | 0.21 (0.18, 0.25) | 0.45 (0.39, 0.50) | <.001 |
| Moderate-high (31%–65%) | 0.11 (0.09, 0.13) | 0.36 (0.32, 0.41) | <.001 |
| High (>65%) | 0.54 (0.49, 0.59) | 0.71 (0.67, 0.76) | <.001 |
| Overall | 0.35 (0.32, 0.38) | 0.58 (0.55, 0.61) | <.001 |
| Management recommendation | | | |
| No action | 0.22 (0.18, 0.27) | 0.32 (0.28, 0.36) | .002 |
| CT surveillance* | 0.36 (0.32, 0.40) | 0.43 (0.38, 0.47) | .02 |
| Diagnostic procedure† | 0.60 (0.55, 0.64) | 0.68 (0.64, 0.72) | .008 |
| Overall | 0.44 (0.41, 0.48) | 0.52 (0.49, 0.55) | .001 |

Note.—Unless otherwise indicated, data are Fleiss κ statistics, and data in parentheses are 95% CIs. κ reported for each separate category is calculated against all remaining categories combined. The overall κ is the weighted average of the individual κ statistics. CAD = computer-aided diagnosis.

* Short-term (6 weeks–6 months) or long-term (≥ 6 months) chest CT follow-up.

† Immediate imaging follow-up (eg, PET/CT scan), nonsurgical biopsy (eg, needle biopsy), or surgical resection.

The LCP-CNN CAD tool estimates malignancy risk based on imaging features without consideration of other clinical information (eg, age, smoking history), so our goal was to determine its impact on clinicians' ability to evaluate IPNs in

the absence of other risk factors. Second, before reviewing any cases, readers were told that the prevalence of malignancy was higher than is normally found in clinical practice, potentially introducing context bias and inflating all risk estimates (34). Third, the modest number of part-solid nodules included in this study limits the generalizability of our results to this subgroup of pulmonary nodules. Fourth, the LCP scores provided to readers were not accompanied by measures of uncertainty. Future studies should further evaluate the reliability of the LCP-CNN CAD tool. Fifth, although we observed significant improvements in diagnostic performance for each reader, the absolute increases in AUC across readers varied, and additional work is necessary to quantify what constitutes a clinically important improvement in discrimination. Sixth, as with all CAD-based studies, our results are applicable only to this software system, and other systems should not be assumed to produce similar results.

In conclusion, our study found that an artificial intelligence–based computer-aided diagnosis (CAD) tool improved the performance of radiologists and pulmonologists when estimating malignancy risk for indeterminate pulmonary nodules (IPNs) on chest CT scans and improved agreement for very low- and high-risk IPN categories. Our findings provide crucial support for bringing CAD tools closer to clinical implementation for IPN risk stratification. Furthermore, our results suggest that the Lung Cancer Prediction Convolutional Neural Network CAD tool may have a meaningful impact on subsequent management decisions. Future prospective studies will be necessary to evaluate the effect of CAD on clinical and patient-centered outcomes in real-world settings.

**Acknowledgments:** The authors thank the following colleagues for providing imaging data for the study: Matthew Callister, PhD; Matthew Clark, BSc; and Victoria Ashford-Turner, RGN (Leeds Teaching Hospital) and David Baldwin, MD, FRCP; Emily Reason; and Alison Clubley, PGCert (Nottingham University Hospital).

**Author contributions:** Guarantors of integrity of entire study, **R.Y.K., A.V.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **R.Y.K., R.F.M., M.J.S., A.V.**; clinical studies, **R.Y.K., R.F.M., A.C., M.J.S., A.V.**; experimental studies, **L.C.P.**; statistical analysis, **R.Y.K., J.L.O., A.V.**; and manuscript editing, **R.Y.K., J.L.O., L.C.P., R.F.M., T.L.D., C.R.B., A.C., M.J.S., F.V.G., A.V.**

**Disclosures of conflicts of interest:** **R.Y.K.** No relevant relationships. **J.L.O.** No relevant relationships. **L.C.P.** Co-founder and employee of and stockholder in Optellum; Optellum holds some patents in this area. **R.F.M.** RSNA R & E Foundation Board of Trustees Treasurer, ARRS Board of Chancellors and Chair of the membership committee, ACR Executive Council, ARRS representative; stock options in Optellum, TheraBionics stockholder. **T.L.D.** No relevant relationships. **C.R.B.** No relevant relationships. **A.C.** No relevant relationships. **M.J.S.** Consulting fees from Intuitive Surgical and Gongwin Biopharm; honoraria from PulmonX for lectures; travel support from Intuitive Surgical; stock options in SpinQ. **P.P.M.** No relevant relationships. **C.F.** No relevant relationships. **F.V.G.** President of the European Society of Thoracic Imaging, Chairman of RAIQC; holds shares in Optellum and RAIQC. **A.V.** Grants from MagArray, Broncus Medical, and PreCyte; consulting fees from Novocure and Johnson & Johnson; on the scientific advisory board of Lungevity Foundation and Delfi Diagnostics (unpaid).

## References

1. 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.
2. Mazzone PJ, Lam L. Evaluating the Patient With a Pulmonary Nodule: A Review. JAMA 2022;327(3):264–273.
3. Gould MK, Tang T, Liu IL, et al. Recent Trends in the Identification of Incidental Pulmonary Nodules. Am J Respir Crit Care Med 2015;192(10):1208–1214.
4. National Lung Screening Trial Research Team; Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.
5. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N Engl J Med 2020;382(6):503–513.
6. US Preventive Services Task Force; Krist AH, Davidson KW, et al. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. JAMA 2021;325(10):962–970.
7. Vachani A, Zheng C, Amy Liu IL, Huang BZ, Osuji TA, Gould MK. The Probability of Lung Cancer in Patients With Incidentally Detected Pulmonary Nodules: Clinical Characteristics and Accuracy of Prediction Models. Chest 2022;161(2):562–571.
8. Ost DE, Gould MK. Decision making in patients with pulmonary nodules. Am J Respir Crit Care Med 2012;185(4):363–372.
9. Penn A, Ma M, Chou BB, Tseng JR, Phan P. Inter-reader variability when applying the 2013 Fleischner guidelines for potential solitary subsolid lung nodules. Acta Radiol 2015;56(10):1180–1186.
10. van Riel SJ, Sánchez CI, Bankier AA, et al. Observer Variability for Classification of Pulmonary Nodules on Low-Dose CT Images and Its Effect on Nodule Management. Radiology 2015;277(3):863–871.
11. Tanner NT, Aggarwal J, Gould MK, et al. Management of Pulmonary Nodules by Community Pulmonologists: A Multicenter Observational Study. Chest 2015;148(6):1405–1414.
12. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. Arch Intern Med 1997;157(8):849–855.
13. 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.
14. Choi HK, Ghobrial M, Mazzone PJ. Models to Estimate the Probability of Malignancy in Patients with Pulmonary Nodules. Ann Am Thorac Soc 2018;15(10):1117–1126.
15. Tanner NT, Porter A, Gould MK, Li XJ, Vachani A, Silvestri GA. Physician Assessment of Pretest Probability of Malignancy and Adherence With Guidelines for Pulmonary Nodule Evaluation. Chest 2017;152(2):263–270.
16. Balekian AA, Silvestri GA, Simkovich SM, et al. Accuracy of clinicians and models for estimating the probability that a pulmonary nodule is malignant. Ann Am Thorac Soc 2013;10(6):629–635.
17. MacMahon H, Li F, Jiang Y, Armato SG 3rd. Accuracy of the Vancouver Lung Cancer Risk Prediction Model Compared With That of Radiologists. Chest 2019;156(1):112–119.
18. Ciompi F, Chung K, van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci Rep 2017;7(1):46479.
19. Way T, Chan HP, Hadjiiski L, et al. Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance. Acad Radiol 2010;17(3):323–332.
20. 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):286–295.
21. Massion PP, Antic S, Ather S, et al. Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeterminate Pulmonary Nodules. Am J Respir Crit Care Med 2020;202(2):241–249.
22. Baldwin DR, Gustafson J, Pickup L, et al. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. Thorax 2020;75(4):306–312.
23. Dotson TL, Filippini C, Arteta C, et al. AI-Based Computer-Aided Diagnosis (CADx) Improves Stratification Decisions on Indeterminate Pulmonary Nodules: An MRMC Reader Study. Abstract presented at: American Thoracic Society 2020 International Conference; May 1, 2020; A7691.
24. Smith BJ, Hillis SL. Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities. *Proceedings Volume 11316, Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*. 2020. Proceedings from the 2020 SPIE Medical Imaging Conference.
25. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Acad Radiol 2008;15(5):647–661.

26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.
27. Farjah F, Monsell SE, Gould MK, et al. Association of the Intensity of Diagnostic Evaluation With Outcomes in Incidentally Detected Lung Nodules. JAMA Intern Med 2021;181(4):480–489.
28. Wiener RS, Gould MK, Slatore CG, Fincke BG, Schwartz LM, Woloshin S. Resource use and guideline concordance in evaluation of pulmonary nodules for cancer: too much and too little care. JAMA Intern Med 2014;174(6):871–880.
29. Lokhandwala T, Bittoni MA, Dann RA, et al. Costs of Diagnostic Assessment for Lung Cancer: A Medicare Claims Analysis. Clin Lung Cancer 2017;18(1):e27–e34.
30. Fox AH, Tanner NT. Approaches to lung nodule risk assessment: clinician intuition versus prediction models. J Thorac Dis 2020;12(6):3296–3302.
31. Kammer MN, Massion PP. Noninvasive biomarkers for lung cancer diagnosis, where do we stand? J Thorac Dis 2020;12(6):3317–3330.
32. Ather S, Kadir T, Gleeson F. Artificial intelligence and radiomics in pulmonary nodule management: current status and future applications. Clin Radiol 2020;75(1):13–19.
33. Paez R, Kammer MN, Massion P. Risk stratification of indeterminate pulmonary nodules. Curr Opin Pulm Med 2021;27(4):240–248.
34. Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. JAMA 1996;276(21):1752–1755.