BMJ Open Respiratory Research

Radiomic 'Stress Test': exploration of a deep learning radiomic model in a high-risk prospective lung nodule cohort

David Xiao ,[1] Yency Forero,[2] Michael N Kammer,[2,3] Heidi Chen,[4] Rafael Paez,[2] Brent E Heideman,[2] Oreoluwa Owoseeni,[2,5] Ian Johnson,[2,5] Stephen A Deppen,[5] Eric L Grogan,[5] Fabien Maldonado[2]

To cite: Xiao D, Forero Y, Kammer MN, *et al*. Radiomic 'Stress Test': exploration of a deep learning radiomic model in a high-risk prospective lung nodule cohort. *BMJ Open Respir Res* 2025;**12**:e002687. doi:10.1136/bmjresp-2024-002687

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjresp-2024-002687).

Received 2 July 2024
Accepted 11 June 2025

[1]Department of General Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[2]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[3]Centre de Recherches en Cancérologie de Toulouse, Toulouse, France
[4]Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[5]Department of Thoracic Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Correspondence to
Dr Fabien Maldonado; fabien.maldonado@vumc.org

ABSTRACT

Background Indeterminate pulmonary nodules (IPNs) are commonly biopsied to ascertain a diagnosis of lung cancer, but many are ultimately benign. The Lung Cancer Prediction (LCP) score is a commercially available deep learning radiomic model with strong diagnostic performance in incidentally identified IPNs, but its potential use to reduce the need for invasive procedures has not been evaluated in patients with nodules for which a biopsy has been recommended.

Methods In this prospectively collected, retrospective blinded evaluation, the probability of cancer in consecutively biopsied IPNs at a tertiary care centre was calculated using the Mayo Clinic prediction model and categorised into low, intermediate and high-probability groups by applying <10% no-test and >70% treatment thresholds per British Thoracic Society guidelines. We evaluated the diagnostic performance of the Mayo Clinic model, the LCP radiomic model and an integrated model combining the LCP score with statistically selected clinical variables (age, spiculation and upper lobe location) using stepwise logistic regression. Performance was assessed using area under the receiver operating characteristic curve (AUC), F1 score and reclassification analysis based on the bias-corrected clinical net reclassification index.

Results The study population included 196 malignant and 125 benign IPNs (61% prevalence of malignancy). The Mayo Clinic model's AUC was 0.69 (0.63–0.75), LCP's AUC was 0.67 (0.61–0.73) and the integrated model combining LCP with statistically selected clinical variables (age, spiculation and upper lobe location) had the highest AUC of 0.75 (0.69–0.80). The integrated model demonstrated improved classification, with an F1 score of 0.645 (0.572–0.716) and a significantly higher AUC compared with the Mayo Clinic model (p=0.046). Reclassification analysis showed a clinical net reclassification index of 0.36 (0.21–0.53) for benign IPNs with eight correctly downgraded intermediate-risk benign nodules and no malignant nodules misclassified into the low-risk category.

Conclusion Incorporating LCP with select clinical variables results in an improvement in malignancy risk prediction and nodule classification and could reduce unnecessary invasive biopsies for IPNs.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Diagnosis and management of indeterminate pulmonary nodules (IPNs) remains a significant challenge.
⇒ Benign nodules are commonly biopsied, and optimised strategies to reduce biopsies of benign nodules are needed to reduce complications.

WHAT THIS STUDY ADDS

⇒ In this prospectively collected, retrospective blinded evaluation study of biopsied IPNs, we evaluated the potential for an integrated model that incorporates a previously-validated deep learning model for lung nodules into the Mayo Clinic model to reduce unnecessary invasive biopsies for benign nodules.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study shows the potential of currently available lung cancer risk prediction models for reducing overtreatment of benign nodules referred to interventional pulmonologists.

INTRODUCTION

Lung cancer remains the most common and deadliest malignancy in the USA and globally.[1] To reduce lung cancer mortality through early detection, lung cancer screening programmes have been implemented, for which 14 million individuals in the USA are estimated to be eligible.[2 3] In addition, the ever-increasing use of cross-sectional imaging results in an estimated 1.6 million new indeterminate pulmonary nodules (IPNs) identified annually in the USA.[4] Management of IPNs is based on estimated pretest probability of malignancy, with low-probability nodules being monitored with follow-up imaging, high-probability nodules referred for treatment and intermediate-probability nodules usually considered for biopsy.[5] Intermediate probability nodules represent a significant diagnostic challenge, with consequent delays in cancer diagnosis and unnecessary biopsies among benign nodules.[6]

Guidelines from organisations such as the American College of Chest Physicians and the British Thoracic Society (BTS) suggest the use of validated risk prediction models to estimate the pretest probability of malignancy.[7–11] These models, with the Brock, Mayo Clinic and Herder models being the most used, are based on clinical, demographic and radiological features.[12–14] The performance of these models has been evaluated in different independent cohorts with area under the receiver operating characteristic curve (AUC) levels reported up to 0.90.[15] However, in a real-world setting of patients with IPNs evaluated by experts and determined to need a biopsy, their discrimination was significantly lower.[16]

Given these limitations, we sought to evaluate if incorporating radiomic imaging analysis could improve risk prediction models in patients referred for biopsy. The utility of current radiomic and clinical models is relatively unknown in this population, which differs in many ways from both the screening-detected and incidentally-detected IPN populations from which these models were derived.[12–14 17] For our study, we evaluated the diagnostic performance of the Mayo Clinic prediction model and a previously validated Lung Cancer Prediction (LCP) score, a deep learning radiomic model cleared for clinical use in IPN evaluation and validated at our institution. We then considered whether combining both models could result in improved diagnostic test performance.

METHODS

Study design

In this study, we used a prospective-specimen collection, retrospective-blinded-evaluation design to externally validate the Mayo Clinic model and LCP Score developed by Optellum (Optellum, Oxford, UK).[18] The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Patient selection

For this study, we used the same cohort used to evaluate lung cancer prediction models in our previously published study,[16] which consecutively enrolled patients with pulmonary nodules referred to the interventional pulmonology service for navigational bronchoscopy between 7 November 2017 and 29 April 2019 at a tertiary medical centre. This cohort included adult patients with nodules less than 30 mm in largest diameter based on clinical radiology reports.[16] All CT scans used for radiomic analysis were non-contrast, axial chest CT series with a slice thickness less than 3 mm and slice spacing less than 2.5 mm. Imaging was performed using standard clinical protocols at a tertiary academic centre. These acquisition parameters were consistent with requirements for the LCP software analysis. Patients for whom an LCP score could not be calculated due to CT scan parameter incongruence with the LCP platform were excluded. All patients lost to follow-up prior to final diagnosis or end of 2-year surveillance period were excluded from the study. Malignant diagnosis was based on expert lung pathologist review of histopathological specimens. Benign diagnosis was based on specific benign histopathological and/or microbiological findings concordant with nodule characteristics or absence of growth after 2 years of follow-up of any nodule with non-specific pathologic findings.

Mayo Clinic model and LCP score

Clinical and demographic variables required to calculate the Mayo Clinic model score were prospectively collected from consultation, procedure and progress notes in electronic health records for each nodule case included in our cohort. The Mayo Clinic model score is based on patient age, smoking history, history of extrathoracic cancer equal to or greater than 5 years prior, nodule size, nodule location and spiculation. The model provides a 0–100% probability of malignancy of a pulmonary nodule.

For LCP scores, CT images were analysed using LCP software developed by Optellum (Optellum, Oxford, UK). The LCP is a deep learning model that provides an estimate of the probability of malignancy, in the form of an integral score from 1 to 10 as well as a percentage score from 0% to 100%, after an investigator selects the region where the nodule of interest is located. The LCP platform is based on the Dense Convolutional Network, a deep learning Convolutional Neural Network (CNN) architecture designed for computer vision tasks, and an eightfold cross-validation strategy was used for training and validation on the National Lung Screening Trial data.[19 20]

Statistically integrated clinical and LCP model

To derive the integrated clinical-radiomic model, we combined the LCP score with select clinical variables from the Mayo Clinic model using logistic regression. Clinical variables were selected through stepwise multivariable logistic regression based on their predictive contribution to malignancy. The final integrated model included age, spiculation and upper lobe location. The model was fit to the study cohort, and the estimated probability of malignancy was adjusted to reflect a disease prevalence of 33% using a Bayes offset, consistent with its intended application in real-world intermediate risk populations.

Statistical analysis

We evaluated the diagnostic performance of the Mayo Clinic model, the LCP model and the integrated model by measuring sensitivity, specificity, positive predictive value, negative predictive value and AUC. The AUC of the best performing integrated model was compared with that of the Mayo Clinic model as comparator using bootstrapping methodology. We also examined their impact on patient stratification by conducting a reclassification analysis. Mayo Clinic risk scores <0.1 were considered to be low probability, >0.7 were considered

high probability and 0.1–0.7 to be intermediate risk for probability of cancer in accordance with the BTS guidelines. Patients were then classified using the same cut-offs using their integrated model risk score. The bias-corrected clinical net reclassification index (cNRI) was then calculated for malignant cases and benign nodules separately by comparing the integrated models (Mayo Clinic model+LCP) with the baseline Mayo Clinic model classifications. Non-parametric bootstrapping with 10 000 samples was used for all confidence intervals and p values. We also performed a subgroup analysis on the cohort with the exclusion of the 26 patients (13.3%) who had metastatic disease. All analyses were conducted in R (R V.4.4.0).[21]

To identify which Mayo Clinic model variables contributed significantly to malignancy prediction when integrated with the LCP score, we conducted a stepwise multivariate logistic regression. The LCP score served as the base predictor, and candidate variables from the Mayo Clinic model—age, smoking history, prior extrathoracic cancer, spiculation, upper lobe location and nodule size—were considered for inclusion. Continuous variables were modelled using restricted cubic splines. Two nodules were excluded from this analysis due to incomplete clinical data.

F1 scores were computed for each model (Mayo Clinic, LCP and integrated LCP+Mayo Clinic model selected variables) to assess classification performance with a focus on balancing sensitivity and precision in this high prevalence cohort.

Additionally, we performed a prespecified subgroup analysis excluding 26 patients (13.3%) with metastatic disease, in order to focus model evaluation on primary pulmonary nodules. All statistical analyses were performed using R V.4.4.0.

Patient and public involvement

While the motivation, design and dissemination of our research were all informed by the meaningful experiences and discussions with patients in clinical settings and previous studies, there was not a formal, systematic collaboration with patients in the design, conduct, reporting, or dissemination plans of this research.

RESULTS

Study population

A total of 321 IPNs were initially included in our analysis after excluding patients with lesions that were>30 mm, patients lost to follow-up and patients for whom nodules could not be scored with the LCP for technical reasons. For the stepwise regression analysis, two additional nodules were excluded due to missing clinical information required for reconstruction of the Mayo Clinic model variables, resulting in 319 nodules analysed in that specific model. The cohort consisted of benign (n=125, 39%) and malignant nodules (n=196, 61%). Among the 125 nodules with a benign diagnosis, 57 (45.6%) did not have histopathological or microbiological specific findings but remained stable or regressed on surveillance imaging, and 24 (19.2%) were due to granulomatous disease. Adenocarcinoma represented the largest number of cancers (n=92, 46.9%), followed by squamous cell carcinoma (n=30, 15.3%), and metastatic cancer to the lung (n=27, 13.8%). The median nodule size for benign was 15 mm (IQR 11–21) and 17 mm (IQR 13–21.2) for cancer. The cancer group had a higher median age at 67 (IQR 60–72), compared with 62 (IQR 53–71) for benign lesions, a higher proportion of smokers with 156 (80%) compared with 85 (68%) in benign lesions, and a higher median Mayo Clinic score of 56% (IQR 34–76) probability of malignancy compared with 33% (IQR 18–56) probability in the benign group. A full description of the patient population is presented in table 1.

Model performances

The AUC of the Mayo Clinic model was 0.69 (95% CI, 0.63 to 0.75), while the model using selected variables from the Mayo Clinic (age, spiculation and upper lobe location) had an AUC of 0.68 (95% CI, 0.61 to 0.74). The LCP had an AUC of 0.67 (95% CI, 0.62 to 0.73). The integrated LCP model with Mayo Clinic score resulted in an AUC of 0.72 (95% CI, 0.67 to 0.78), while the integrated LCP model and the statistically selected Mayo Clinic variables performed the best with an AUC of 0.75 (95% CI, 0.69 to 0.80). The AUC of this integrated model was significantly higher than that of the Mayo Clinic model alone (p=0.046), as demonstrated by bootstrapping (table 2, figure 1, and online supplemental figure 1).

Integrated model

In the stepwise multivariable logistic regression, the variables retained alongside the LCP score were age (p=0.004), spiculation (p=0.002) and upper lobe location (p=0.026). These variables demonstrated independent predictive value and were included in the final integrated model. Prior cancer history showed marginal significance (p=0.066) and was not retained. The model excluding two nodules with missing Mayo Clinic model inputs yielded similar results to the full model. Full regression coefficients are provided in online supplemental table 1.

The F1 score of the integrated model (LCP+selected Mayo Clinic variables) was 0.645 (95% CI, 0.572 to 0.716), higher than that of the LCP model alone (0.598 (95% CI, 0.522 to 0.672)) and the Mayo Clinic model (0.211 (95% CI, 0.137 to 0.284)), demonstrating superior balance of sensitivity and precision.

Table 3 summarises the F1 scores for the Mayo Clinic model, LCP and the integrated model. The integrated LCP+Mayo Clinic select variable model demonstrated the highest F1 score (0.645 (95% CI, 0.572 to 0.716)), followed by the LCP model (0.598 (95% CI, 0.522 to 0.672)) and the Mayo Clinic model (0.211 (95% CI, 0.137 to 0.284)).

Table 1 Continuous variables expressed as median (IQR)

Characteristic	Benign	Cancer	
	n=125	n=196	
Age, years	62 (53–71)	67 (60–72)	
Sex, male	55 (44%)	87 (44%)	
Smoking	85 (68%)	156 (80%)	
Nodule size (mm)	16 (11–21)	17 (13–21)	
Mayo probability	33% (18–56%)	56% (34–76%)	
Histopathology		**Histopathology**	
Non-specific/miscellaneous	57 (45.6%)	Adenocarcinoma	92 (46.9%)
Granulomatous	24 (19.2%)	Squamous cell carcinoma	30 (15.3%)
Organising pneumonia	10 (8%)	NSCLC, not specified	21 (10.7%)
Purulent	10 (8%)	Small cell carcinoma	8 (4.1%)
Inflammation, atypical	8 (6.4%)	Carcinoid	6 (3.1%)
Pathogen	5 (4%)	Metastatic cancer	26 (13.3%)
Fibrosis/fibroelastotic scar	11 (8.8%)	Renal cell	6 (3.1%)
Other		Pancreatic	6 (3.1%)
		Colorectal	4 (2%)
		Breast	2 (1%)
		Other primaries	8 (4.1%)
		Lymphoma	6 (3.1%)
		Other	7 (3.5%)

NSCLC, non-small cell lung cancer.

Model calibration

Using the entire cohort, the models were recalibrated to the clinically relevant prevalence of 33%. The Brier score for the Mayo Clinic model was 0.24, and integrated LCP with Mayo Clinic was 0.21. The integrated LCP with Mayo Clinic select variables had a Brier score of 0.193.

Reclassification of prevalence adjusted model

Using categories defined by decision thresholds of 0.1 and 0.7 in accordance with the BTS guidelines, nodules were deemed reclassified if their prevalence adjusted radiomic based model score put them in a different probability group than the Mayo Clinic model adjusted for prevalence. We observed the cNRI bias-corrected cNRI of

Table 2 Diagnostic performance of the Mayo Clinic, LCP and integrated models

Model	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	AUC (95% CI)
Mayo	0.57 (0.50 to 0.64)	0 (0 to 0)	0.75 (0.68 to 0.82)	—	0.69 (0.63 to 0.75)
Mayo Select	0.54 (0.47 to 0.61)	0.05 (0.01 to 0.09)	0.72 (0.65 to 0.80)	1 (1 to 1)	0.68 (0.61 to 0.74)
LCP	0.49 (0.42 to 0.56)	0 (0 to 0)	0.78 (0.71 to 0.85)	—	0.67 (0.62 to 0.73)
Mayo +LCP	0.49 (0.42 to 0.56)	0 (0 to 0)	0.81 (0.74 to 0.88)	—	0.72 (0.67 to 0.78)
Mayo Select +LCP	0.54 (0.47 to 0.61)	0.06 (0.02 to 0.11)	0.80 (0.73 to 0.87)	1 (1 to 1)	0.75 (0.69 to 0.80)

Values shown include AUC, sensitivity, specificity, PPV and NPV with 95% CI.
AUC, area under the receiver operating characteristic curve; LCP, Lung Cancer Prediction score; Mayo, Mayo model; Mayo Select, Mayo model excluding all radiographic variables; NPV, negative predictive value; PPV, positive predictive value.

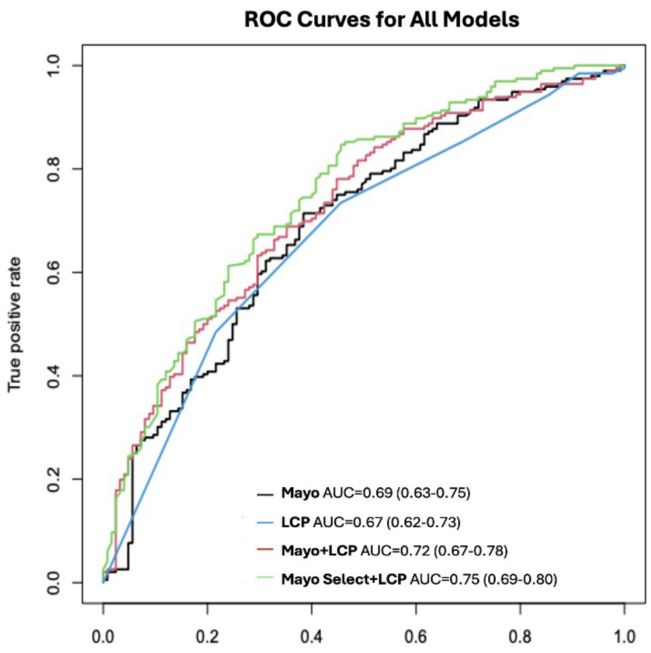

Figure 1 Receiver operating characteristic (ROC) curves for all models. AUC, area under the receiver operating characteristic curve; LCP, Lung Cancer Prediction score; Mayo, Mayo model; Mayo Select, Mayo model excluding all radiographic variables; ROC, receiver operating characteristic.

0.36 (95% CI, 0.21 to 0.53) for benign cases, and cNRI of −0.32 (95% CI, −0.49 to −0.18) for malignancy cases. For cancers, the LCP with Mayo Clinic select variables model classified 106 IPNs as high, 90 as intermediate and 0 as low risk. The Mayo Clinic model classified 158 cancerous nodules as high probability, 38 as intermediate and 0 as low probability. For benign nodules, the LCP with Mayo Clinic select variables model classified 27 nodules as high risk, 90 as intermediate and 8 as low risk. The Mayo Clinic model classified 37 benign nodules as high, 88 as intermediate, and 0 as low risk (table 4).

Subgroup analysis

A total of 295 IPNs were included in our subgroup analysis after excluding patients with non-primary, metastatic lung cancers, which represented 26 nodules (13.3%) of the full cohort. In this subgroup analysis, the AUC values for all models remained unchanged compared with analysis in the full cohort. The Mayo Clinic model achieved an AUC of 0.69 (95% CI, 0.63 to 0.75), LCP with an AUC of 0.67 (95% CI, 0.61 to 0.73), integrated LCP with Mayo Clinic model with an AUC of 0.72 (95% CI, 0.67 to 0.78) and LCP with Mayo Clinic select variables with an AUC of 0.75 (95% CI, 0.69 to 0.80). The integrated LCP with Mayo Clinic select model achieved cNRI of 0.31 (95% CI, 0.17 to 0.48) for benign nodules and cNRI of −0.31 (95% CI, −0.47 to −0.17) for malignant nodules. The Brier score for integrated LCP with Mayo Select was 0.20.

Table 3 F1 score with 95% CI for each model

Model	F1 score	95% CI
Mayo	0.211	(0.137 to 0.284)
LCP	0.598	(0.522 to 0.672)
Mayo Select+LCP	0.645	(0.572 to 0.716)

The F1 score represents the balance between sensitivity and positive predictive value in this high-prevalence cohort.
LCP, Lung Cancer Prediction score; Mayo, Mayo model; Mayo Select, Mayo model excluding all radiographic variables.

DISCUSSION

In this study, we evaluated the diagnostic performance of several prediction models: the Mayo Clinic model, a deep learning-based model for lung nodule prediction (LCP) and models that integrated the LCP with either the full Mayo Clinic model or only its clinical variables. The deep learning radiomic model was previously validated at and is currently in clinical use at our institution. We discovered that integrating the LCP with select non-radiographic variables of the Mayo Clinic model could serve as a last-step filter to correctly reclassify benign nodules in patients referred to bronchoscopic evaluation.

A large number of lung nodules are biopsied every year. While a significant number of patients undergoing lung biopsies and resections have malignant disease, far too many of these invasive procedures are done for benign lesions.[6] The patients in our cohort all underwent biopsies after evaluation by specialists at a high-volume academic centre, and nearly 40% of this cohort had benign nodules. In our previous study, we demonstrated that the available clinical prediction models provided insufficient discrimination between cancer and benign nodules in this cohort.[16] This led us to investigate whether the addition of radiomics could provide potential clinical utility, specifically using the LCP, the only Food and Drug Administration-cleared radiomic prediction model for incidentally detected lung nodules.

Our study resulted in interesting findings. To start, this cohort represents the most challenging set of nodules to adjudicate, as they represent nodules deemed suspicious enough by nodule experts to warrant biopsy. This cohort differs from the populations that were used to train both the Mayo Clinic model and the LCP, but our previous work with deep learning radiomic models encouraged us to evaluate the potential of these models in this context.[20 22] In this cohort, both the LCP and the Mayo Clinic model alone showed modest discrimination, with an AUC of 0.69 for the Mayo Clinic model and 0.68 for LCP, without significant difference between the two. However, we demonstrated that integrating the LCP with the Mayo Clinic model improves diagnostic accuracy. We observed improvement in diagnostic performance with statistically significant higher AUC values when integrating the LCP with the Mayo Clinic model. We evaluated two approaches: (1) using the full Mayo

Table 4 Risk stratification and reclassification of nodules using the integrated LCP model compared with the Mayo Clinic model, based on British Thoracic Society thresholds

Model	Cancer risk category			Benign risk category			cNRI Cancer (95% CI)	cNRI Benign (95% CI)
	Low	Intermediate	High	Low	Intermediate	High		
Mayo	0	84	112	0	88	37	Base	Base
LCP	0	101	95	0	96	27		
Mayo+ LCP	0	100	96	0	102	23	−0.44 (−0.59 to −0.28)	0.30 (0.19 to 0.41)
Mayo Select+ LCP	0	90	106	8	90	27	−0.32 (−0.49 to −0.19)	0.36 (0.21 to 0.53)

cNRI values reflect net reclassification of benign and malignant cases.
cNRI, clinical net reclassification index; LCP, Lung Cancer Prediction score; Mayo, Mayo model; Mayo Select, Mayo model excluding all radiographic variables.

Clinic model score, and (2) the 'Mayo Clinic select variables', which calculated the Mayo Clinic probability score using only the clinical variables (age, smoking history and prior cancer). The rationale behind this was that the LCP model likely already accounts for these radiographic variables, thus making the Mayo Clinic model radiographic variables redundant. Additionally, the specific radiographic variables used in the Mayo Clinic model (size, spiculation and lobe location) are balanced between cancers and benign nodules in this cohort. This integrated model with selected Mayo Clinic variables resulted in an AUC of 0.75, as compared with an AUC of 0.69 for the Mayo Clinic model alone ($p<0.05$). Although the absolute increase in AUC from 0.69 to 0.75 when integrating clinical variables with the LCP model may appear modest, this improvement was statistically significant. In the context of pulmonary nodule management, even small improvements in discrimination can meaningfully impact patient care by correctly reclassifying benign nodules and potentially avoiding unnecessary invasive procedures. Specifically, in our study, we calculated the bias-corrected clinical net reclassification indices for the two LCP integrated models compared with the Mayo Clinic model alone as the base comparator and demonstrated beneficial reclassification of benign nodules for both LCP integrated models. When using decision thresholds of 0.1 and 0.7 in accordance with BTS guidance, the LCP+Mayo Clinic Select model appropriately downgraded eight benign nodules from intermediate and high-risk categories into the low-probability category, which would result in a recommendation for CT surveillance rather than for a biopsy.[8] This would mean that, using this integrated model, we would have theoretically spared 8 of the 125 patients with benign nodules from unnecessary invasive and costly procedures, possible associated complications and procedure-related anxiety. While the absolute effect size is small and modest, each benign nodule correctly reclassified in this context would represent an important clinical implication in high-risk populations where the cost of unnecessary procedures is significant.[23 24] The integrated model resulted in a negative bias-corrected cNRI for cancers; however, in this context, only the rule-out classification is clinically meaningful as all patients are already referred to bronchoscopy. Therefore, a movement from intermediate-to-high or high-to-intermediate would not result in a change in patient management.

When developing the original LCP, clinical variables were not determined to contribute significantly to the performance of the model and thus were not included. However, the addition of clinical variables in this context may provide improvement in performance greater in real-world, high-risk populations. Additionally, the observation that the Mayo Clinic model excluding radiographic variables (nodule size, spiculation and lobe location) performed similarly to the full Mayo Clinic model suggests that in this high-risk cohort, traditional radiographic features may have diminished incremental value. Clinical and epidemiological characteristics, such as age, smoking history and prior cancer history, appeared to be more discriminative in predicting malignancy. Furthermore, the LCP deep learning-based radiomic model likely captures complex imaging features beyond conventional visual assessment, thus reducing the need for explicit inclusion of basic radiographic parameters in integrated models.

We chose to use the Mayo Clinic prediction model for this study based on our previous evaluation of this same cohort comparing the Mayo Clinic, Brock, Veterans' Affairs and Herder models.[16]

All nodules were evaluated by lung nodule experts and deemed suspicious enough to warrant biopsy despite comprehensive patient work-up, making this cohort particularly challenging. Both benign and malignant cohorts were similar in terms of radiographic and clinical features. The nodule size was one of the most notable similarities, with median nodule size for benign of 15 mm (IQR 11–21) versus 17 mm (IQR 13–21.2) for malignant ones. Nodule size is one of the most relied on and predictive variables for the determination of malignancy and is used in almost every clinical and radiomic risk prediction model. Furthermore, the cancer and benign groups were

also similar in age, with the median age of 67 for those with cancer, compared with 62 for those with benign lesions, with the same percentage of males and former or active smokers. These similarities alone would make it challenging to risk-stratify indeterminate pulmonary nodules using clinical prediction models. The specific challenge for the radiomic LCP model, which does not use any clinical data for risk stratification, likely stems from two important factors: a high prevalence of granulomatous disease and metastatic disease in this cohort. Our tertiary referral is an area where histoplasmosis is endemic. Lung granulomas caused by histoplasmosis infection pose significant diagnostic challenges due to their similar appearance to lung cancer on diagnostic imaging, including guideline-recommended fludeoxyglucose-18 positron emission tomography (FDG-PET)/CT scans. These similarities in image characteristics, such as nodule size, spiculation, tendency for upper lobe location, associated hilar and mediastinal lymphadenopathy and FDG-avidity, make histoplasmosis a great masquerader of lung cancer.[25 26] Furthermore, 13% of the cancer lesions were metastatic lesions, which were not included in training cohorts for the Mayo Clinic model and the LCP and are not clinically indicated to evaluate metastatic lung nodules. While metastatic lesions may differ in appearance from primary lung cancers, these differences are not usually sufficient to inform the use of prediction models, and correctly characterising their malignant phenotype is of utmost importance. Interestingly, in our subgroup analysis excluding all metastatic lung cancers, the performance of the models remained unchanged with the same AUC values.

This study is unique for several reasons. We had previously published similar integrated prediction models using non-Food and Drug Administration (FDA) cleared radiomic models[27] in the incidentally detected nodule setting, but at the time of writing this manuscript, our study is the first to evaluate an integrated model specifically using LCP as the radiomic component. Furthermore, this is also the first study that evaluated and developed prediction models using a high-risk cohort that consisted of all suspicious nodules referred for invasive biopsies after extensive evaluation by lung nodule experts. Furthermore, the similarities in clinical and radiographic characteristics across benign and malignant nodules make this cohort particularly challenging but also very relevant for the development and validation of prediction models. Moreover, the inclusion of patients from a region endemic with histoplasmosis, a great masquerader of lung cancer and significant cause of unnecessary benign lung nodule biopsies, contributes further to both the difficulty and relevance of this cohort for risk prediction model development.

To strengthen the integration strategy, we performed a multivariable stepwise logistic regression to identify the most informative Mayo Clinic model variables to combine with the LCP score. This analysis revealed that age, speculation and upper lobe location were independent predictors of malignancy when added to the LCP model, and these were retained in the final integrated model. In addition to AUC, we also calculated F1 scores to better capture model performance in terms of the balance between sensitivity and precision, particularly relevant in high-prevalence diagnostic settings. These analytical steps provide additional statistical support for the performance and potential clinical utility of the integrated model.

This study has several limitations, one of which is the generalisability of our cohort. All patients included were referred to one academic centre located in the Southeastern USA in an area that is endemic to histoplasmosis. Our results may not be generalisable to nodules managed in other locations with potentially different demographics, risk factors, histoplasmosis rates and different hospital settings with disparate practice patterns. In addition, our study evaluated a very specific nodule population consisting of highly suspicious nodules, and our model may not be suitable for other populations with different probability of malignancy. Finally, while we externally validated the Mayo Clinic and LCP models on our cohort, the integrated models we developed were trained and validated only on our cohort, thus lacking external validation.

Our study further contributes to the growing body of literature evaluating the promising benefits and limitations of radiomics in the diagnosis of lung cancer in indeterminate pulmonary nodules. Our study is unique in that it is one of the first studies using a cohort of prospectively collected highly suspicious nodules that have all undergone biopsy with diagnostic bronchoscopy. We demonstrated the shortcomings of clinical and radiomic models when used independently on this challenging cohort. Ultimately, we were able to develop a predictive model integrating LCP with select clinical variables that may prevent unnecessary biopsies of highly suspicious but ultimately benign nodules without delaying the timely work-up of malignant ones. Most importantly, as the clinical variables are already used routinely in the workup of pulmonary nodules and the LCP is cleared for use in both the USA and Europe, this model could potentially be implemented in clinical practice more promptly.

X David Xiao @chowda90

Acknowledgements A research version of the software licensed by Optellum was used for this study. Optellum had no role in the design, data collection, analysis, decision to publish or preparation of the manuscript. The authors have no financial relationship with Optellum.

Contributors DX, YF, MNK, HC, RP, BEH, SAD, ELG and FM contributed to conceptualisation and design of the study. DX, YF, HC, FM, OO and IJ collected, managed and interpreted the data. SAD, ELG and FM contributed to the supervision and study coordination. HC, MNK, DX, YF and FM performed the formal analysis and interpretation of the data. DX and YF wrote the first draft of the manuscript. All authors have accessed and verified all the data in the study and take responsibility for the integrity of the data and accuracy of the data analysis. All authors reviewed and edited the manuscript. All authors approved the final version of the manuscript. DX is the guarantor.

Funding This study was supported by NIH grants (T32CA106183-19 to DX, U01CA152662 to SAD and ELG, R01CA252964 to ELG and 1R01CA253923 to FM). The other authors have no conflicts of interest to declare.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Vanderbilt University Medical Center Review Board (IRB Study #140274). Written informed consent was obtained from all patients prior to collection of any data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The data that support the findings of this study are available from the corresponding author, (FM), upon reasonable request.

ORCID iD

David Xiao http://orcid.org/0000-0002-1423-8146

REFERENCES

1 Mariotto AB, Enewold L, Zhao J, *et al*. Medical Care Costs Associated with Cancer Survivorship in the United States. *Cancer Epidemiol Biomarkers Prev* 2020;29:1304–12.
2 Jonas DE, Reuland DS, Reddy SM, *et al*. Screening for Lung Cancer With Low-Dose Computed Tomography: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* 2021;325:971–87.
3 American Cancer Society. American cancer society: cancer facts and figures 2021. Atlanta, Ga American Cancer Society; 2021.
4 Gould MK, Tang T, Liu I-LA, *et al*. Recent Trends in the Identification of Incidental Pulmonary Nodules. *Am J Respir Crit Care Med* 2015;192:1208–14.
5 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:e93S–e120S.
6 Tanner NT, Aggarwal J, Gould MK, *et al*. Management of Pulmonary Nodules by Community Pulmonologists: A Multicenter Observational Study. *Chest* 2015;148:1405–14.
7 MacMahon H, Naidich DP, Goo JM, *et al*. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology* 2017;284:228–43.
8 Baldwin DR, Callister MEJ, Guideline Development Group. The British Thoracic Society guidelines on the investigation and management of pulmonary nodules. *Thorax* 2015;70:794–8.
9 American College of Radiology. Lung CT screening reporting & data system (Lung-RADS). 2022.
10 NCCN. NCCN lung cancer screening guideline V2.2022. Available: https://www.nccn.org/guidelines/guidelines-detail?category=2&id=1441 [Accessed 08 Aug 2022].
11 Swensen SJ, Silverstein MD, Ilstrup DM, *et al*. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med* 1997;157:849–55.
12 Herder GJ, van Tinteren H, Golding RP, *et al*. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. *Chest* 2005;128:2490–6.
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;369:910–9.
14 Winter A, Aberle DR, Hsu W. External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data. *Thorax* 2019;74:551–63.
15 Al-Ameri A, Malhotra P, Thygesen H, *et al*. Risk of malignancy in pulmonary nodules: A validation study of four prediction models. *Lung Cancer (Auckl)* 2015;89:27–30.
16 Heideman BE, Kammer MN, Paez R, *et al*. The Lung Cancer Prediction Model "Stress Test": Assessment of Models' Performance in a High-Risk Prospective Pulmonary Nodule Cohort. *CHEST Pulm* 2024;2:100033.
17 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;75:306–12.
18 Hansson L, Hedner T, Dahlöf B. Prospective randomized open blinded end-point (PROBE) study. A novel design for intervention trials. Prospective Randomized Open Blinded End-Point. *Blood Press* 1992;1:113–9.
19 The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N Engl J Med* 2011;365:395–409.
20 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;202:241–9.
21 R: The R project for statistical computing. Available: https://www.r-project.org/ [Accessed 22 May 2024].
22 Gao R, Tang Y, Khan MS, *et al*. Cancer risk estimation combining lung screening CT with clinical data elements. *Radiol Artif Intell* 2021;3.
23 Marmor HN, Jackson L, Gawel S, *et al*. Improving malignancy risk prediction of indeterminate pulmonary nodules with imaging features and biomarkers. *Clinica Chimica Acta* 2022;534:106–14.
24 Kammer MN, Massion PP. Noninvasive biomarkers for lung cancer diagnosis, where do we stand? *J Thorac Dis* 2020;12:3317–30.
25 Deppen SA, Blume JD, Kensinger CD, *et al*. Accuracy of FDG-PET to diagnose lung cancer in areas with infectious lung disease: a meta-analysis. *JAMA* 2014;312:1227–36.
26 Maiga AW, Deppen SA, Mercaldo SF, *et al*. Assessment of Fluorodeoxyglucose F18-Labeled Positron Emission Tomography for Diagnosis of High-Risk Lung Nodules. *JAMA Surg* 2018;153:329–34.
27 Kammer MN, Lakhani DA, Balar AB, *et al*. Integrated Biomarkers for the Management of Indeterminate Pulmonary Nodules. *Am J Respir Crit Care Med* 2021;204:1306–16.