

A Retrospective External Validation of the Cleveland Clinic Malignancy Probability Prediction Model for Indeterminate Pulmonary Nodules

Michal M. Reid, MD; Jack J. Amja, MD; Irene T. Riestra Guance, MD; Rupesh R. Andani, MBBS; Robert A. Vierkant, MS; Amit Goyal, MD; and Janani S. Reisenauer, MD

Abstract

Objective: To perform a retrospective, multicenter, external validation of the Cleveland Clinic malignancy probability prediction model for incidental pulmonary nodules.

Patients and Methods: From July 1, 2022, to May 31, 2023, we identified 296 patients who underwent tissue acquisition at Mayo Clinic (MC) (n=198) and Loyola University Medical Center (n=98) with histopathology indicating malignant (n=195) or benign (n=101). Data was collected at initial radiographic identification (point 1) and at the time of intervention (point 2). Point 3 represented the most recent data. The areas under the receiver operating characteristics were calculated for each model per time point. Calibration was evaluated by comparing the predicted and observed rates of malignancy.

Results: The areas under the receiver operating characteristics at time points 1, 2, and 3 for the MC model were 0.67 (95% CI, 0.61-0.74), 0.67 (95% CI, 0.58-0.77), and 0.70 (95% CI, 0.63-0.76), respectively. The Cleveland Clinic model (CCM) was 0.68 (95% CI, 0.61-0.74), 0.75 (95% CI, 0.65-0.84), and 0.72 (95% CI, 0.66-0.78), respectively. The mean \pm SD estimated probability for malignant pulmonary nodules (PNs) at time points 1, 2, and 3 for the CCM was 64.2 \pm 25.9, 65.8 \pm 24.0, and 64.7 \pm 24.4, which resembled the overall proportion of malignant PNs (66%). The mean estimated probability of malignancy for the MC model at each time point was 38.3 \pm 27.4, 36.2 \pm 24.4, and 42.1 \pm 27.3, substantially lower than the observed proportion of malignancies.

Conclusion: The CCM found discrimination similar to its internal validation and good calibration. The CCM can be used to augment clinical and shared decision-making when evaluating high-risk PNs.

© 2024 THE AUTHORS. Published by Elsevier Inc on behalf of Mayo Foundation for Medical Education and Research. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>) ■ Mayo Clin Proc Inn Qual Out 2024;8(4):375-383

The differential diagnosis for pulmonary nodules (PNs) is broad, and although the majority are ultimately found to be benign, invasive procedures such as biopsy or resection are often required to rule out malignancy. Currently, decisions to acquire tissue are driven by expert opinion, guidelines,^{1,2} and validated malignancy probability risk prediction models.³⁻⁶ These guidelines and risk prediction models assess a patient's risk and the likelihood that a nodule is malignant based on clinical and radiographic characteristics. However, despite these tools, a considerable number of patients with benign PNs still

undergo invasive procedures.⁵ These interventions carry inherent risks to the patient and increased expense for both the patient and the health system; therefore, better risk stratification of malignant PNs would have important clinical and economic implications.

Prediction models are most accurate when used within populations similar to those they were developed with. Currently, no externally validated model exists for incidental PNs deemed to have a high likelihood of malignancy to consider either biopsy or resection. The clinical models that do exist were developed from different populations with often

From the Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN (M.M.R., I.T.R.G., R.R.A., J.S.R.); Division of Pulmonary, Critical Care, and Sleep Medicine, University of Kansas Medical Center, Kansas City, KS (M.M.R.); Division of Pulmonary and Critical Care Medicine, Loyola University Medical Center, Maywood, IL (J.J.A., A.G.); Division of Pulmonary, Critical Care,

Affiliations continued at the end of this article.

TABLE 1. Validated Malignancy Probability Prediction Models for the Evaluation of Pulmonary Nodules^{7,8}

Model	Cleveland Clinic Model ¹⁰	Mayo Clinic Model ⁹	Herder Model ³	Veterans Affairs Model ⁴	Brock University Model ⁵
Population	Incidental PN referred for biopsy or resection	Incidental PN identified on CXR -Did not use CT, excluded patients with history of lung cancer or extrathoracic cancer within 5 years	Incidental PN, further evaluated with PET scan -Limited by nonstandardized PET reporting and variation in data acquisition and reconstruction techniques	Incidental PN on CXR, confirmed on CT imaging +/- PET -Predominantly male cohort, and majority current or former smokers	PN detected on LDCT as part of lung cancer screening program
Prevalence of malignancy in model development cohort	66.50%	23%	57%	54%	5.50%
Variables	<ul style="list-style-type: none"> ·Age ·Smoking history ·Emphysema ·Upper lobe location ·Solid and irregular/spiculated edges ·History of cancer other than lung ·FDG-PET avidity ·Change in PN size 	<ul style="list-style-type: none"> ·Age ·Smoking history ·History of extrathoracic malignancy ≥5 years ago ·Nodule diameter ·Spiculation ·Upper lobe location 	<ul style="list-style-type: none"> ·Same as Mayo Clinic Model, added FDG-PET uptake (none/faint/moderate/intense) 	<ul style="list-style-type: none"> ·Age ·Smoking history ·Time since quitting smoking ·Nodule diameter 	<ul style="list-style-type: none"> ·Age ·Sex ·Family history of lung cancer ·Emphysema ·Nodule size ·Nodule type ·Location ·Nodule count
Abbreviations: CT, computed tomography; CXR, chest radiograph; FDG, fludeoxyglucose F18; LDCT, low dose computed tomography; PET, positron emission tomography; PN, pulmonary nodule.					

different characteristics than those with a higher risk of malignancy. Please see [Table 1^{1-5,7-10}](#) for a description of validated malignancy probability prediction models for PNs. More recently, models using artificial intelligence have also been developed.⁹ In 2019, a model was developed at the Cleveland Clinic specifically targeting a population with incidental PNs with a higher likelihood of malignancy.¹⁰ It is composed of 8 different models, and the algorithm selects the most appropriate model based on the clinical and radiographic variables available. Please see [Table 1](#) for further details. The model performed well on the developmental data set, with a concordance index (C-index), statistical

equivalent to the area under the receiver operating characteristics curve (AUC), for the models ranging from 0.75-0-0.81, and on a small (n=45), internal, independent dataset (C-index 0.67), but it has yet to be externally validated.¹⁰ The analysis included a cohort of patients from Mayo Clinic (MC) in Rochester, Minnesota, and Loyola University Medical Center (LUMC) in Maywood, Illinois, who underwent either a biopsy or resection of an incidental PN and had a confirmed histopathologic diagnosis. The goal of this study was to perform a multicenter, retrospective, external validation of the Cleveland Clinic Model (CCM) and additionally, to compare the results with those derived from the Mayo

Clinic Model (MCM) to determine which model performs better in this patient population.

PATIENTS AND METHODS

The current study received institutional review board approval from internal institutional review boards at MC and LUMC. Data were collected retrospectively and included the clinical and radiographic variables used in the CCM and MCM, as detailed in Table 1. The same inclusion and exclusion criteria that were used for the development of the CCM were applied as follows. A convenience sample of patients aged 18 years or older with a PN measuring <30 mm and having a definitive histopathologic diagnosis, malignant or benign, resulting from a lung biopsy or resection at the MC from 2015-2021 and the LUMC from 2010-2021 were included. Individuals younger than 18 years, PNs found by lung cancer screening scans, PNs associated with pathologic adenopathy (short-axis diameter of >1 cm), or those without a histopathologic diagnosis were excluded. Demographic and clinical data was collected by chart review. Imaging was directly reviewed by the research team, and the imaging report was also reviewed. If there was any discrepancy between the radiology report and the researcher's direct review, the interpretation of the researcher was used.

Demographic, clinical, and radiographic data were collected for both malignant and benign PNs at various time points. Point 1 is identified when the PN was first seen on the chest computed tomography (CT) (point 1); if observation was recommended with at least a 3-month interval, subsequent imaging was labeled as point 2. A third time point (point 3) was also examined at the time of intervention. If a patient proceeded directly to intervention after time point 1, then no data exists for time point 2, and intervention data would exist at time point 3, as described above. Only positron emission tomography scans performed within 3 months of the CT being used were considered to be from the same time point. The MCM, a commonly used model for incidental PNs, was used as a comparison in the original manuscript¹⁰ and was again used as a comparison benchmark in this study. Patient data were inserted into

the CCM and the MCM for each time point to calculate the estimated probabilities of malignancy.

Data were summarized using means and standard deviations for continuous variables and frequencies and percentages for categorical variables. We compared demographic and clinical characteristics and CCM and MCM scores across benign vs malignant PN status using 2-sample *t* tests for continuous variables and Fisher exact tests for categorical variables. Three sets of comparisons were carried out: one using MC patients only, one using LUMC patients only, and one combining the 2 study groups. The discriminatory capabilities of the CCM and MCM scores to predict benign vs malignant PN status were examined by calculating sensitivity, specificity, positive predictive value, and negative predictive value for all observed model values. The AUC (also called the concordance index, or C-index) were generated using these sensitivity and specificity estimates. Power calculations run a priori revealed that a sample size of 100 benign nodules and 200 malignant nodules would result in a 95% CI with a half-width of 0.05 and an AUC of 0.80. Model discrimination was further examined by dividing the CCM and MCM model scores into approximate quintiles and examining associations with malignancy. Calibration for the CCM and MCM was assessed by dividing the sample into approximate quintiles based on predicted probabilities and plotting the median predicted value by the observed frequency of malignancy within each quintile.

RESULTS

A total of 296 patients were included in this study (MC, n=198; LUMC, n=98). One hundred and ninety-five PNs (65.9%) were malignant, and 101 (34.1%) were benign. The cohort included 174 (58.8%) females, and 268 (90.5%) were White. The mean age at intervention was 65.6 years (standard deviation [SD]=11.3), 188 (63.5%) patients had a positive smoking history, and the mean pack years were 21.7 years (SD = 24.7). The mean PN size on the initial CT was 18.2 mm (SD=11.0) and 231 were located in the upper lobe. At the time of intervention, 233 (78.7%) were solid, 50 (16.9%) were part-solid, and 13 (4.4%) were ground glass opacities. One

TABLE 2. Associations of Benign Versus Malignant Diagnosis With Demographic and Clinical Characteristics, Overall and by Study Site^a

	MC (n=198)			LUMC (n=98)			Total (N=296)		
	Benign (n=72)	Malignant (n=126)	P	Benign (n=29)	Malignant (n=69)	P	Benign (n=101)	Malignant (n=195)	P
Age at first CT			<.001 ^b			.171 ^b			<.001 ^b
N	72	126		29	69		101	195	
Mean ± SD	59.8±13.19	65.8±9.98		64.4±12.76	67.7±9.37		61.1±13.17	66.5±9.78	
Age at diagnosis			<.001 ^b			.181 ^b			<.001 ^b
N	71	125		29	69		100	194	
Mean ± SD	60.7±13.12	66.9±10.10		65.1±12.76	68.2±9.54		62.0±13.11	67.4±9.90	
Sex, n (%)			.372 ^c			.182 ^c			.142 ^c
Female	38 (52.8%)	75 (59.5%)		15 (51.7%)	46 (66.7%)		53 (52.5%)	121 (62.1%)	
Male	34 (47.2%)	51 (40.5%)		14 (48.3%)	23 (33.3%)		48 (47.5%)	74 (37.9%)	
Race, n (%)			.812 ^c			.012 ^c			.822 ^c
Asian	1 (1.4%)	4 (3.2%)		2 (6.9%)	1 (1.4%)		3 (3.0%)	5 (2.6%)	
Black	0 (0.0%)	2 (1.6%)		5 (17.2%)	4 (5.8%)		5 (5.0%)	6 (3.1%)	
Hispanic	0 (0.0%)	2 (1.6%)		2 (6.9%)	1 (1.4%)		2 (2.0%)	3 (1.5%)	
White	70 (97.2%)	116 (92.1%)		19 (65.5%)	63 (91.3%)		89 (88.1%)	179 (91.8%)	
Other	1 (1.4%)	2 (1.6%)		1 (3.4%)	0 (0.0%)		2 (2.0%)	2 (1.0%)	
Smoking status, n (%)			.232 ^c			.0032 ^c			.0042 ^c
Current	14 (19.4%)	27 (21.4%)		2 (6.9%)	13 (18.8%)		16 (15.8%)	40 (20.5%)	
Former	23 (31.9%)	53 (42.1%)		12 (41.4%)	44 (63.8%)		35 (34.7%)	97 (49.7%)	
Never	35 (48.6%)	46 (36.5%)		15 (51.7%)	12 (17.4%)		50 (49.5%)	58 (29.7%)	
Pack years smoked			.041 ^b			<.001 ^b			<.001 ^b
N	37	80		29	69		101	195	
Mean ± SD	16.0±22.38	23.8±27.44		11.0±19.01	28.5±21.10		14.5±21.49	25.5±25.42	
Previous lung cancer, n (%)			.132 ^c			NA			.122 ^c
No	70 (97.2%)	126 (100.0%)		29 (100.0%)	69 (100.0%)		99 (98.0%)	195 (100.0%)	
Yes	2 (2.8%)	0 (0.0%)		0 (0.0%)	0 (0.0%)		2 (2.0%)	0 (0.0%)	
History of cancer other than lung, n (%)			.632 ^c			.312 ^c			.342 ^c
No	52 (72.2%)	83 (68.6%)		24 (82.8%)	49 (71.0%)		76 (75.2%)	132 (69.5%)	
Yes	20 (27.8%)	38 (31.4%)		5 (17.2%)	20 (29.0%)		25 (24.8%)	58 (30.5%)	

^aAbbreviations: CT, computed tomography; LUMC, Loyola University Medical Center; MC, Mayo Clinic; SD, standard deviation.

^bTwo-sample t test

^cFisher exact test.

hundred and forty-one (47.6%) had follow-up CT scans, and 244 (82.4%) had positron emission tomography scans. Emphysema was identified on 103 (34.8%) of CT scans. Patient demographic characteristics and PN radiographic features are described in Tables 2 and 3.

Adenocarcinoma was the most common malignancy (n=138, 70.8%), followed by squamous (n=26, 13.3%), carcinoid (n=18, 9.2%), metastatic nonlung malignancy (n=7, 3.6%), small cell (n=2, 1.0%), lymphoma (n=1, 0.5%), and other (n=3, 1.5%). Among the benign nodules, granuloma (n=47,

46.5%) was the most common diagnosis, followed by other (n=29, 28.7%), hamartoma (n=16, 15.8%), organizing pneumonia (n=7, 6.9%), and pneumonia (n=2, 2.0%). Among the malignancies, PN size distributions were as follows: <1 cm (n=41, 21%), 1-2 cm (n=88, 45%), 2-3 cm (n=50, 26%), and >3 cm (n=16, 8%). Ten patients had a stage III disease, and 3 patients had a stage IV disease.

We generated an AUC for each model at each time point, and the discrimination for both models was similar. The AUC for the MCM on the combined data set at time points

TABLE 3. Associations of Benign Versus Malignant Diagnosis With Radiologic Characteristics, Overall and by Study Site^a

	MC (n=198)			LUMC (n=98)			Total (N=296)		
	Benign (n=72)	Malignant (n=126)	P	Benign (n=29)	Malignant (n=69)	P	Benign (n=101)	Malignant(n=195)	P
Location, n (%)			.35 ^b			0.88 ^b			.45 ^b
Upper lobe	38 (52.8%)	64 (50.8%)		20 (68.9%)	39 (56.5%)		58 (57.4%)	103 (52.8%)	
Other	34 (47.2%)	62 (49.2%)		9 (31.1%)	30 (43.5%)		43 (42.6%)	92 (47.2%)	
Size (mm) (point 1)			.007 ^c			0.15 ^c			.004 ^c
N	72	126		29	69		101	195	
Mean ± SD	14.6±9.04	19.6±13.92		13.2±6.18	15.4±7.19		14.2±8.31	18.1±12.13	
Size (mm) (point 2)			.11 ^c			0.66 ^c			.11 ^c
N	36	62		8	24		44	86	
Mean ± SD	13.8±5.62	16.8±10.34		14.0±3.74	15.2±7.64		13.9±5.29	16.4±9.65	
Density (point 1), n (%)			<.001 ^b			.89 ^b			.004 ^b
GGO	4 (5.6%)	11 (8.7%)		1 (3.4%)	5 (7.2%)		5 (5.0%)	16 (8.2%)	
Part-solid	4 (5.6%)	32 (25.4%)		2 (6.9%)	4 (5.8%)		6 (5.9%)	36 (18.5%)	
Solid	64 (88.9%)	83 (65.9%)		26 (89.7%)	60 (87.0%)		90 (89.1%)	143 (73.3%)	
Density (point 2), n (%)			.049 ^b			.00 ^b			.07 ^b
GGO	1 (2.8%)	6 (9.7%)		1 (12.5%)	4 (16.7%)		2 (4.5%)	10 (11.6%)	
Part-solid	7 (19.4%)	23 (37.1%)		0 (0.0%)	2 (8.3%)		7 (15.9%)	25 (29.1%)	
Solid	28 (77.8%)	33 (53.2%)		7 (87.5%)	18 (75.0%)		35 (79.5%)	51 (59.3%)	
Solid nodule border (point 1), n (%)			.04 ^c			.57 ^b			.009 ^b
Irregular/spiculated	27 (42.2%)	50 (61.0%)		16 (61.6)	40 (69.0%)		43 (47.8)	90 (64.3%)	
Lobulated	7 (10.9%)	12 (14.6%)		4 (15.4%)	11 (19.0%)		11 (12.2%)	23 (16.4%)	
Smooth	30 (46.9%)	20 (24.4%)		6 (23.1%)	7 (12.1%)		36 (40.0%)	27 (19.3%)	
Solid nodule border (point 2), n (%)			.003 ^b			.44 ^b			.007 ^b
Irregular/spiculated	11 (39.2%)	23 (69.7%)		3 (42.9%)	13 (72.2%)		14 (40.0%)	36 (70.6%)	
Lobulated	3 (10.7%)	5 (15.2%)		2 (28.6%)	2 (11.1%)		5 (14.3%)	7 (13.7%)	
Smooth	14 (50.0%)	5 (15.2%)		2 (28.6%)	3 (16.7%)		16 (45.7%)	8 (15.7%)	
Quantity, n (%)			<.001 ^b			0.67 ^b			.001 ^b
Multiple (1 dominant nodule)	4 (5.6%)	30 (23.8%)		1 (3.4%)	6 (8.7%)		5 (5.0%)	36 (18.5%)	
Solitary	68 (94.4%)	96 (76.2%)		28 (96.6%)	63 (91.3%)		96 (95.0%)	159 (81.5%)	
Emphysema present, n (%)			.21 ^b			.003 ^c			.005 ^b
Yes	19 (26.4%)	45 (35.7%)		5 (17.2%)	34 (49.3%)		24 (23.8%)	79 (40.5%)	
PET avidity, n (%)			.07 ^b			.39 ^b			
Avid	20 (51.3%)	55 (69.6%)		16 (69.6%)	46 (79.3%)		41 (51.9%)	114 (69.1%)	
Not avid	19 (48.7%)	24 (30.4%)		7 (30.4%)	12 (20.7%)		38 (48.1%)	51 (30.9%)	
Change in size			.80 ^b			.39 ^b			.83 ^b
No	8 (22.2%)	14 (19.7%)		1 (12.5%)	8 (33.3%)		9 (20.5%)	22 (23.2%)	
Yes	28 (77.8%)	57 (80.3%)		7 (87.5%)	16 (66.7%)		5 (79.5%)	73 (76.8%)	
N/A	36	55		21	45		57	100	

^aAbbreviations: GGO, ground glass opacity; LUMC, Loyola University Medical Center; MC, Mayo Clinic; PET, positron emission tomography; SD, standard deviation.
^bFisher exact test
^cTwo-sample t test

1, 2, and 3 were 0.67 (95% CI, 0.61-0.74), 0.67 (95% CI, 0.58-0.77), and 0.70 (95% CI, 0.63-0.76). The AUC for the CCM at the same time points for the combined dataset

were 0.68 (95% CI, 0.61-0.74), 0.75 (95% CI, 0.65-0.84), and 0.72 (95% CI, 0.66-0.78), respectively. The AUC for the CCM on time point 3 using the LUMC dataset was

0.73 (95% CI, 0.63-0.83) and for the MCM was 0.66 (95% CI, 0.54-0.78). The AUC for the CCM and MCM at time point 3 on the MC dataset was 0.72 (95% CI, 0.64-0.79) and 0.71 (95% CI, 0.63-0.78), respectively. The AUC for the CCM and MCM at time points 1 and 3 for the combined data set is presented in Figure A, and the full set of sensitivity, specificity, positive predictive value, and negative predictive value estimates is provided in [Supplemental Tables 1-4](#), available online at <http://www.mcpiqjournal.org>. Quintile-based associations of malignancy with model score values at time point 3 are provided in [Table 4](#). Compared with patients with CCM model scores in quintile 1 (predicted probability of malignancy $\leq 41\%$, 23 malignancies in 57 patients), those with model scores in quintiles 4 (predicted probability of malignancy 79%-86%; 52 of 59 patients, odds ratio 0.98) and 5 (predicted probability of $>86\%$; 47 of 57 patients, odds ratio 6.95) were much more likely to have malignant disease. As shown in [Table 4](#), 72% of the malignant nodules had scores $\geq 63\%$ using the CCM, whereas only 27% of the malignant nodules had scores $>68\%$ using the MCM. Most guidelines use a threshold of 65% or 70% probability of malignancy to characterize a nodule as having a high probability of malignancy and recommend intervention,^{1,2} and [Supplemental Table 5](#), available online at <http://www.mcpiqjournal.org>, demonstrates

that the CCM identified more of the malignant nodules as having a high probability when compared with the MCM.

In addition, we compared the calibration for both models by comparing predicted and observed rates of malignancy. The mean \pm SD estimated probability for malignant nodules from the combined dataset at time points 1, 2, and 3 for the CCM were 64.2 (25.9), 65.8 (24.0), and 64.7 (24.4), respectively, which closely resembled the overall proportion of PNs with malignant disease (66%). By contrast, the mean estimated probability of malignancy for the MCM at each time point was 38.3 (27.4), 36.2 (24.4), and 42.1 (27.3), respectively, substantially lower than the observed proportion of malignancies. In addition, there was a statistically significant difference ($P<.001$) between the estimated probabilities of malignancy for malignant and benign PNs for both models in each dataset and at each time point. A more detailed breakdown of the mean estimated probabilities of malignancy for malignant nodules at each time point for each site and the combined data set can be found in [Supplemental Table 6](#), available online at <http://www.mcpiqjournal.org>. In Figure B, we present a calibration plot comparing the mean probability of malignancy for each model with the proportion of patients who have malignant disease, demonstrating much better calibration of the CCM when compared with the MCM for this dataset. A summary assessment of the discrimination and calibration for each model at each site and time point reveals consistent patterns across the MC and LUMC study sites ([Supplemental Table 5](#)).

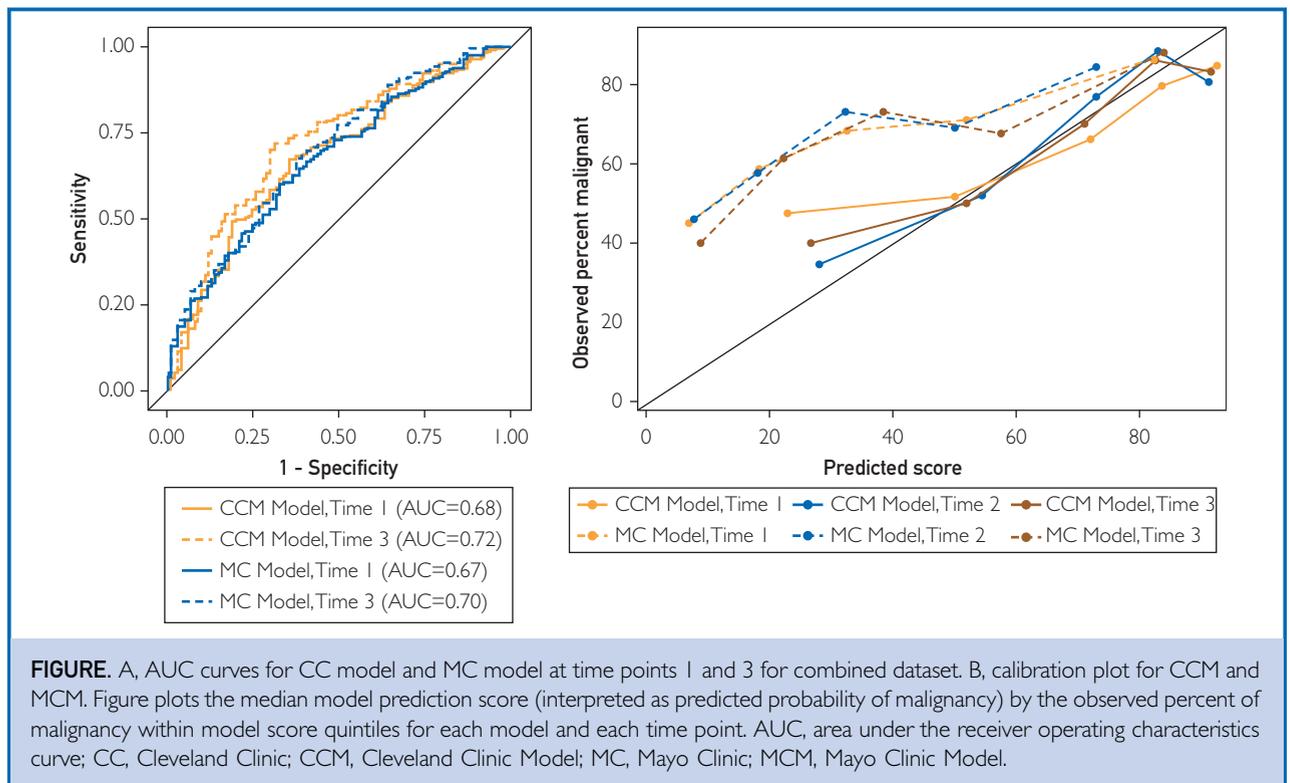
TABLE 4. Odds of Malignant Diagnosis by Quintiles of CCM and MCM at Time 3

Model score at time 3	No. malignancies (n=195)	Total no. patients (N=296)	OR (95% CI)	P
CCM				<.001
≤ 41	23	57	1.00 (ref)	
(41,63)	32	63	1.53 (0.74-3.15)	
(63,79)	41	60	3.19 (1.49-6.81)	
(79,86)	52	59	10.98 (4.25-28.40)	
>86	47	57	6.95 (2.93-16.48)	
MCM				<.001
≤ 14	21	55	1.00 (ref)	
(14,29)	39	63	2.63 (1.25-5.54)	
(29,48)	43	60	4.10 (1.87-8.95)	
(48,68)	39	58	3.32 (1.54-7.20)	
>68	53	60	12.26 (4.71-31.94)	

Abbreviations: CCM, Cleveland Clinic Model; MCM, Mayo Clinic Model; OR, odds ratio; ref, reference.

DISCUSSION

The CCM was developed to estimate the probability of malignancy in patients with PNs considered to have high enough likelihood of malignancy to recommend biopsy or resection. We collected retrospective data from 2 high-volume centers, MC and LUMC. Data were collected from the time the PN was first identifiable on imaging and from the most recent data available before biopsy or resection. The AUC generated for the CCM was similar to the value from the previous internal validation on an independent sample (C-index 0.67).



The mean estimated probabilities of malignancy for malignant PNs in the combined data set at all 3 time points for the MCM were <50%, which is in the indeterminate range for PNs. However, the mean estimated probabilities of malignancy for malignant PNs using the CCM at each time point was 69.6%, 72.9%, and 71.0%, respectively. These are all at or above most guideline thresholds of >65% or >70% to recommend resection.^{1,2} We observed a strong dose-response association, such that patients with model scores (predicted probability of malignancy) in the third (63%-79%), fourth (79%-86%), or fifth (>86%) quintile, representing 72% of all malignant nodules, had predicted malignancy probabilities at or above the guideline threshold to recommend resection and had between 3 and 11 times the odds of having malignancy when compared with those in the first quintile. Meanwhile, using the MCM, only the malignant nodules in the upper range of the fourth quintile (48%-68%) and those in the fifth quintile (>68%), were correctly classified as having a high probability of malignancy. The CCM more often correctly

classified malignant PNs as high-risk when compared with the MCM.

The calibration plot in Figure 1B shows that the CCM is well calibrated to this dataset made up of patients with high-risk PNs from 2 different external sites. Meanwhile, the MCM consistently underestimated the probability of malignancy. This suggests that using the CCM as an adjunct for malignancy likelihood assessment in PNs with a higher pre-test probability of malignancy can augment the clinical decision of when to proceed with biopsy or resection. Conversely, neither model proved effective at shifting benign nodules out of the indeterminate range into the low-risk category.

The retrospective data for this study came from 2 high-volume, academic centers in the Midwest. Moreover, the model itself was developed from a cohort at a high-volume center. The mean estimated probabilities and AUCs for the CCM were similar for both the LUMC and MC datasets, suggesting good generalizability to other populations at high-volume academic centers. Whether similar results could be expected at smaller centers or community hospitals is unclear.

In addition, data were collected by multiple contributors at both sites. Although everyone was instructed on how to collect the data to ensure consistency and the same database structure in RedCap was used, there was almost certainly some variability in data collection. However, we believe this is a strength of the study rather than a limitation. It is more representative of how the model would be used in real-world practice by multiple users with varying levels of experience. Therefore, this analysis is more reflective of how the model would perform in real-world clinical practice.

We chose to compare the CCM to the MCM for 2 reasons. First, the initial model development study compared the CCM to the MCM, so continuing that comparison with that benchmark already in place seemed appropriate. Furthermore, although the MCM was derived from and intended for a different population, it continues to be widely used in clinical practice for indeterminate, incidental PNs and is often used to risk stratify PNs similar to those that were included in our analysis. That being said, it is not surprising that the MCM consistently underestimated the probability of malignancy, as reported in the calibration plot (Figure 1B) given that the prevalence of malignant PNs included in the cohort for model development was markedly lower in the MCM when compared with the cohort used to develop the CCM (27% vs 66.5%).^{6,10} The prevalence of malignancy in the cohort for this study intentionally mirrored that of the development dataset for the CCM (65.9% and 65.5%, respectively), as this model is intended for PNs that have already been identified as having a higher likelihood of malignancy. This underscores how important it is to use a model developed from a population similar to that which one is evaluating.

We elected to use 2 time points for 2 reasons: first, to evaluate whether 1 model was able to better identify malignant nodules at either time point; and second, because size is a predictor in the MCM but not in the CCM, it was possible that using only the earliest data available, when the PNs are more likely to be smaller, could skew the analysis in favor of the CCM. We additionally used a third time point (point 3), to evaluate only the data available before the intervention to reflect a more real-world application of the model. However, there was no clinically important difference between

the models at each time point, suggesting that the CCM could be helpful at any time as part of a clinician's evaluation. It should also be noted that these time points were captured on the basis of the investigator's independent review of imaging, ie if in retrospect, the nodule was present. This may be discordant with real-world physician interpretation, which reflects the retrospective nature of this study.

Our study does have several limitations. Although it contained both males and females, smokers and nonsmokers, the cohort was predominantly White (90.5%). Therefore, it is unclear how its accuracy would be affected when used in other racial groups. Furthermore, it was developed and validated in a cohort from large academic centers; therefore, its accuracy in smaller community centers may be different. In addition, a benign nodule had definitive benign histopathologic diagnosis; however, benignity was not confirmed on serial follow-up, so it is conceivable that some of the benign PNs were ultimately found to be malignant. Last, this is a retrospective study and only evaluated malignancy probability estimates and model AUCs, and it is unclear how this model would affect clinical decision-making. Models are only helpful if they affect clinical decision-making in a way that benefits patients, such as better identification of PNs that should undergo intervention and reduction of interventions on benign PNs. Our observed AUCs of ~ 0.7 suggest that the model results may not be applicable for all patients; however, the CCM correctly identified a substantially larger proportion of patients with malignant incidental nodules as having a high probability of malignancy when compared with the MCM. Ultimately, the next step in assessing how this model could affect clinical decision-making would be a clinical utility study.

CONCLUSION

Multiple malignancy probability estimation models are available, and they are most accurate when used in populations similar to those from which they were developed. The CCM is intended for patients with incidental PNs predetermined to have a high enough likelihood of malignancy to consider biopsy or resection. The model is freely available and easily accessible online and uses common clinical and radiographic variables. This study demonstrates a successful

external validation of the CCM. On external validation, the CCM produced results similar to those on previous internal validation in terms of discrimination and found good calibration for this data set composed of data from 2 different external academic centers. Furthermore, the CCM identified a greater proportion of malignant PNs as having a high probability of malignancy when compared with the MCM. If the CCM had been applied to these PNs, it would have shifted two-thirds of the malignant PNs into the high-risk category compared with less than one-third if the MCM had been applied. These findings demonstrate the CCM can be used to augment clinical and shared decision-making when evaluating PNs being considered for biopsy or resection.

POTENTIAL COMPETING INTERESTS

Dr Reisenauer reports research grant from Intuitive, royalties from Imvana, lazarro, Mediview, consulting fees for active consulting from elucent and noah, honoraria for lectures from Astra Zenec, and participated on the Mauna Kea Advisory board. The other authors report no competing interests.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at <http://www.mcpiqjournal.org>. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AUC**, area under the receiver operating characteristics curve; **CC**, Cleveland Clinic; **CCM**, Cleveland Clinic Model; **C-index**, concordance index; **CT**, chest computed tomography; **LUMC**, Loyola University Medical Center; **MC**, Mayo Clinic; **MCM**, Mayo Clinic Model; **PN**, pulmonary nodule; **SD**, standard deviation

Affiliations (Continued from the first page of this article.): and Sleep Medicine, Hartford Healthcare Medical

Group, Hartford, CT (J.J.A.); Division of Clinical Trials and Biostatistics, Mayo Clinic, Rochester, MN (R.A.V.); and Division of Thoracic Surgery, Mayo Clinic, Rochester, MN (J.S.R.).

Correspondence: Address to Michal Reid, Division of Pulmonary, Critical Care, and Sleep Medicine, University of Kansas Medical Center, 4000 Cambridge St, Mailstop 3007 Kansas City, KS (mreid4@kumc.edu).

REFERENCES

- 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. <https://doi.org/10.1378/chest.12-2351>.
- Callister ME, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax*. 2015;70(12 suppl 2):ii1-ii54. <https://doi.org/10.1136/thoraxjnl-2015-207168>.
- 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(4):2490-2496. <https://doi.org/10.1378/chest.128.4.2490>.
- Gould MK, Ananth L, Barnett PG; Veterans Affairs SNAP Cooperative Study Group. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest*. 2007;131(2):383-388. <https://doi.org/10.1378/chest.06-1261>.
- 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. <https://doi.org/10.1056/NEJMoa1214726>.
- 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. <https://doi.org/10.1001/archinte.1997.00440290031002>.
- Mazzone PJ, Lam L. Evaluating the patient with a pulmonary nodule: a review. *JAMA*. 2022;327(3):264-273. <https://doi.org/10.1001/jama.2021.24287>.
- 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. <https://doi.org/10.1513/AnnalsATS.201803-173CME>.
- 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. <https://doi.org/10.1164/rccm.201903-0505OC>.
- Reid M, Choi HK, Han X, et al. Development of a risk prediction model to estimate the probability of malignancy in pulmonary nodules being considered for biopsy. *Chest*. 2019;156(2):367-375. <https://doi.org/10.1016/j.chest.2019.01.038>.