



Post-imaging pulmonary nodule mathematical prediction models: are they clinically relevant?

Johanna Uthoff^{1,2} · Nicholas Koehn¹ · Jared Larson¹ · Samantha K. N. Dilger^{1,2} · Emily Hammond^{1,2} · Ann Schwartz³ · Brian Mullan¹ · Rolando Sanchez⁴ · Richard M. Hoffman⁴ · Jessica C. Sieren^{1,2} · for the COPDGene Investigators

Received: 18 October 2018 / Revised: 6 February 2019 / Accepted: 14 March 2019 / Published online: 1 April 2019
© European Society of Radiology 2019

Abstract

Objectives Post-imaging mathematical prediction models (MPMs) provide guidance for the management of solid pulmonary nodules by providing a lung cancer risk score from demographic and radiologists-indicated imaging characteristics. We hypothesized calibrating the MPM risk score threshold to a local study cohort would result in improved performance over the original recommended MPM thresholds. We compared the pre- and post-calibration performance of four MPM models and determined if improvement in MPM prediction occurs as nodules are imaged longitudinally.

Materials and methods A common cohort of 317 individuals with computed tomography-detected, solid nodules (80 malignant, 237 benign) were used to evaluate the MPM performance. We created a web-based application for this study that allows others to easily calibrate thresholds and analyze the performance of MPMs on their local cohort. Thirty patients with repeated imaging were tested for improved performance longitudinally.

Results Using calibrated thresholds, Mayo Clinic and Brock University (BU) MPMs performed the best (AUC = 0.63, 0.61) compared to the Veteran's Affairs (0.51) and Peking University (0.55). Only BU had consensus with the original MPM threshold; the other calibrated thresholds improved MPM accuracy. No significant improvements in accuracy were found longitudinally between time points.

Conclusions Calibration to a common cohort can select the best-performing MPM for your institution. Without calibration, BU has the most stable performance in solid nodules ≥ 8 mm but has only moderate potential to refine subjects into appropriate workup. Application of MPM is recommended only at initial evaluation as no increase in accuracy was achieved over time.

Key Points

- *Post-imaging lung cancer risk mathematical prediction models (MPMs) perform poorly on local populations without calibration.*
- *An application is provided to facilitate calibration to new study cohorts: the Mayo Clinic model, the U.S. Department of Veteran's Affairs model, the Brock University model, and the Peking University model.*
- *No significant improvement in risk prediction occurred in nodules with repeated imaging sessions, indicating the potential value of risk prediction application is limited to the initial evaluation.*

Keywords Risk assessment · Lung neoplasms · Tomography, x-ray computed · Logistic models · Area under the curve

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00330-019-06168-x>) contains supplementary material, which is available to authorized users.

✉ Jessica C. Sieren
jessica-sieren@uiowa.edu

¹ Department of Radiology, University of Iowa, 200 Hawkins Drive cc704 GH, Iowa City, IA 52242, USA

² Department of Biomedical Engineering, University of Iowa, 5601 Seamans Center, Iowa City, IA 52242, USA

³ Karmanos Cancer Institute, Wayne State University, 4100 John R St, Detroit, MI 48201, USA

⁴ Department of Internal Medicine, University of Iowa, 200 Hawkins Drive, Iowa City, IA 52242, USA

Abbreviations

AUC	Area under the curve
BTS	British Thoracic Society
BU	Brock University model
COPD	Chronic obstructive pulmonary disorder
CT	Computed tomography
MAD	Median absolute deviation
MC	Mayo Clinic model
MPMs	Mathematical prediction models
PR	Precision recall
PU	Peking University model
ROC	Receiver-operator characteristic
TP_1	Initial imaging encounter on which the pulmonary nodule was identified
TP_F	Final imaging encounter before pulmonary nodule diagnosis
VA	U.S. Department of Veterans Affairs model

Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide [1]. Computed tomography (CT) imaging is used to characterize lung nodules. Size-based guidelines exist to provide clinicians with criteria to assess the potential malignancy of pulmonary nodules including Lung-RADS Assessment Categories, American College of Chest Physicians Clinical Practice Guidelines, and Fleischner Society Follow-Up Guidelines [2–4]. However, these have the potential to misclassify small malignant nodules and large benign nodules leading to suboptimal treatment plans [5–7]. This is particularly true of first encounters, or “de novo” nodules, which often fall into CT surveillance recommendations without access to growth information.

Pre-imaging lung cancer risk models have been produced which seek to stratify the individual’s benefit from screening thereby reducing unnecessary radiation exposure on subjects with limited benefit from CT imaging [8]. To better characterize imaging-detected nodules, post-imaging mathematical prediction models (MPMs) developed using multivariate logistic regressions of known lung cancer risk factors including family history, demographics, and radiologist-defined imaging characteristics to provide a malignancy risk stratification after an imaging encounter [9–13]. MPMs have been utilized on an ad hoc basis by clinicians seeking standardized input from evidence-based models. However, recently, an MPM was incorporated into the British Thoracic Society’s (BTS) Guidelines for Nodule Follow-up following an initial size-based stratification of risk (grade C recommendation) indicating a growing interest in the increased use of MPMs for day-to-day management of pulmonary nodule subjects [14].

This study compares four previously published post-imaging MPMs: the Mayo Clinic model (MC) [9], the U.S.

Department of Veterans Affairs model (VA) [10], the Peking University model (PU) [12], and the Brock University model (BU) [11], on a large cohort of trial subjects and a longitudinal cohort of retrospective clinical subjects. As these MPMs were developed using different imaging parameters (clinical chest radiographs [9, 12], clinical CT scans [10, 12], or lung cancer screening CT scans [11]), different proportions of malignant cases (MC, 35%; VA, 54%; PU, 61%; BU, 6%), and variable size distributions (mean size malignant/benign; MC, 17.8 mm/11.6 mm; VA, 18.9 mm/14.8 mm; PU, 21.3 mm/17.2 mm; BU, 15.7 mm/4.1 mm), we expect significant cohort dependence to be seen when each MPM was applied to an independent dataset.

While several studies have attempted to compare the accuracy of various post-imaging MPMs, they have reported performance (sensitivity, specificity) based on optimized cutoff points for their unique study cohorts as opposed to the recommended thresholds associated with a given MPM [15–17]. These studies reported that independent cohort-optimized thresholds can vary greatly from the MPM thresholds and adjustments to the cutoff used affects sensitivity and specificity values [17]. This presents a lack of clarity in the appropriate cutoff point for a given MPM to be applied in the clinical context [18]. Here, we evaluate the current clinical usefulness of MPMs using the recommended thresholds and compare the performance to our study-optimized cutoffs.

Materials and methods

Study cohorts

As mentioned previously, the MPMs investigated here have been built and tested in diverse datasets. For this study, two cohorts of subjects with pulmonary nodules were investigated: a research cohort and a longitudinal clinical cohort (Table 1).

Research trial cohort

The research cohort consisted of 317 subjects (80 malignant, 237 benign) retrospectively included from two separate prospective trials collecting high-resolution CT scans (217 COPDGene [19], 100 INHALE [20]). Data was collected with informed consent by the parent studies, and use of the de-identified data for our study was approved by the University of Iowa Institutional Review Board. While neither study was specifically aligned with the recommendations for screening lung cancer, both had de novo nodules encountered during imaging. The primary goal of the COPDGene study was to find underlying genetic factors of chronic obstructive pulmonary disorder (COPD); however, as COPD is a known comorbidity of lung cancer, an ongoing ancillary study was established to track participants with malignant nodules. The

Table 1 Subject and nodule demographics of study cohorts

Cohort	Demographics	Malignant	Benign
Research	Number of subjects	80	237
	Age (years) (mean, range)	64.0 (41–87)	62.1 (40–86)
	Sex	54F:26M	113F:124M
	Pack-years (mean, range)	40.7 (0–80)	15.8 (0–50)
	Nodule size (mean, range)	16.3 mm (4–30 mm)	9.2 mm (4–30 mm)
	< 6 mm	5	59
	≥ 6 mm to < 8 mm	6	56
	≥ 8 mm to < 15 mm	27	92
	≥ 15 mm	42	30
	LDCT screening eligible (yes:no)	48:32	85:152
	Longitudinal clinical	Subjects	16
Age (years) (mean, range)		46.5 (23–64)	61.1 (40–74)
Sex		9F:7M	10F:4M
Pack-years (mean, range)		21.2 (0–50)	14.2 (0–25)
Nodule size (mean, range)		18.9 mm (3–48 mm)	13.3 mm (3–29 mm)

LDCT screening eligibility criteria based on age between 55 and 80, and ≥ 30 pack-years who currently smoke or have quit within the past 15 years

F female, M male, LDCT low-dose computed tomography

primary goal of the INHALE study was to evaluate lung cancer risk with measures of COPD from spirometry and imaging. Demographic and historical information was collected from participants in these trials, and radiologist reports were generated to include descriptions of nodule findings. For the COPDGene study, diagnosis for each subject was either confirmed malignant (pathology) or confirmed benign (pathology, resolution, and/or 2-year stability). The INHALE study confirmed malignancy through histological confirmation accessed through the Detroit area Surveillance, Epidemiology, and End Results registry, and benign cases were selected to match size characteristics. Further information about these studies is included in the [Supplementary Information](#).

Longitudinal clinical cohort

The longitudinal clinical cohort was included as a proof of concept on MPM prediction performance improvement over time and repeated imaging. The cohort consisted of 30 subjects (16 malignant, 14 benign) with 92 clinical CT scans (Table 1). With Institutional Review Board approval, the medical records of the University of Iowa Hospitals and Clinics patients with nodules indicated were retrospectively reviewed.

Medical histories and radiological reports were reviewed for the following inclusion criteria: (1) a solid pulmonary nodule in repeated CT imaging and (2) confirmed malignant (pathology) or confirmed benign (pathology, resolution, and/or 2-year stability). For this assessment, we compared the performance of MPM predictions at (a) the initial (incidental) imaging encounter on which the pulmonary nodule was identified (TP_I), (b) the final imaging encounter before diagnosis (TP_F), and (c) across all the imaging encounters between detection and diagnosis.

Mathematical prediction models

Four MPMs were assessed: the Mayo Clinic model (MC) [9], the U.S. Department of Veterans Affairs model (VA) [10], the Peking University model (PU) [12], and the Brock University model (BU) [11]. Pertinent risk variables were manually extracted from subject records and a risk score was calculated for each CT scan (Table 2); detailed descriptions of the MPM equations and variable descriptions are provided in the [Supplementary Information](#). Unless the radiological report specifically indicated the presence of calcification, spiculation, or the absence of a border, nodules were considered non-calcified, non-spiculated, and smooth bordered. The MC and VA models discussed stratified risk into three tiers based on malignancy probability value [21, 22]. The PU model specified two categories split by a single probability cutoff value [12]. Similarly, the BU model is incorporated into the British Thoracic Society (BTS) Guidelines for nodules ≥ 8 mm in diameter with a single threshold splitting routine follow-up and additional workup [14, 23].

Statistical analysis

MPMs raw prediction performance was assessed using area under the receiver-operator characteristic curve (ROC-AUC) with 95% confidence intervals. Measures of sensitivity and specificity were calculated from Youden's J statistic optimal threshold, a common method for determining the best cutoff point, which maximizes the balance of sensitivity and specificity [24]. The stability of the Youden thresholds was assessed using median absolute deviation (MAD) below 0.05 on subset sizes between 50 and 250 subjects using 41,000 naïve bootstrapping trials sampling without replacement. As our research cohort contained class imbalance (more benign than malignant nodules), the area under the precision-recall curve (PR-AUC) was also assessed to provide a more robust analysis of performance; similar to ROC-AUC, PR-AUC is optimal at 1.0 and evaluates the distribution of separation between classes [25]. ROC-AUC and PR-AUC are not suitable assessment measurements when the predictions are discrete categories, as they are in the MPM-recommended thresholds. Instead, we assessed the performance by recommendation-induced misclassification of nodule or delay

Table 2 Tabular form of mathematical prediction model's (MPMs) base equations. Risk variables are categorized into demographic (subject-reported) and radiological (clinician-reported) factors. Units are coded in clinical terms; for use in the equation(s), sex (F = 1, M = 0) and presence (Y = 1, N = 0) are numerically coded. To obtain a prediction value for a given MPM, multiply each coefficient by the subject's risk variable value and take the summation with the base intercept/offset. The resulting number is the x in the logistic equation $e^x/(1 + e^x) =$ risk prediction. For example, performing the VA MPM prediction for a 62-year-old never smoker with a 10-mm nodule would yield $x = (62 \cdot 0.0779 + 0 \cdot 2.061 + 0 \cdot 0.0567 + 10 \cdot 0.112 - 8.404) = -2.454$; plugging into the logistic equation would yield a risk prediction = 0.079

Risk variable		Units	MPM coefficient			
			MC MPM	VA MPM	PU MPM	BU MPM
Demographical	Age	Years	0.0391	0.0779	0.07	0.0287
	Sex	F/M				0.6011
	Ever smoker	Y/N	0.7917	2.061		
	Time of smoking cessation	Years		0.0567		
	Cancer history	Y/N	1.3388			
	Family history of cancer	Y/N			1.267	
	Family history of lung cancer	Y/N				0.2961
Radiological	Emphysema	Y/N				0.2953
	Upper lobe	Y/N	0.7838			0.6581
	Diameter ^a	MM	0.1274	0.112	0.0676	-5.3854 ^a
	Spiculation	Y/N	1.0407		0.736	0.7729
	Smooth border	Y/N			-1.408	
	Calcification	Y/N			-1.615	
	Nodule type	Solid: Y/N				0
		Part solid: Y/N				0.377
		Non-solid: Y/N				-0.1276
	Nodule count	Count				-0.0824
Base intercept/offset			-6.872	-8.404	-4.496	0.2761

MPM mathematical prediction model, MC Mayo Clinic, VA Veteran's Affairs, BU Brock University, PU Peking University, F female, M male, Y presence, N absence

^a In the BU model, nodule size is defined by (diameter in millimeters/10)^{-0.5}

in ground-truth diagnosis. MPM-recommended categories were binarized into benign tagged ("low risk" or "watchful waiting") and malignant tagged ("high risk" or recommended immediate additional workup). Statistical differences between MPM classifications (inter-MPM and intra-MPM) were analyzed using McNemar and DeLong tests [26].

In conjunction with this paper, we have developed an easy-to-use application which allows independent researchers and clinicians to perform the analysis detailed in this study using their local population, including exploring calibrated thresholds and comparison of MPM performance. The application is hosted at <https://www.i-clic.uihc.uiowa.edu/resources/sieren/mpm/>.

Results

Research cohort comparison of MPMs

Calibrated thresholds equalize performance among MPMs

The four models (MC, VA, PU, BU) were applied to the research cohort ($N = 317$, 80 malignant, 237 benign) yielding four risk scores (one per MPM) per subject which were compared with

the nodule's known diagnosis (Fig. 1, solid line). The optimal AUC cutoff (Fig. 1, dashed line) was derived for each of the models. The MC (AUC, 0.63) and BU (AUC, 0.61) MPMs achieved the best separation between classes on this cohort compared to the PU (AUC, 0.55) and VA (AUC, 0.51) MPMs. The MC and BU MPMs were both statistically significantly better than the VA MPM ($p = 0.02$); all other pairwise comparisons of significance yielded p values above the assigned alpha (0.05). No MPM significantly outperformed all others, revealing relative similarity in their calibrated discriminatory capability between malignant and benign nodules. Testing the Youden threshold stability (MAD, < 0.05) at different calibration set sizes demonstrated stability at 100 subjects for three MPMS (MC, BU, PU) and stability at 145 subjects for all four MPMs (see Supplemental Information, Fig. A1).

Calibrated thresholds outperform the original recommended thresholds in workup categorization

The impact of risk stratification based on the calibrated threshold (Table 3) and MPM-associated categories (Table 4) were applied to the predictions (Fig. 1). Using the MPM-associated categories, up to 25% of the malignant lesions would have

Research Cohort MPM Predictions

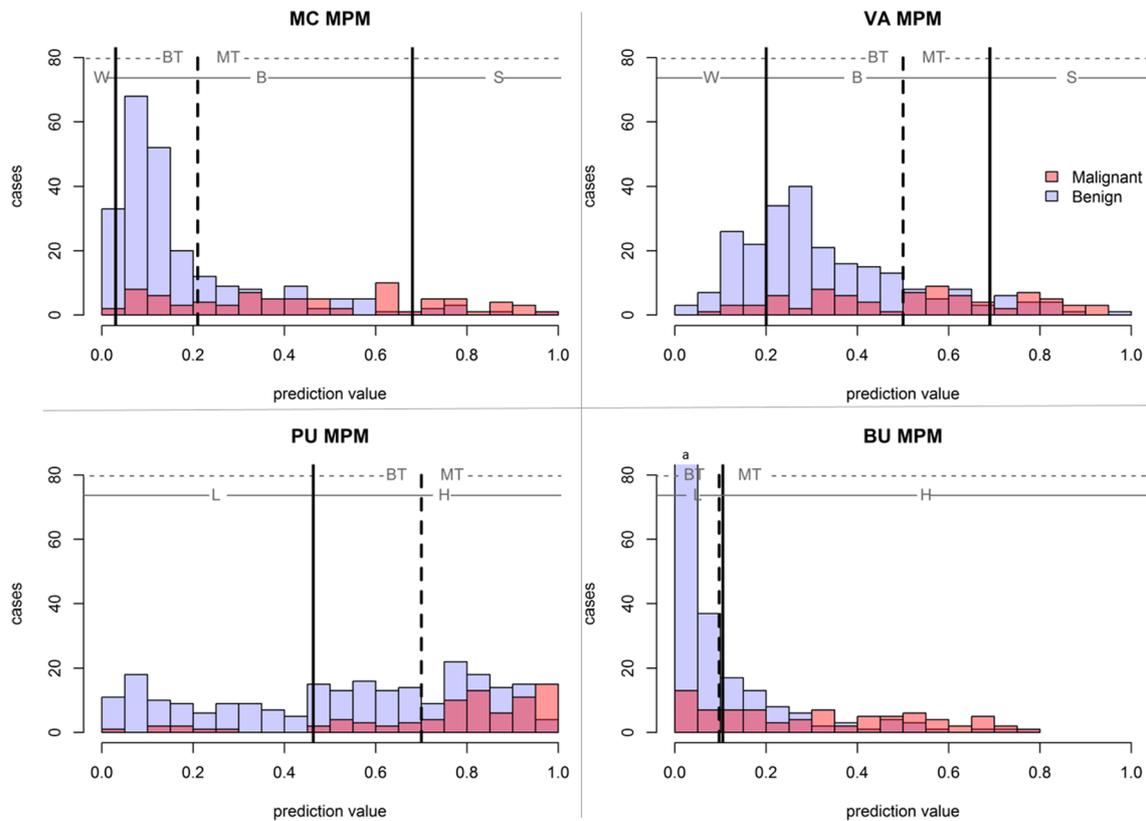


Fig. 1 Histograms of MPM predictions split based on true nodule classification. Solid lines indicate MPM-derived thresholds with MPM-assigned categories of watchful waiting (W), biopsy (B), surgery (S), low risk (L), or high risk (H). The dashed line indicates cohort AUC-derived

threshold for optimal separation of classes, with cases to the left of the line assigned “benign-tagged” and cases to the right of the line assigned “malignant-tagged”

been assigned low risk, while 25.3% to 97.5% of benign lesions would have been recommended for further workup. The BU MPM was the only model to have agreement between the Youden-optimized calibrated threshold (0.10) and the MPM-associated guidelines (0.10) for the full cohort; however, in nodules ≥ 15 mm, the Youden-optimized threshold was much higher (0.32). Furthermore, McNemar’s comparison between the optimal and recommended thresholds demonstrated significant difference between the classification accuracy of three of the MPMs (MC, VA, PU) with $p < 0.001$, indicating that calibration to the local dataset improves discriminative prowess over original MPM-associated risk categorizations. As the BU Youden optimal threshold was nearly identical to the recommended, there was no statistical significance $p = 0.99$; this stability indicates the BU MPM-associated thresholds were already well calibrated for this cohort.

Comparison to Fleischner size-based clinical management recommendations

The Fleischner Guidelines for Management of Incidental Pulmonary Nodules Detected on CT indicates that solid

pulmonary nodules have a differential follow-up using three size-based thresholds (< 6 mm; 6–8 mm; > 8 mm). Table A.4 in the Supplemental Material shows the breakdown for these categories and the clinical consequences of the follow-up recommendations. To compare the Fleischner to the calibrated MPMs, the size threshold of ≥ 8 mm was used for “high-risk” prediction and < 8 mm for “low-risk” prediction. McNemar’s analysis demonstrated that the Youden calibrated predictions for all four MPMs were statistically superior ($p < 0.01$) than the Fleischner predictions.

Calibrated thresholds improve specificity in nodules ≥ 8 mm

Size is a common variable among the MPMs and is prominent in current management guidelines. An accurate MPM risk assessment would be most clinically interesting and powerful on the nodules ≥ 8 mm to < 15 mm at baseline with 5–15% probability of malignancy in Lung-RADS—in this study, 119 nodules (27 malignant, 92 benign). The best compromising MPM at this size category was the PU model, which using MPM-associated thresholds achieved 97% sensitivity but only 36% specificity; applying the Youden optimal threshold

Table 3 Performance measures using cohort-derived optimized Youden thresholds (Fig. 1, dashed lines). Refer to Supplemental Table A.1 for complete size-based breakdown

MPM	Nodule size	Optimized threshold performance and recommendation for treatment/nodule evaluation	
MC MPM	All	< 21%	≥ 21%
		Low risk	High risk
		19M:180B	61M:57B
	≥ 8 mm to < 15 mm	24% malignant wait	24% benign immediate workup
		8M:67B	19M:25B
		30% malignant wait	30% benign immediate workup
	≥ 15 mm	2M:2B	40M:28B
		5% malignant wait	93% benign immediate workup
		< 50%	≥ 50%
VA MPM	All	Low risk	High risk
		34M:197B	46M:40B
		43% malignant wait	17% benign immediate workup
	≥ 8 mm to < 15 mm	16M:75B	11M:17B
		59% malignant wait	18% benign immediate workup
		7M:8B	35M:22B
	≥ 15 mm	17% malignant wait	73% benign immediate workup
		< 10%	≥ 10%
		Low risk	High risk
BU MPM	All	19M:178B	61M:59B
		24% malignant wait	25% benign extra procedures
		9M:61B	18M:31B
	≥ 8 mm to < 15 mm	33% malignant wait	34% benign immediate workup
		0M:1B	42M:29B
		0% malignant wait	97% benign immediate workup
	≥ 15 mm	< 70%	≥ 70%
		Low risk	High risk
		18M:154B	62M:83B
PU MPM	All	22% malignant wait	35% benign immediate workup
		9M:56B	18M:36B
		33% malignant wait	39% benign immediate workup
	≥ 8 mm to < 15 mm	6M:18B	36M:12B
		14% malignant wait	39% benign immediate workup
		39% malignant wait	39% benign immediate workup

MPM mathematical prediction model, MC Mayo Clinic, VA Veteran's Affairs, BU Brock University, PU Peking University, M malignant, B benign

achieved 67% sensitivity and improved specificity to 61% (Tables 3 and 4). Using the MPM-associated threshold, the VA model would have only missed one malignant nodule, but at the cost of 79 benign nodules undergoing biopsy (75 cases) or surgery (4 cases); the optimized threshold improved VA MPM specificity for the nodules between 8 and 15 mm to 82%. The MC model was the only MPM to completely reduce wait time on malignant lesions, sending 26 to biopsy and 1 to surgery; however, all benign lesions would have also been assigned to biopsy (91 cases) or surgery (1 case); here, applying optimized thresholds significantly improves specificity to 70% with sensitivity of 70%. In considering nodules between 8 and 15 mm in diameter, the MPM-associated recommendation thresholds for workup have little benefit in tradeoff

between sensitivity and specificity. Applying optimized thresholds improves specificity at the cost of some sensitivity.

Size exclusion prior to MPM in BTS guidelines appropriate

The BU model is unique as it has been incorporated into the BTS guidelines for management of nodules; per BTS decision flowchart, only nodules ≥ 8 mm are to be assessed with the BU MPM [27]. Tables 3 and 4 demonstrate the BU accuracy for that size-based subset. On our cohort, following the BTS exclusion of nodules < 8 mm in diameter would have meant 11 malignant and 115 benign nodules would not be assessed with the BU due to size exclusion. Applying the BU to the size excluded, no malignant and nine benign nodules are labeled

Table 4 MPM assigned categories breakdown of nodule risk prediction. Refer to Supplemental Table A.2 for complete size-based breakdown

MPM	Nodule size	Malignancy probability and associated clinical guidelines for treatment/nodule evaluation			
MC MPM		< 3%	3–68%	> 68%	
		Watchful waiting	Needle biopsy	Surgery	
	All	0M:6B	61M:225B	19M:6B	
		0% malignant wait	98% benign immediate workup		
	≥ 8 mm to < 15 mm	0M:0B	26M:91B	1M:1B	
		0% malignant wait	100% benign immediate workup		
	≥ 15 mm	0M:0B	24M:24B	18M:6B	
		0% malignant wait	100% benign immediate workup		
	VA MPM		< 20%	20–69%	> 69%
			Watchful waiting	Needle biopsy	Surgery
		All	7M:58B	51M:163B	22M:16B
			9% malignant wait	76% benign immediate workup	
≥ 8 mm to < 15 mm		1M:13B	25M:75B	1M:4B	
		4% malignant wait	86% benign immediate workup		
≥ 15 mm		3M:0B	18M:18B	21M:12B	
		7% malignant waits	100% benign immediate workup		
BU MPM			< 10%	≥ 10%	
			Low risk	Higher risk	
		All	20M:177B	60M:60B	
			25% malignant wait	25% benign extra procedures	
	≥ 8 mm to < 15 mm	9M:61B	18M:31B		
		33% malignant wait	34% benign immediate workup		
	≥ 15 mm	0M:1B	42M:29B		
		0% malignant wait	97% benign immediate workup		
	PU MPM		< 46.3%	≥ 46.3%	
			Nodule considered benign	Nodule considered malignant	
		All	8M:87B	72M:150B	
			10% malignant wait	63% benign immediate workup	
≥ 8 mm to < 15 mm		3M:33B	24M:59B		
		7% malignant wait	64% benign immediate workup		
≥ 15 mm		3M:4B	39M:26B		
		7% malignant wait	87% benign immediate workup		

MPM mathematical prediction model, MC Mayo Clinic, VA Veteran’s Affairs, BU Brock University, PU Peking University, M malignant, B benign

“high risk” by the BU MPM. Of the 11 malignant size-excluded nodules, one is recommended to be “discharged,” four are recommended for a 1-year follow-up CT, and six are recommended for a 3-month CT—indicating the need for more sophisticated discrimination techniques geared towards small nodules. The BTS recommendation to not include BU prediction on small nodules is appropriate, and as the BU threshold did not change with calibration, the recommended decision of 10% risk (0.1 prediction value) is well founded.

Longitudinal cohort

We investigated the improvement in MPM performance over the repeated imaging time point on a clinical, longitudinal

dataset of nodules imaged up to six times (average 3.1 ± 1.1) prior to diagnosis (Fig. 2). The average number of days between sequential patient imaging encounters was 214 days (± 338 days) with malignant nodules tending to have a slightly longer time between scans (218 days ± 368) compared to benign nodules (197 days ± 305).

The VA model was the only MPM to also decrease the percentage of benign nodules at TP_F that were categorized as high risk. The TP_I AUCs (MC, 0.62–0.96; VA, 0.65–0.96; BU, 0.51–0.90; PU, 0.70–0.98) were consistently higher than the TP_F AUCs in three of the MPMs (MC, 0.56–0.94; VA, 0.34–0.78; BU, 0.53–0.92; PU, 0.44–0.88). McNemar’s *p* value between TP_I and TP_F showed no statistical significance between MPM predictions at TP_I and TP_F (MC, 0.76; VA,

Longitudinal MPM Predictions

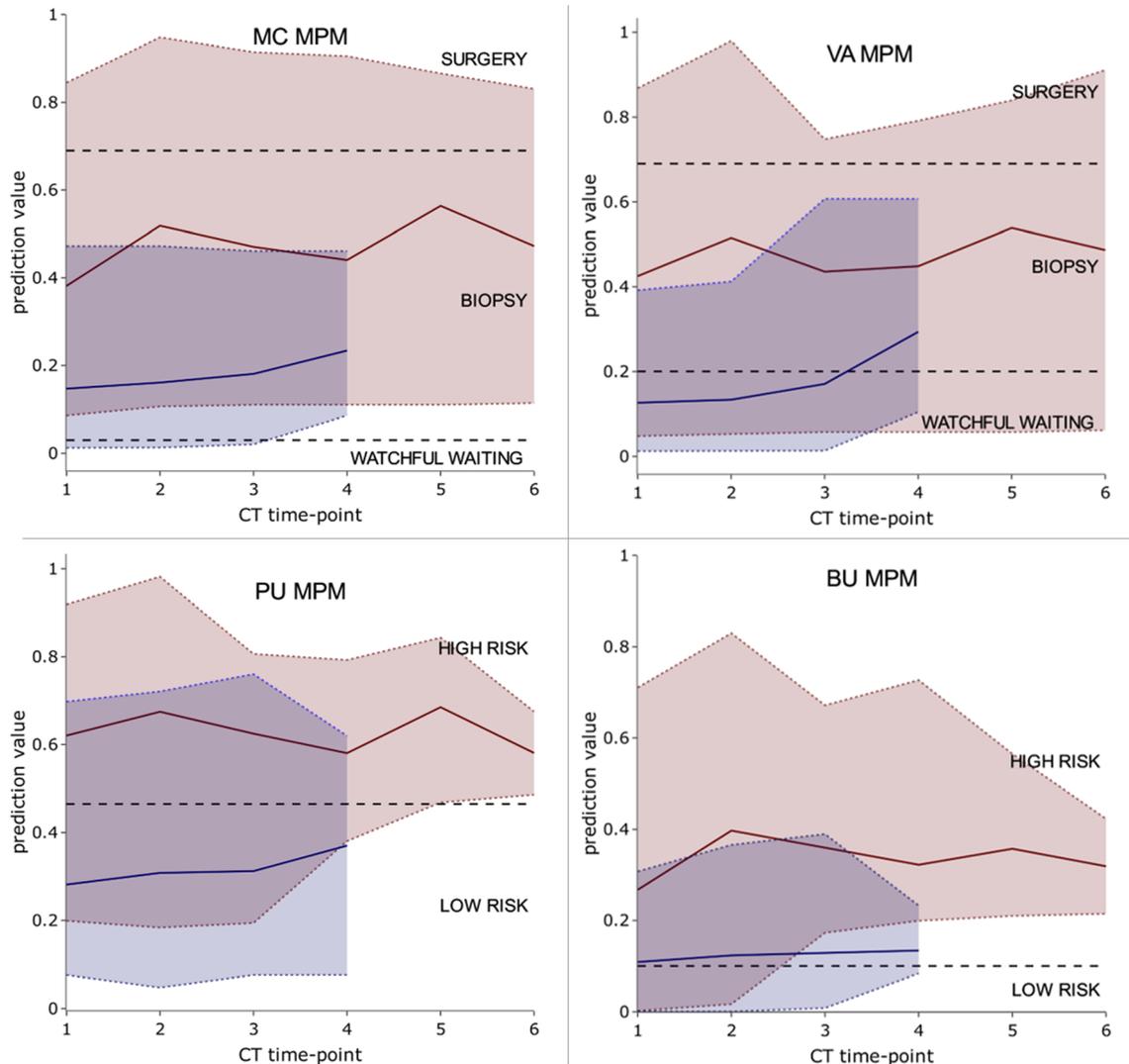


Fig. 2 MPM prediction value over CT number on longitudinal cohort. The range in prediction values for malignant (red) and benign (blue) is shown with minimum and maximum values indicated by dashed colored

lines. The average prediction value for the two classes is shown with the solid colored lines. Black dashed lines indicate MPM-derived thresholds

0.08; BU, 0.91; PU, 0.18), indicating no improvement to MPM risk assessment closer to diagnosis. This data suggests that MPM risk should not be incorporated into longitudinal evaluation of detected pulmonary nodules.

Discussion

We have applied four post-imaging MPMs to a large cohort of trial subjects and to a longitudinal cohort of clinical subjects. To our knowledge, this is the first study to compare MPMs by both the MPM-associated categories and AUC-derived (calibrated) classifications and to observe MPM stability over longitudinal scans.

Recent alignment of size-based recommendations indicates that nodules ≥ 8 mm in maximum diameter are at a heightened risk of malignancy [2, 3, 14]. Hammer et al investigated eight risk calculators on a cohort of 86 nodules (59 malignant), showing a consistent underestimation of malignancy risk. Here, we have a smaller proportion (25%) of malignancies in our cohort, yet our results concur with the assessment that care needs to be taken when assessing larger nodules (≥ 8 mm) with these MPMs [15]. The applied BU model on the ≥ 8 mm subcohort also demonstrated an underestimation of true malignancy risk with an overestimation of risk on benign nodules. Given the average nodule size in the MPM development cohorts was larger than 8 mm, it would be likely that the development-cohorts size bias would lead to more large benign nodules being tagged as suspicious.

Chung et al recently validated the BU model on two large clinical cohorts showing that while the full model achieved AUCs of 0.901–0.911, the AUC-derived optimal threshold was 1.8–4% lower than the recommended BTS guidelines; this is a difference of 4–9% in sensitivity [23]. However, that study contained a significant size bias between benign and malignant cases. While nodule diameter is not a variable in the BU model, the BTS flow diagram applies the BU model only to nodules ≥ 8 mm in diameter (≥ 300 mm³ volume). Here, we have applied the BU model in the manner recommended by BTS and demonstrated that all 11 below the size-stratified malignant nodules had a BU less than the threshold 10%. In practice, these malignant lesions would have remained untreated for at least 3 months before additional imaging.

While the BTS closely followed the original BU model study for this risk threshold, many independent surveys of MPMs have relied solely on the threshold derived from their cohort's AUC optimum [15–17]. Here, we have displayed both the AUC-derived threshold from our cohort as well as the MPM-derived thresholds. When using our cohort-derived optimal cutoff point, MPM specificity was higher (65.0–83.0%) than through using the MPM-derived assigned categories (2.5–74.7%), but MPM sensitivity was lower (58.0–78%) compared to MPM assigned categories (75.0–100%). Based on MPM assigned categories, only the MC model would have detected 100% of malignancies at the imaging time point, but this is at the cost of requiring biopsy/surgery for all benign lesions. It is important to note that some studies have reported high AUCs of MPMs in their independent cohorts, but these studies have looked solely at the AUC-derived thresholds to assess MPM performance [16, 18].

Our study has several limitations. First, the mean nodule size of the cohorts was smaller than those used to develop the MPMs. As nodule size was a common variable among the MC, VA, and PU MPMs, this could have affected the prediction results. Second, the MPMs investigated here use subject-provided demographic/historical information and radiologist-described image characteristics, both of which can suffer from subjective variability and completeness. Radiologist variability is more easily investigated and has been shown to be different between radiologists as well as within a radiologist on “coffee-break” reads in which a period of time is placed between repeated analyses [28, 29]. While to a certain extent, the variability is built into the risk models in the development dataset, the modeling of noisy data is likely different between the development cohort and the user-end radiologist. Maiga et al compared the MC model with clinician-assigned risk from qualitative statements of cancer risk, showing that the current trend of qualitative risk statements for malignancy is highly variable, and recommend a standardized scale for clinicians to follow [30]. Recent advances in CT, including dose reduction techniques and reconstruction algorithms, have the

potential to affect the signal-to-noise ratio within the scan, thereby a potential source of variation that could affect both radiologist/reader efficiency and consistency. We do believe some of this variation is already contained within the development of the MPMs given the diverse (often clinical) datasets on which they were developed. Interesting to this point, the Mayo Clinic model (chest radiographs) performs on par with the Brock University model (low-dose CT). Our cohort included only solid nodules; further studies are required to determine if MPM performance is affected when used on a cohort of subsolid tumors. Our research cohort consisted of 25.2% malignant cases and the longitudinal clinical cohort 53.3% malignant cases; the MPMs compared here were developed on cohorts of subjects with different malignancy rates (MC, 35%; VA, 54%; BU, 6%; PU, 61%). We have included the PR-AUC measure to further describe the discrimination ability of MPMs in cohorts with disproportionate numbers of malignant and benign cases.

With the move towards digitized healthcare reporting and standardization of care, computer-based risk models have a natural place in the decision pipeline. There is a benefit to adding fully automated, non-subjective systems with high performance to supplement radiologist reads with additional risk assessments. Efforts to develop tools which do not incur user subjectivity have been previously described; Mehta et al compared the MC MPM with three multivariate models developed with volumetric features extracted from semi-automatic (single click) segmentation of the nodule [17]. Machine learning for the assessment of lung cancer risk has been further developed to reduce extraction variability [31–37].

The number of lung nodules detected is set to increase with increased access to screening and clinical CT scanning. To make the screening and detection power of CT efficient and safe in practice, there is a great need for better informed decision-making. Given proper assessment and application, post-imaging risk models have the potential to improve decision-making processes. While standardization and widespread usage of these automated techniques have yet to happen, MPMs are being utilized in clinics today. This paper has demonstrated the need for clarification in malignancy thresholds reported and demonstrated the cohort dependence built into these MPMs. We thereby recommend if an MPM is to be utilized for newly detected pulmonary nodules that it is first calibrated with a retrospectively collected dataset (≥ 100 subjects) from the utilizing institution to ensure a locally optimal threshold value. We have developed an easy-to-use web-based application to assist institutions in performing MPM calibration and comparison of performance metrics between models. The application allows MPM discriminative power to be assessed using either ROC-AUC (balanced cohort) or PR-AUC (unbalanced cohort) measures and provides sensitivity and specificity. The lack of improvement in risk prediction from these MPMs over time suggests caution in the utility of these tools during the surveillance stage of clinical management.

Acknowledgements We thank Kimberly Sprenger, Debra O’Connell-Moore, Mark Escher, Patrick Thalken, and Kimberly Schroeder for technical assistance.

Funding This work was supported in part by Grant IRG-77-004-34 from the American Cancer Society, administered through the Holden Comprehensive Cancer Center at the University of Iowa. The COPDGene Study was supported by NHLBI U01 HL089897 and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprised of AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens, and Sunovion. The INHALE study was supported by Award Number R01CA141769 and P30CA022453 from the National Cancer Institute, Health and Human Services Award HHSN26120130011I, and the Herrick Foundation.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Dr. Jessica C. Sieren.

Conflict of interest The authors declare that they have no conflict of interest.

Statistics and biometry The first and last authors, as biomedical engineers, have experience with biostatistics methods. No complex statistical methods were necessary for this paper.

Informed consent The University of Iowa Institutional Review Board has approved this study (IRB 201603824). Informed consent was obtained from the research cohort participants through the parent studies, COPDgene and INHALE (including the approval of collected data for expanded research questions beyond the parent study purpose). For the retrospective clinical cohort, written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Forty subjects from study subjects or cohorts have been previously reported by our lab in a machine learning approach development [19, 20, 31].

Methodology

- Retrospective
- observational
- performed at one institution

References

1. Siegel RL, Miller KD, Jemal A (2017) Cancer statistics, 2017. *CA Cancer J Clin* 67:7–30
2. MacMahon H, Naidich DP, Goo JM et al (2017) Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology* 284:228–243
3. American College of Radiology (2014). Lung CT screening reporting and data system (Lung-RADS). Available via: <https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads>. Accessed 12/12/2018
4. Gould MK, Donington J, Lynch WR et al (2013) 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* 143:e93S–e120S
5. Mehta HJ, Mohammed TL, Jantz MA (2017) The American College of Radiology Lung Imaging Reporting and Data System: potential drawbacks and need for revision. *Chest* 151:539–543
6. Pinsky PF, Gierada DS, Black W et al (2015) Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. *Ann Intern Med* 162:485–491
7. van Riel SJ, Ciompi F, Jacobs C et al (2017) Malignancy risk estimation of screen-detected nodules at baseline CT: comparison of the PanCan model, Lung-RADS and NCCN guidelines. *Eur Radiol* 27:4019–4029
8. Gray EP, Teare MD, Stevens J, Archer R (2016) Risk prediction models for lung cancer: a systematic review. *Clin Lung Cancer* 17:95–106
9. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES (1997) The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med* 157:849–855
10. Gould MK, Ananth L, Barnett PG, Veterans Affairs SNAP Cooperative Study Group (2007) A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest* 131:383–388
11. McWilliams A, Tammemagi MC, Mayo JR et al (2013) Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med* 369:910–919
12. Li Y, Wang J (2012) A mathematical model for predicting malignancy of solitary pulmonary nodules. *World J Surg* 36:830–835
13. Herder GJ, van Tinteren H, Golding RP et al (2005) Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. *Chest* 128:2490–2496
14. Baldwin DR, Callister ME (2015) The British Thoracic Society guidelines on the investigation and management of pulmonary nodules. *Thorax* 70:794–798
15. Hammer MM, Nachiappan AC, Barbosa EJM Jr (2018) Limited utility of pulmonary nodule risk calculators for managing large nodules. *Curr Probl Diagn Radiol* 47:23–27
16. Al-Ameri A, Malhotra P, Thygesen H et al (2015) Risk of malignancy in pulmonary nodules: a validation study of four prediction models. *Lung Cancer* 89:27–30
17. Mehta HJ, Ravenel JG, Shaftman SR et al (2014) The utility of nodule volume in the context of malignancy prediction for small pulmonary nodules. *Chest* 145:464–472
18. Perandini S, Soardi GA, Motton M, Montemezzi S (2015) Critique of Al-Ameri et al. (2015) - risk of malignancy in pulmonary nodules: a validation study of four prediction models. *Lung Cancer* 90:118–119
19. Regan EA, Hokanson JE, Murphy JR et al (2010) Genetic epidemiology of COPD (COPDGene) study design. *COPD* 7:32–43
20. Schwartz AG, Lusk CM, Wenzlaff AS et al (2016) Risk of lung cancer associated with COPD phenotype based on quantitative image analysis. *Cancer Epidemiol Biomarkers Prev* 25:1341–1347
21. Swensen SJ, Silverstein MD, Edell ES et al (1999) Solitary pulmonary nodules: clinical prediction model versus physicians. *Mayo Clin Proc* 74:319–329
22. Cummings SR, Lillington GA, Richard RJ (1986) Managing solitary pulmonary nodules. The choice of strategy is a “close call”. *Am Rev Respir Dis* 134:453–460
23. Chung K, Mets OM, Gerke PK et al (2018) Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population. *Thorax* 73:857–863
24. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3:32–35

25. Jesse D, Goadrich M (2006) The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 25–29 June 2006, pp 233–240
26. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12: 153–157
27. Callister ME, Baldwin DR, Akram AR et al (2015) British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax* 70(Suppl 2):ii1–ii54
28. McNitt-Gray MF, Kim GH, Zhao B et al (2015) Determining the variability of lesion size measurements from CT patient data sets acquired under “no change” conditions. *Transl Oncol* 8:55–64
29. Lin H, Huang C, Wang W, Luo J, Yang X, Liu Y (2017) Measuring interobserver disagreement in rating diagnostic characteristics of pulmonary nodule using the lung imaging database consortium and image database resource initiative. *Acad Radiol* 24:401–410
30. Maiga AW, Deppen SA, Massion PP et al (2018) Communication about the probability of cancer in indeterminate pulmonary nodules. *JAMA Surg* 153:353–357
31. Dilger SK, Uthoff J, Judisch A et al (2015) Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. *J Med Imaging (Bellingham)* 2:041004
32. Dhara AK, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N (2016) A combination of shape and texture features for classification of pulmonary nodules in lung CT images. *J Digit Imaging* 29: 466–475
33. Ferreira JR Jr, Oliveira MC, de Azevedo-Marques PM (2017) Characterization of pulmonary nodules based on features of margin sharpness and texture. *J Digit Imaging* 31:451–463
34. Nibali A, He Z, Wollersheim D (2017) Pulmonary nodule classification with deep residual networks. *Int J Comput Assist Radiol Surg* 12:1799–1808
35. Sun W, Zheng B, Qian W (2017) Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis. *Comput Biol Med* 89:530–539
36. Way TW, Sahiner B, Chan HP et al (2009) Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. *Med Phys* 36: 3086–3098
37. Zhu Y, Tan Y, Hua Y, Wang M, Zhang G, Zhang J (2010) Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. *J Digit Imaging* 23: 51–65

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.