ABSTRACT

Purpose: Nodule evaluation is challenging and critical to diagnose multiple pulmonary nodules (MPNs). We aimed to develop and validate a machine learning-based model to estimate the malignant probability of MPNs to guide decision-making.

Experimental Design: A boosted ensemble algorithm (XGBoost) was used to predict malignancy using the clinicoradiologic variables of 1,739 nodules from 520 patients with MPNs at a Chinese center. The model (PKU-M model) was trained using 10-fold cross-validation in which hyperparameters were selected and fine-tuned. The model was validated and compared with solitary pulmonary nodule (SPN) models, clinicians, and a computer-aided diagnosis (CADx) system in an independent transnational cohort and a prospective multicentric cohort.

Results: The PKU-M model showed excellent discrimination (area under the curve; AUC (95% confidence interval (95% CI)), 0.909 (0.854–0.946)) and calibration (Brier score, 0.122) in the development cohort. External validation (853 nodules) revealed that the AUC of the PKU-M model was 0.890 (0.859–0.916), higher than those of the Brock model [0.806 (0.771–0.838)], PKU model [0.780 (0.743–0.817)], Mayo model [0.739 (0.697–0.776)], and VA model [0.682 (0.640–0.722)]. Prospective comparison (200 nodules) showed that the AUC of the PKU-M model [0.871 (0.815–0.915)] was higher than that of surgeons [0.790 (0.711–0.852), 0.741 (0.662–0.804), and 0.727 (0.650–0.788)], radiologist [0.748 (0.671–0.814)], and the CADx system [0.757 (0.682–0.818)]. Furthermore, the model outperformed the clinicians with an increase of 14.3% in sensitivity and 7.8% in specificity.

Conclusions: After its development using machine learning algorithms, validation using transnational multicentric cohorts, and prospective comparison with clinicians and the CADx system, this novel prediction model for MPNs presented solid performance as a convenient reference to help decision-making.

Introduction

Multiple pulmonary nodules (MPNs) are radiographic opacities up to 3 cm in the lung, with an associated atelectasis, hilar enlargement, or pleural effusion (1, 2). Because of the widespread use of thoracic CT, MPNs have become an increasingly recognized phenomenon, with a detection rate ranging from 6.8% to 50.9% in lung cancer screening trials (3–5). Previous studies have reported the benign rate as high as 40% for a diagnostic operation after nodule detection (6–8), emphasizing the importance of careful nodule evaluation before invasive procedures to minimize surgical risk and reduce unnecessary pulmonary function loss. Unlike solitary pulmonary nodules (SPNs) in which PET/CT and percutaneous biopsy can help in the differential diagnosis, the accuracy of PET/CT in MPNs, particularly in multiple ground-glass opacities (GGOs), is unsatisfactory (6, 9). In addition, performing a biopsy for every nodule shown in a CT scan is impossible. Consequently, a tool to pretest the malignancy for MPNs to guide subsequent management is needed.

Existing guidelines recommended several clinical prediction models based on logistic regression (10–13) to help estimate the risk of cancer before decision-making (1, 14, 15). Although these models are externally validated, they almost focus on SPNs, with no evidence of accuracy in MPNs. The Brock model is the only one that includes an analysis of MPNs; however, it is designed for screening populations with an extremely low prevalence of cancer (3.7%–5.5%), limiting its-use to evaluate patients for surgery. In recent years, some artificial intelligence (AI) products have been applied clinically, including computer-aided detection/diagnosis systems (CADe/x), which help identify or diagnose nodules from CT scans (16–18). However, none of
Chen et al.

Translational Relevance

This study developed the first prediction model for multiple pulmonary nodules (MPNs) using a novel machine learning algorithm. After external validation in independent transnational cohorts, the model showed excellent discrimination and calibration, outperforming existing multivariable risk prediction models (solitary pulmonary nodule or screening models). Furthermore, the model performed better than radiologists, surgeons, and a well-trained artificial intelligence diagnosis system in prospective comparison. This model has been implemented as a web-based model so that clinicians can obtain the estimated probability of malignancy by filling in the clinical and radiographic features of the patients. Therefore, the established model will conveniently guide precision diagnosis of MPNs before surgical treatment, thus decreasing unnecessary invasive procedures.

Materials and Methods

Study design

MPNs are defined as more than one nodule in the lung parenchyma detected on thoracic CT scans, with each nodule ranging from 4 to 30 mm. Our study comprised three key components (Fig. 1). First, we constructed a prediction model to pretest the nodule-level probability of malignancy using 11 typical machine learning algorithms and selecting the best performing algorithm for further optimization. Second, the model was externally validated in an independent transnational MPN cohort to test the performance and compare it to the existing mathematical SPN models. Third, we conducted a prospective comparison in predictive accuracy among our model, clinicians, and a well-trained CADx system using another multicentric cohort.

Variable collection

The sociodemographic variables were extracted through the electronic medical record system and included age, sex, family history of lung cancer, smoking status, smoking quantity (pack-year), time since quitting smoking, and elevation of a subset of tumor markers [carcinoembryonic antigen (normal: <5 ng/mL), cytokeratin 19 fragment (normal: <30 ng/mL), neuron-specific enolase (normal: <12.5 U/mL), cancer antigen 199 (normal: <27 U/mL), cancer antigen 125 (normal: <35 U/mL)].

The radiographic characteristics of the MPNs on CT scans were extracted by two board-certified clinicians (a radiologist with 10 years of experience and a thoracic surgeon with 15 years of experience) who were blinded to the final pathologic diagnoses and our model development process; any conflicts were adjudicated by a third investigator. The radiographic characteristics included the nodule size (maximum transverse size measured on the lung window setting), visual nodule type (pure GGO, part-solid, or solid), nodule location (upper, middle or lower lobe; left or right side), nodule distribution (unilateral or bilateral), and nodule count per scan. The presence of emphysema, spiculation, lobulation, pleural retraction sign, unclear border, and calcification component was also recorded. Perifissural nodules were excluded because they were demonstrated to be benign in previous studies (13–15).

Machine learning

Model selection

A random subset of four of five of the nodules in the development cohort was selected as the training set, and the remaining nodules comprised the test set. We ran a series of typical and recent machine learning algorithms on the training set to identify the best model. After selecting the model, the performance was assessed using the test set.

These products have been verified in MPNs, resulting in the lack of evidence regarding surgical evaluation for such nodules.

In this study, we sought to develop a web-based machine learning model, using clinical and radiographic characteristics to predict the probability of malignancy exclusively for MPNs, validate them with a contemporary, transnational and multicentric cohort, and prospectively compare the performance with clinicians and a well-trained CADx system using another multicentric cohort.

The independent validation cohort comprised consecutive patients who had undergone surgical treatment at four independent hospitals [Beijing Haidian Hospital (China), Beijing Chuiyangliu Hospital (China), Aerospace 731 Hospital (China), Beijing Aerospace General Hospital (China), Shijiazhuang People’s Hospital (China), Seoul National University Hospital (Korea)] from January 2016 to December 2018. Considering that including nodules without resection would increase the bias in the accurate validation process, only nodules with a definite pathology were recorded to validate the model.

The prospective comparison cohort was registered before the initiation of the study and comprised consecutive patients with MPNs who had undergone surgical treatment at a single institution and comprised consecutive patients with MPNs who had undergone surgical treatment at four independent hospitals [Beijing Haidian Hospital (China), Beijing Chuiyangliu Hospital (China), Aerospace 731 Hospital (China), Shijiazhuang People’s Hospital (China)] from January 2019 to March 2019. Nodules with explicit pathology were labeled and recorded.

All CT scans were obtained at 120 kVp, 40 to 60 mA, and rotation times of up to 1 second. Images were reconstructed at a 0.8-mm, 1.0-mm, 1.25-mm, or 1.5-mm slice width using standard mediastinal (width, 350 HU; level, 40 HU) and lung (width, 1500 HU; level, –600 HU) window-width and window-level settings. Patients with presence of pneumonia, pleural effusion, or a benign calcification pattern (fully or popcorn calcification) on thoracic CT scans were excluded. In addition, patients with a history of malignancy within 5 years, initial adjuvant therapy, or unavailable thoracic CT scans were excluded.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (19) was used as the reporting guide for our study. This study was performed in accordance with the Helsinki Declaration and was approved by the Institutional Review Board (IRB) of each center (2019PHB111–01). The prospective comparison cohort was registered at ClinicalTrials.gov (NCT03795181). Informed consent was waived by the IRB because the study was observational and noninvasive.

Study design

MPNs are defined as more than one nodule in the lung parenchyma detected on thoracic CT scans, with each nodule ranging from 4 to 30 mm. Our study comprised three key components (Fig. 1). First, we constructed a prediction model to pretest the nodule-level probability of malignancy using 11 typical machine learning algorithms and selecting the best performing algorithm for further optimization. Second, the model was externally validated in an independent transnational MPN cohort to test the performance and compare it to the existing mathematical SPN models. Third, we conducted a prospective comparison in predictive accuracy among our model, clinicians, and a well-trained CADx system using another multicentric MPN cohort.

To complete this study, multiple patient cohorts were acquired as follows.

The development cohort comprised consecutive patients with newly discovered MPNs by thin-slice thoracic CT scans who had been treated at Peking University People’s Hospital (Beijing, China) from January 2007 to December 2018. All the nodules found on CT scans were labeled and recorded to develop the prediction model. Notably, the diagnosis of lung cancer was made by pathologic examination after surgical resection. A nodule was considered benign by pathologic examination after resection or radiographic surveillance for at least 2 years if not undergoing a procedural biopsy. Including nodules without a definite pathology increased the missing data in the development process but minimized the bias in the spectrum of risk encountered by clinicians and the additional bias arising from nodules not undergoing resection.

The independent validation cohort comprised consecutive patients with MPNs who had undergone surgical treatment at six independent transnational hospitals [Beijing Haidian Hospital (China), Beijing Chuiyangliu Hospital (China), Aerospace 731 Hospital (China), Beijing Aerospace General Hospital (China), Shijiazhuang People’s Hospital (China), Seoul National University Hospital (Korea)] from January 2016 to December 2018. Considering that including nodules without resection would increase the bias in the accurate validation process, only nodules with a definite pathology were recorded to validate the model.

Published OnlineFirst February 24, 2021; DOI: 10.1158/1078-0432.CCR-20-4007
learning algorithms (adaptive boosting (AdaBoost), decision tree, logistic regression, linear support vector machine (Linear SVM), radial basis function kernel support vector machine (RBF SVM), naive Bayes, nearest neighbors, neural net, quadratic discriminant analysis (QDA), random forest, and extreme gradient boosting (XGBoost)) on the training set using all the sociodemographic and radiographic variables as the features to train each model 50 times on each default setting, and evaluated each algorithm using the test set to compare the performance (Fig. 2A). Finally, XGBoost (20)—a novel tree boosting algorithm to develop prediction models that has gained wide popularity in the machine learning community and has excellent scalability, high running speed, and state-of-the-art accuracy—had a better average performance than the other algorithms in the test; thus, it was selected for further optimization.

Model construction

The development cohort was randomly divided into 10 equal-sized folds, of which 8-folds constituted the training set, 1-fold constituted the tuning set, and 1-fold constituted the test set. Optimal model hyperparameters (e.g., number of trees or depth of each tree) were selected and fine-tuned by grid search using a 10-fold cross-validation procedure. The early-stop method was used to control overfitting (21). Finally, the machine learning model (PKU-M model) was developed and evaluated in the test set. Feature ranking was obtained by computing Shapley Additive Explanations (SHAP) values (22), which can help understand how a single feature affects the output of the model. All the machine learning procedures were implemented in Python using open-source libraries.

Statistical analysis

Analysis of the correlations between malignancy and the categorical variables was conducted using Fisher exact test or Pearson $\chi^2$ test. Student $t$ test or the Mann–Whitney $U$ test was used for continuous variables, which were expressed as means $\pm$ SD and medians with range. Receiver operating characteristic (ROC) curve analysis was performed to evaluate the discriminatory performance, and the areas under the curve (AUCs) were reported with 1,000 bootstrap bias-corrected 95% confidence intervals (95% CI, ref. 23). The Youden index, which maximizes the sum of the sensitivity and specificity, was defined to calculate the sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV; ref. 24). The Brier score (on a scale ranging from 0 to 1), which calculates the difference between the estimated and observed risk for malignancy with values closer to 0 indicating better calibration, was used to evaluate model calibration.

External validation was conducted using the independent validation cohort. The AUC of the PKU-M model was compared with that of four prior mathematical models: The Mayo model (10), Veterans Affairs (VA) model (11), Peking University (PKU) model (12), and Brock model (13). Paired comparisons of AUCs were computed using the same nonparametric bootstrap technique as that for 95% CIs (23). Decision curve analyses that showed the net benefit of using a model at different thresholds were performed to evaluate the clinical value of the models (25, 26).

A comparison among the PKU-M model, four clinicians (a chest radiologist with 5 years of experience and three thoracic surgeons with 3, 5, and 10 years of experience, respectively) and a CADx product (RX) based on three-dimensional convolutional neural network (CNN) developed by a Chinese technique company (RX) was conducted to compare the diagnostic accuracy with that of the prospective comparison cohort. The clinicians were informed to participate in a study to predict malignancy for MPNs. They were asked to do the following: (i) to estimate the probability of malignancy for each target nodule and (ii) judge the malignant risk of...
nODULES IN THREE CATEGORIES: HIGH RISK, MEDIUM RISK, AND LOW RISK. THE CLINICIANS WERE PROVIDED ACCESS TO THE SOCIODEMOGRAPHIC DATA AND WERE ALLOWED TO EVALUATE CT SCANS BY CHANGING THE LEVEL AND WIDTH OF THE WINDOW ON THE MONITOR WITHOUT LIMITATION IN READING TIME BUT WERE BLINDED TO THE PATHOLOGY AND ENTIRE MODEL DEVELOPMENT PROCESS. THE CONSISTENCY OF RISK PREDICTION AMONG THE CLINICIANS WAS ASSESSED BY KENDALL'S COEFFICIENT OF CONCORDANCE (KENDALL'S W; REF. 27). IN ADDITION, THE_PICTURE ARCHIVING AND COMMUNICATION SYSTEMS (PACS) OF CT SCANS WERE INPUT INTO THE RX TO PREDICT THE NODULES TO SIMULTANEOUSLY PRESENT THE PREDICTED PROBABILITY AND MALIGNANT RISK CATEGORY FOR EACH TARGET NODULE. A DETAILED DESCRIPTION OF THE RX IS GIVEN IN THE SUPPLEMENTARY FILE AS A REFERENCE. FINALLY, THE ESTIMATIONS OF THE PKU-M MODEL, CLINICIANS, AND RX WERE COMPARED. TO CALCULATE THE SENSITIVITY AND SPECIFICITY, ONLY THE JUDGEMENT OF “HIGH RISK” WAS CONSIDERED “TEST POSITIVE” (POSITIVE FOR MALIGNANCY) TO ASSESS THE DIAGNOSTIC PERFORMANCE. THE SENSITIVITY AND SPECIFICITY WERE COMPARED USING McNEMAR’S χ² TEST.

TO EXPLORE THE GENERALIZABILITY OF THE PKU-M MODEL IN PATIENTS WITH SPN, WE COLLECTED AN EXPLORATION COHORT COMPRISING CONSECUTIVE PATIENTS WITH SPN WHO HAD RECEIVED SURGICAL RESECTION AT PEKING UNIVERSITY PEOPLE’S HOSPITAL (CHINA) FROM OCTOBER 2018 TO DECEMBER 2018. THE PKU-M MODEL WAS VALIDATED USING THIS COHORT AND COMPARED WITH THE FOUR PRIOR MATHEMATICAL MODELS MENTIONED ABOVE.

ALL STATISTICAL ANALYSES WERE PERFORMED USING STATA/MP (VERSION 16.0) AND R SOFTWARE (VERSION 3.6.3).

RESULTS

FIVE HUNDRED TWENTY PATIENTS WERE INCLUDED IN THE DEVELOPMENT COHORT, WITH 1,739 NODULES FOUND ON CT SCANS, OF WHICH 876 NODULES

---

**Figure 2.**

Selection of machine learning algorithms and performance metrics of models in the development cohort. **A**, Selection of the machine learning algorithms based on their performance. All the models were trained and tested 50 times. The AUCs are illustrated by the boxplot. The XGBoost algorithm had the best AUC, so it was selected for further analysis. **B**, Performance of the PKU-M model by ROC curve analysis. **C**, Sensitivity and specificity versus cut-off probability plot of the PKU-M model. Decreasing sensitivity and increasing specificity are shown for increasing probability thresholds for malignancy, with a histogram for the distribution of the predicted probabilities. **D**, Feature importance plot for the PKU-M model. All the features are shown in this figure. The blue and red points in each row represent nodules having low to high values of the specific feature, while the x-axis shows the SHAP value, indicating the impact on the model (i.e., does it tend to drive the predictions towards malignancy (positive value of SHAP) or benignity (negative value of SHAP)?). Acc, accuracy; AdaBoost, adaptive boosting; FN, false negative; FP, false positive; RBF SVM, radial basis function kernel SVM; QDA, quadratic discriminant analysis; TN, true negative; TP, true positive.
Table 1. Distribution of nodule variables according to lung malignancy status (nodule-level data).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Development cohort</th>
<th>Independent validation cohort</th>
<th>Prospective comparison cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Benign N = 863</td>
<td>Malignancy N = 876</td>
<td>Total N = 1,739</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left upper lobe</td>
<td>199 (48.7%)</td>
<td>210 (51.3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.620</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left lower lobe</td>
<td>144 (49.7%)</td>
<td>146 (50.3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45 (49.5%)</td>
<td>46 (50.5%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>63 (38.0%)</td>
<td>103 (62.0%)</td>
</tr>
<tr>
<td></td>
<td>Right upper lobe</td>
<td>241 (47.5%)</td>
<td>266 (52.5%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32 (53.3%)</td>
<td>28 (46.7%)</td>
</tr>
<tr>
<td></td>
<td>Right middle lobe</td>
<td>96 (52.8%)</td>
<td>86 (47.2%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>74 (49.3%)</td>
<td>76 (50.7%)</td>
</tr>
<tr>
<td></td>
<td>Right lower lobe</td>
<td>183 (52.1%)</td>
<td>168 (47.9%)</td>
</tr>
<tr>
<td>Nodule location</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodule distribution</td>
<td>Ulnilateral</td>
<td>288 (40.5%)</td>
<td>424 (59.6%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bilateral</td>
<td>575 (36.0%)</td>
<td>452 (44.0%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pure GGO</td>
<td>388 (54.3%)</td>
<td>326 (45.7%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part-solid</td>
<td>37 (10.5%)</td>
<td>315 (89.5%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9 (5.9%)</td>
<td>144 (94.1%)</td>
</tr>
<tr>
<td></td>
<td>Solid</td>
<td>438 (65.1%)</td>
<td>235 (34.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>155 (66.8%)</td>
<td>77 (33.2%)</td>
</tr>
<tr>
<td></td>
<td>Border</td>
<td>34 (27.2%)</td>
<td>388 (72.8%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unclear</td>
<td>718 (59.5%)</td>
<td>488 (40.5%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clear</td>
<td>829 (55.3%)</td>
<td>670 (44.7%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spiculation</td>
<td>103 (23.6%)</td>
<td>333 (76.4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lobulation</td>
<td>760 (58.3%)</td>
<td>543 (41.7%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pleural retraction sign</td>
<td>34 (14.2%)</td>
<td>206 (85.8%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>103 (51.0%)</td>
<td>97 (49.0%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9 (5.9%)</td>
<td>144 (94.1%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>155 (66.8%)</td>
<td>77 (33.2%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34 (27.2%)</td>
<td>388 (72.8%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>718 (59.5%)</td>
<td>488 (40.5%)</td>
</tr>
<tr>
<td>Nodule size</td>
<td>Mean ± SD</td>
<td>6.7 ± 3.7</td>
<td>12.7 ± 6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median (range)</td>
<td>6 (4–30)</td>
<td>10 (4–30)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 (2–40)</td>
<td>10 (4–30)</td>
</tr>
<tr>
<td>Nodule count</td>
<td>Mean ± SD</td>
<td>6.2 ± 6.0</td>
<td>3.7 ± 2.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median (range)</td>
<td>4 (2–28)</td>
<td>3 (2–14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 (2–40)</td>
<td>3 (2–40)</td>
</tr>
</tbody>
</table>
(50.4%) were confirmed to be malignant. The independent validation cohort comprised 220 patients with 583 nodules having an explicit pathologic diagnosis, of which 318 nodules (54.5%) were malignant. The prospective comparison cohort included 78 patients with 200 nodules having a definite pathologic diagnosis, of which 126 nodules (63.0%) were malignant. The patient sociodemographic characteristics are shown in Supplementary Table S1, and the nodule variables are mentioned in Table 1.

**Model performance**

Machine learning produced an excellent performance in terms of predicting the probability of malignancy for MPNs with an AUC of 0.909 (95% CI, 0.854–0.946; Fig. 2B). The Brier score was 0.122, indicating a minimal difference between the predicted and observed probability of malignancy and good model calibration. As illustrated in Fig. 2C, decreasing sensitivity and increasing specificity are shown for an increasing probability of malignancy, with a histogram for the distribution of the predicted probability. We defined an optimal cut-off probability of 0.447 for the PKU-M model according to the Youden index so that the sensitivity, specificity, PPV, NPV, and accuracy were 0.807, 0.849, 0.845, 0.811, and 0.828, respectively.

SHAP values revealed the distribution of the impacts each feature had on the PKU-M model output (Fig. 2D). The nodule size, nodule type, nodule count, border, age, spiculation, lobulation, emphysema, nodule location, and nodule distribution were the top 10 most predictive features in the model. A larger nodule size, a part-solid/pure GGO nodule, an unclear border, an older patient, spiculation, lobulation, and a nodule scattered unilaterally were associated with malignant nodules, whereas the presence of a solid nodule, a larger nodule count, and emphysema were associated with benign nodules.

**External validation**

The independent validation cohort was used for external validation. The PKU-M model performed well, with an AUC of 0.890 (95% CI, 0.859–0.916), significantly higher than the AUCs of the Brock model (0.806; 95% CI, 0.771–0.838), PKU model (0.780; 95% CI, 0.743–0.817), Mayo model (0.739; 95% CI, 0.697–0.776), and VA model (0.682; 95% CI, 0.640–0.722; P < 0.001 for all; Fig. 3A). The AUC of the PKU-M model was consistently better than those of other models at each center (Supplementary Fig. S1). The Brier score was 0.1426, lower than the scores of the Brock model (0.400), PKU model (0.216), Mayo model (0.366), and VA model (0.370). Decision curve analysis (Fig. 3B) depicted that the net benefit of the PKU-M model exceeded that of any other models, indicating that it had better clinical impact at a wide range of probability thresholds.

The distribution of the probabilities of malignancy according to the different models is shown in Fig. 3C. The PKU-M model discriminated between malignant and benign nodules at a wide range of probabilities. However, the differences in the probabilities between malignant and benign nodules estimated by other models were small, limiting their use in differentiation.

The performance was subsequently evaluated in subgroups stratified by sex, age, family history of lung cancer, smoking history, tumor marker, nodule distribution, nodule type, and nodule size, and was compared with that of the prior models (Supplementary Fig. S2). The PKU-M model showed excellent discrimination ability with an AUC higher than those of the Brock model, PKU model, Mayo model, and VA model in almost all the subgroups.

![Figure 3](image-url)
Prospective comparison

The prospective comparison cohort comprising 200 nodules from 78 patients with MPN was used for prospective comparison. The AUC of the PKU-M model was 0.871 (95% CI, 0.815–0.915), significantly higher than the AUCs of the three thoracic surgeons (0.790 (95% CI, 0.711–0.852), P = 0.016; 0.741 (95% CI, 0.662–0.804), P < 0.001; and 0.727 (95% CI, 0.650–0.788), P < 0.001), the radiologist (0.748, (95% CI, 0.671–0.814, P < 0.001), and RX (0.757 (95% CI, 0.682–0.818), P < 0.001; Fig. 4A). The results were consistent at each center and are shown in Supplementary Fig. S3.

We displayed the sensitivity and specificity of the PKU-M model according to the Youden index, as well as the sensitivity and specificity of the clinicians and RX according to the risk category they assessed (Fig. 4B). The risk category evaluated by four clinicians is shown in Supplementary Fig. S4. The radiologist and thoracic surgeon (3 years of experience) classified more nodules as medium risk than the other two thoracic surgeons, indicating a significant number of nodules that seemed challenging to diagnose subjectively. Kendall’s W was 0.469, suggesting only moderate risk judgment consistency among the four clinicians. Using McNemar χ² test, the PKU-M model outperformed the average clinicians with an absolute increase of 0.143 (0.091–0.195; P < 0.001) in sensitivity and 0.078 (0.010–0.145; P = 0.018) in specificity. The PKU-M model had comparable sensitivity and specificity with RX, RX had similar specificity with average clinicians, while it outperformed the average clinicians with an absolute increase of 0.103 (P < 0.001) in sensitivity (Fig. 4C).

Generalizability exploration

To preliminarily explore the generalizability of the new model, it was evaluated in the exploration cohort, which comprised 195 surgical patients with SPN, of whom 123 (63.1%) were patients with lung cancer (Supplementary Table S2). Surprisingly, the PKU-M model maintained the remarkable discrimination competence with an AUC of 0.871 (95% CI, 0.815–0.917; Supplementary Fig. S5), comparable with the AUC of the PKU model (0.812; 95% CI, 0.750–0.871; P = 0.105) but significantly higher than the AUCs of the Brock model (0.780; 95% CI, 0.713–0.841; P = 0.013), Mayo model (0.762; 95% CI, 0.689–0.827; P = 0.006), and VA model (0.681; 95% CI, 0.597–0.755; P < 0.001). This result illustrated the excellent generalizability of our model and broadened the scope of application in patients with SPN.

Web-based model

For user-friendly access, the PKU-M model was also implemented as a web-based model. Figure 5 shows a screenshot of the models that are available on https://mpn.pkuphmodel.com/. Users can estimate the probability of cancer by filling in the patient and nodule characteristics.

Discussion

Clinicians are usually faced with the dilemma when managing pulmonary nodules: surgical resection will pose the risk of morbidity to patients, whereas CT surveillance may lead to delayed diagnosis and treatment. The problem becomes more serious when encountering MPNs. It is often impractical to perform biopsy or resection for each nodule detected on CT scans because the distribution is mostly bilateral and scattered. Besides, it is also impossible to remove all the nodules, particularly in patients with marginal pulmonary function and preoperative comorbidities, because each procedure will increase the likelihood of a poor outcome. Current guidelines recommended a
risk-based algorithm to help decision-making that a wait-and-see strategy can be performed when a low risk is present, and resection should be performed when a high risk is present (1, 14, 15). However, no tools have been designed for the risk prediction of MPNs at the point of surgical evaluation, causing confusion in managing such nodules. In this study, we used machine learning as a novel analytic approach to establish the first model that can estimate the probability of malignancy for MPNs. The machine learning model showed consistent excellent discrimination and calibration, even in external transnational validation, helping to automate the process of nodule evaluation, particularly for patients with MPNs.

Several models have been reported to predict the malignancy for incidental SPNs. The most widely used model is the Mayo model (10), which is recommended by the American College of Chest Physicians (ACCP) guidelines (1). The British Thoracic Society (BTS) guidelines (14) for the management of pulmonary nodules also recommended using composite prediction models based on clinicoradiologic factors to estimate the malignant probability of a pulmonary nodule. However, no model has been established for MPNs. Considering the heterogeneous characteristics between SPNs and MPNs, whether models for SPNs can be applied to MPNs is unknown. Therefore, we evaluated the predictive accuracy for MPNs using three representative SPN models that were mostly externally validated (10–12). The results showed that all the SPN models displayed low accuracy, indicating that previous SPN models may not be suitable for evaluating MPNs.

Machine learning algorithms specifically suited to find associations between data beyond the one-dimensional statistical approaches currently used (e.g., logistic regression). The increasing availability of computational power and storage space allows machine learning algorithms to analyze more complex data and output instantaneously (28, 29). XGBoost is a novel and popular algorithm in recent years and has won numerous machine learning competitions (20). In our study, XGBoost exceeded the other 10 algorithms, including logistic regression; thus, it was selected to construct the model. The PKU-M model was developed after considering all the candidate variables derived from previously published models. Additional variables...
specific to MPNs, such as the nodule count and nodule distribution (bilateral or unilateral), were also analyzed. Furthermore, because pulmonary GGOs have been increasingly diagnosed because of the wide adoption of thoracic CT in the past 10 years, we also included nodule type (GGO/part-solid/solid) in the process. The other strength of our model is that we included nodules that had undergone resection and radiographic surveillance together in the development cohort to minimize the bias in the spectrum of risk encountered by clinicians and additional bias arising from nodules that had not undergone resection. In contrast, all the nodules in the validation cohorts had an explicit pathology to ensure the accuracy of the validation process. The results of both the development cohort and external validation cohort indicated that the PKU-M model significantly outperformed the SPN models with high accuracy, even in different subgroups stratified in several clinical and radiographic features.

Notably, the suggested population for clinical application is different between our model and the Brock model (13), which is the only multivariable model that involved some MPNs. The Brock model is established in a lung cancer screening population in which the majority (>90%) were benign. Most of these nodules presented typical imaging features that were relatively easy to assess by experienced doctors who did not need assistance by the prediction model. In addition, the probabilities estimated by models derived from populations with a lower malignancy prevalence tended to be lower than those estimated by models derived from populations with a high malignancy prevalence (14). In contrast, nodules in our development cohort had a higher prevalence of cancer (50%) because patients were evaluated by surgeons. Therefore, the PKU-M model has more clinical relevance than the Brock model at the point of surgical evaluation.

CADx systems have been reported to assist in detecting and diagnosing pulmonary nodules on CT scans. Although the false-positive rate remains a concern, these systems have indeed decreased the workload of radiologists, particularly for patients with multiple small nodules. However, few devices for nodule malignancy evaluation have been validated with a satisfactory result except the report from the Google team (30) and a recent retrospective study by Baldwin and colleagues (31). Although the result is promising, it is still far from clinical application. First, such reported deep learning systems were only based on CNN, which lack clinical characteristics such as smoking history, tumor history, age, and other data that are also crucial for diagnosis. Second, the final clinical decisions are made by thoracic surgeons or respiratory physicians, not radiologists; thus, comparisons with these clinicians should be more persuasive. Third, Asian patients are significantly different from Western patients regarding pulmonary nodules in that they have many more GGOs (32). Finally, these two AI products were both developed using the National Lung Cancer Screening Trial (NLST) dataset of screening populations. However, whether they can perform well in patients with incidental pulmonary nodules that were the most common in clinical practice remains unknown. Therefore, we conducted the first prospective comparison among clinicians, our machine learning model, and a CADx system (RX) to diagnose pulmonary nodules, particularly multiple nodules. Because RX was developed by a Chinese company specializing in lung cancer diagnosis through image analysis, using a machine learning algorithm called three-dimensional CNN (33, 34), similar to Google AI and Baldwin’s AI, we considered that RX could represent current AI products to compare with our model. The AUC of the PKU-M model was significantly higher than that of each clinician and CADx system, indicating that the XGBoost model discriminates MPNs better than clinicians and the CADx system. McNemar test showed that our model outperformed the clinicians with an absolute increase of 0.143 in sensitivity and 0.078 in specificity. In addition, the CADx system outperformed the clinicians with an absolute increase of 0.103 in sensitivity but had similar specificity. Considering that the CADx system in this study had only learned approximately 11,000+ cases, following its continuing progress, its accuracy is expected to be further improved and its performance is highly possible to exceed that of the models in the future (30, 31). However, currently, it is not the most validated way for nodule evaluation.

After analysis of SPNs, the discrimination of prior mathematical models improved, suggesting that they are more reliable in patients with SPNs but not MPNs. It is inspiring that we found that the PKU-M model can perform well even in solitary nodules. The AUC was 0.871, comparable with that of our previously established PKU model and higher than those of the Brock model, Mayo model, and VA model. Compared with prior models, the new model included more clinical and radiographic features related to lung cancer, such as tumor markers, lobulation, and pleural retraction sign, and it was developed by a state-of-the-art machine learning algorithm instead of logistic regression; thus, it can assess the solitary nodules more comprehensively.

Although larger multicentric datasets are needed for validation, this result preliminarily demonstrates the PKU-M model’s generalizability in predicting patients with SPN.

Models are developed to assist in clinical diagnosis, indicating that they should be practical. Because of the consideration of the nonlinear relationship between lung cancer and some clinicoradiologic variables, the use of a nomogram is not suitable for the model. Moreover, nomograms require manual calculation according to the scale, which is complicated to some extent. Therefore, we designed a web-calculated version of the model. When clinicians diagnose MPNs, they can only input several clinicoradiologic characteristics, and the software will automatically display the risk of malignancy.

The limitation of this study is that it did not include PET/CT results, which may decrease the accuracy of this model. However, because a large proportion of nodules in this study comprised small (<1 cm) solid or GGO/part-solid nodules, the accuracy of PET/CT is likely unsatisfactory (6, 9, 35, 36); it is therefore not a regular examination. The established model can be more widely used regardless of whether the patient has undergone PET/CT. The other limitation is that the generalizability of this model in Western heterogeneous populations is unclear. Several thoracic CT datasets for Western ancestry individuals are publicly available, with the possibility to use these datasets for validation. However, to complete this study, CT images and detailed clinical information are required to input into our model. To our knowledge, none of the available public datasets meet the need. Therefore, a future study is expected to collect Western cohorts that meet the criteria to validate our model. The third limitation is that patients in this study were patients with high-risk cancer but not a screening cohort. Whether the model could be applied to a widely extended cohort must be verified. However, for the screening cohort, most of the nodules are benign with significant features on CT scans that are easy to diagnose. The true dilemma that puzzles clinicians is presented by suspicious nodules like those included in this study. Finally, although we included as many clinicoradiologic features as possible in our model, it is still a generally comprehensive analysis method. Because of the nodule heterogeneity among patients, this model might not be accurate for an atypical individual. The role of the model is to provide only a relatively reliable reference for clinical judgment; it cannot completely replace clinicians in making the final decision.

In summary, after its development using a novel machine learning algorithm, validated with transnational multicentric cohorts and...
prospectively compared with clinicians and a deep learning system, this first cancer prediction model specifically for MPNs demonstrated solid performance as a convenient reference to help decision-making.

Authors' Disclosures

Y.T. Kim reported personal fees from Johnson and Johnson outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

K. Chen: Conceptualization, resources, formal analysis, funding acquisition, validation, investigation, methodology, writing—original draft, project administration, writing—review and editing. Y. Nie: Conceptualization, data curation, software, formal analysis, investigation, visualization, methodology, writing—original draft, writing—review and editing. S. Park: Conceptualization, resources, data curation, validation, investigation, writing—original draft, project administration. K. Zhang: Data curation, software, formal analysis, visualization, methodology. Y. Zhang: Software, formal analysis, investigation, visualization, methodology. Y. Liu: Software, formal analysis, visualization, methodology. B. Hui: Resources, Data curation, investigation. L. Zhou: Data curation, investigation. X. Wang: Investigation. Q. Qi: Investigation. H. Li: Investigation. G. Kang: Investigation. Y. Huang: Resources, data curation. Y. Chen: Resources, data curation. J. Liu: Resources, data curation. J. Cui: Resources, data curation. M. Li: Resources, data curation. J.K. Park: Resources, data curation. C.H. Kang: Resources, data curation. H. Shen: Data curation. Y. Yang: Resources, data curation. T. Guan: Data curation. Y. Zhang: Resources, data curation. F. Yang: Conceptualization, resources, supervision, validation, project administration, writing—review and editing. Y.T. Kim: Conceptualization, resources, data curation, supervision, validation, methodology, project administration, writing—review and editing. J. Wang: Conceptualization, resources, supervision, project administration, writing—review and editing.

Acknowledgments

We thank American Journal Experts (http://www.aje.cn) for providing language help during the preparation of this manuscript. This work was supported by the National Natural Science Foundation of China (no. 82072566, to K. Chen) and Peking University People’s Hospital Research and Development Funds (RS2019-01, to K. Chen).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received October 11, 2020; revised December 8, 2020; accepted January 21, 2021; published first February 24, 2021.

References


Clinical Cancer Research

Development and Validation of Machine Learning–based Model for the Prediction of Malignancy in Multiple Pulmonary Nodules: Analysis from Multicentric Cohorts

Kezhong Chen, Yuntao Nie, Samina Park, et al.

Clin Cancer Res 2021;27:2255-2265. Published OnlineFirst February 24, 2021.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-20-4007

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2021/01/27/1078-0432.CCR-20-4007.DC1

Cited articles
This article cites 34 articles, 3 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/27/8/2255.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://clincancerres.aacrjournals.org/content/27/8/2255.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.