OXFORD

A pairwise radiomics algorithm–lesion pair relation estimation model for distinguishing multiple primary lung cancer from intrapulmonary metastasis

Ting-Fei Chen[1,§], Lei Yang[1,§], Hai-Bin Chen[2,§], Zhi-Guo Zhou[3], Zhen-Tian Wu[4], Hong-He Luo[1,*], Qiong Li[5,*] and Ying Zhu[6,*]

[1]Department of Thoracic Surgery, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510000, China
[2]Breax Laboratory, PCAB Research Center of Breath and Metabolism, Beijing 100017, China
[3]Reliable Intelligence and Medical Innovation Laboratory (RIMI Lab), Department of Biostatistics & Data Science, University of Kansas Medical Center, and University of Kansas Cancer Center, Kansas City, KS 66160, USA
[4]Center for Information Technology & Statistics, Statistics Section, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510000, China
[5]Department of Radiology, Sun Yat-sen University Cancer Center; State Key Laboratory of Oncology in South China; Collaborative Innovation Center for Cancer Medicine, Guangzhou 510000, China
[6]Department of Radiology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510000, China
*Correspondce: Hong-He Luo, luohongh@mail.sysu.edu.cn; Qiong Li, liqiong@sysucc.org.cn; Ying Zhu, zhuy45@mail.sysu.edu.cn
§Ting-Fei Chen, Lei Yang, and Hai-Bin Chen contributed equally to this work

Abstrac

Background: Distinguishing multiple primary lung cancer (MPLC) from intrapulmonary metastasis (IPM) is critical for their disparate treatment strategy and prognosis. This study aimed to establish a non-invasive model to make the differentiation pre-operatively.

Methods: We retrospectively studied 168 patients with multiple lung cancers (307 pairs of lesions) including 118 cases for modeling and internal validation, and 50 cases for independent external validation. Radiomic features on computed tomography (CT) were extracted to calculate the absolute deviation of paired lesions. Features were then selected by correlation coefficients and random forest classifier 5-fold cross-validation, based on which the lesion pair relation estimation (PRE) model was developed. A major voting strategy was used to decide diagnosis for cases with multiple pairs of lesions. Cases from another institute were included as the external validation set for the PRE model to compete with two experienced clinicians.

Results: Seven radiomic features were selected for the PRE model construction. With major voting strategy, the mean area under receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity of the training versus internal validation versus external validation cohort to distinguish MPLC were 0.983 versus 0.844 versus 0.793, 0.942 versus 0.846 versus 0.760, 0.905 versus 0.728 versus 0.727, and 0.962 versus 0.910 versus 0.769, respectively. AUCs of the two clinicians were 0.619 and 0.580.

Conclusions: The CT radiomic feature-based lesion PRE model is potentially an accurate diagnostic tool for the differentiation of MPLC and IPM, which could help with clinical decision making.

Keywords: multiple primary lung cancer, radiomics, intrapulmonary metastasis

Introduction

As one of the most common cancers worldwide,[1] lung cancer is a threat to people's health and life. As a result of the popularization of high-resolution computed tomography (CT) scanning, more patients with lung cancer are diagnosed at an early stage and are thus able to receive curative surgery.[2–4] Along with this trend is the increasing incidence of those finding more than one lesion in their lungs.

Multiple primary lung cancer (MPLC) was first reported in 1924[5] and has become increasingly common since that time.[6,7] With many unsolved problems for MPLC, one consensus is that whenever possible, surgical resection should be considered, in that survival is excellent and even comparable to solitary lung cancer.[8–11] Yet for intrapulmonary metastatic patients, radical surgery may not be the optimal choice. For this reason, it is of great importance to distinguish MPLC from intrapulmonary metastasis (IPM) so that the appropriate treatment strategy may be applied. The majority of previous studies have primarily relied on pathological analysis for differentiation, emphasizing the need for a diagnostic tool capable of addressing this issue preoperatively. Recent studies[12,13] have revealed imaging features indicative of multiple primaries. The role of positron emission tomography (PET-CT) with standard uptake value (SUV) maximal ratio in differentiating synchronous MPLC from IPM appears to be inconsistent among studies.[13–15] Other emerging techniques such as machine-learning tools have been applied in the diagnosis of pulmonary nodules, yet mostly focusing on single-lesion cases.[16] One study[17] tried to apply an artificial intelligence method for single lung-nodule diagnosis in MPLC without attention to the differentiation between pairs of lesions. The application of radiomics has been diverse in lung cancer and has included histopathology, genetic mutations, and even immune phenotype.[18]

With the application of radiomics, we hypothesize that by comparing the intrinsic features between paired lesions, the

discrimination between MPLC and IPM may be achieved with better accuracy. The difference of imaging features between paired lesions of MPLC would likely be greater than the differences between pairs that include a primary tumor and an intrapulmonary metastasis. We believe that this method could be helpful for pre-operative differential diagnosis of MPLC and the subsequent treatment decision-making.

Materials and methods

Patients and pairing

Patients with two or more suspicious lesions receiving surgery for lung cancer in the First Affiliated Hospital of Sun Yat-sen University (SYSUFH) from October 2014 to October 2020 were enrolled. Their electronic medical records were extracted for further investigation, including clinicopathologic characteristics, radiological data, operation records, molecular genetic testing results, etc. Based on the Martini and Melamed criteria[19] and the 2013 American College of Chest Physicians criteria,[20] patient inclusion criteria for MPLC in this study were as follows.

 (i) The time gap between incidence of two lesions was >4 years (metachronous).

(ii) Histopathological results of lesions were obtained by lung resection, bronchoscopy, biopsy, or aspiration:

  (a) lesions of the same patient were of different pathological types or subtypes, or with *in situ* histology;
  (b) lesions of proven atypical adenomatous hyperplasia were ruled out;
  (c) for those of the same histology other than *in situ* types, molecular genetic testing results showed different driver mutations or any difference in fusion sites.[15]

Cases were excluded for:

  (i) a lack of pre-operative chest CT images;
  (ii) distant metastases;
  (iii) both lesions displayed as the same rare histological type such as neuroendocrine tumors or adenosquamous carcinoma;[21]
  (iv) gene testing showed lesions share the same uncommon driver mutation or fusion of the same breakpoint.[22]

For the last two situations, IPM would be considered. The diagnosis of IPM was based on the algorithm combining histopathology and molecular results,[13,21,23] including patients with similar but uncommon histology or mutations, or at least two mutations in common.

A flowchart of the patient enrollment process is showed in Fig. 1. The same criteria were used for the validation cohort from Sun Yat-sen University Cancer Center (SYSUCC) between January 2020 and March 2021. This retrospective study was approved by the respective Institutional Ethics Commissions of SYSUFA {No.[2020]371} and SYSUCC {No. YB2018-13} with a waiver of informed consent. The information of CT scan protocols are detailed in the online supplementary material.

For further analysis, lesions of each individual case were paired. For example, two lesions of an MPLC patient would make one pair, while four lesions of an MPLC patient would make six pairs (pairs of lesions a-b, a-c, a-d, b-c, b-d, and c-d). Based on these lesion-pairs, the study calculated their differences and managed to distinguish MPLC from IPM.

Histopathological assessment and molecular genetic testing

Resected lesion samples underwent histopathological examination by at least two experienced pathologists, following the 2015 World Health Organization classification guidelines.[24] Immunohistochemistry was employed as needed. Additionally, when deemed necessary and with patient consent, next generation sequencing testing was conducted. Formalin-fixed, paraffin-embedded specimens were utilized for DNA isolation, enabling the detection of specific gene mutations, gene rearrangements, etc. These molecular genetic testing results were obtained to support the diagnosis of MPLC.

Machine learning based model development

With the collected patient dataset, a machine-learning method based on radiomic features[25] was developed for the differential diagnosis of MPLC and IPM. The main steps to develop the model included lesion segmentation, radiomic feature calculation, lesion pair feature deviation calculation, feature selection and lesion pair relation estimation (PRE) model training, and MPLC and IPM classification, as illustrated in Fig. 2.

Lesion segmentation

All lung cancer lesions were manually delineated using open-source software ITK-Snap (Penn Image Computing and Science Laboratory, Philadelphia, PA, USA)[26] by two experienced chest radiologists independently, both blinded to the histologic diagnosis. Firstly, the minimal bounding box was calculated in three dimensions based on the delineated contours. The minimal region of interes was extracted from the normalized image by clipping the image intensities in the range [−1024, 1024] and then mapping them to the range [−1, 1]. Each lesion was processed via this method to get cropped region of interest images.

Radiomic feature calculation

The feature vector F with 107 radiomic features ($F = [F_1, \ldots, F_i, \ldots, F_{107}]$) was calculated for each lesion, which included 18 first-order features, 14 shape-based features, 24 gray level co-occurrence matrix (GLCM) features, 16 gray level run length matrix (GLRLM) features, 16 gray level size zone matrix (GLSZM) features, 5 neighboring gray-tone difference matrix (NGTDM) features, and 14 gray level dependence matrix (GLDM) features.

Feature deviation calculation of lesion pairs

MPLC lesions, according to the definition itself, originated from multiple sources independent of each other, while IPM lesions share the same origin. Thus, it is reasonable to hypothesize that MPLC lesions could be different in intensity, properties, and material construction, etc. On the other hand, IPM lesions could be more closely related in terms of radiomic features. Based on this hypothesis, we calculated the absolute deviation (AD) of feature vectors $F^{AD} = [F_1^{AD}, \ldots, F_i^{AD}, \ldots, F_{107}^{AD}]$ for each lesion pair of each patient, where $F_i^{AD} = F_i^a - F_i^b$, F_i^a, and F_i^b represent the i_{th} radiomic feature of lesions a and b from the same patient. Thus, F_i^{AD} was defined as the representative similarity measure of the lesion pair.

Feature selection

Based on the similarity feature vectors calculated above, a PRE model was built via the random forest classifier[27] for pairwise lesions. In this study, 70% of patients were randomly split as
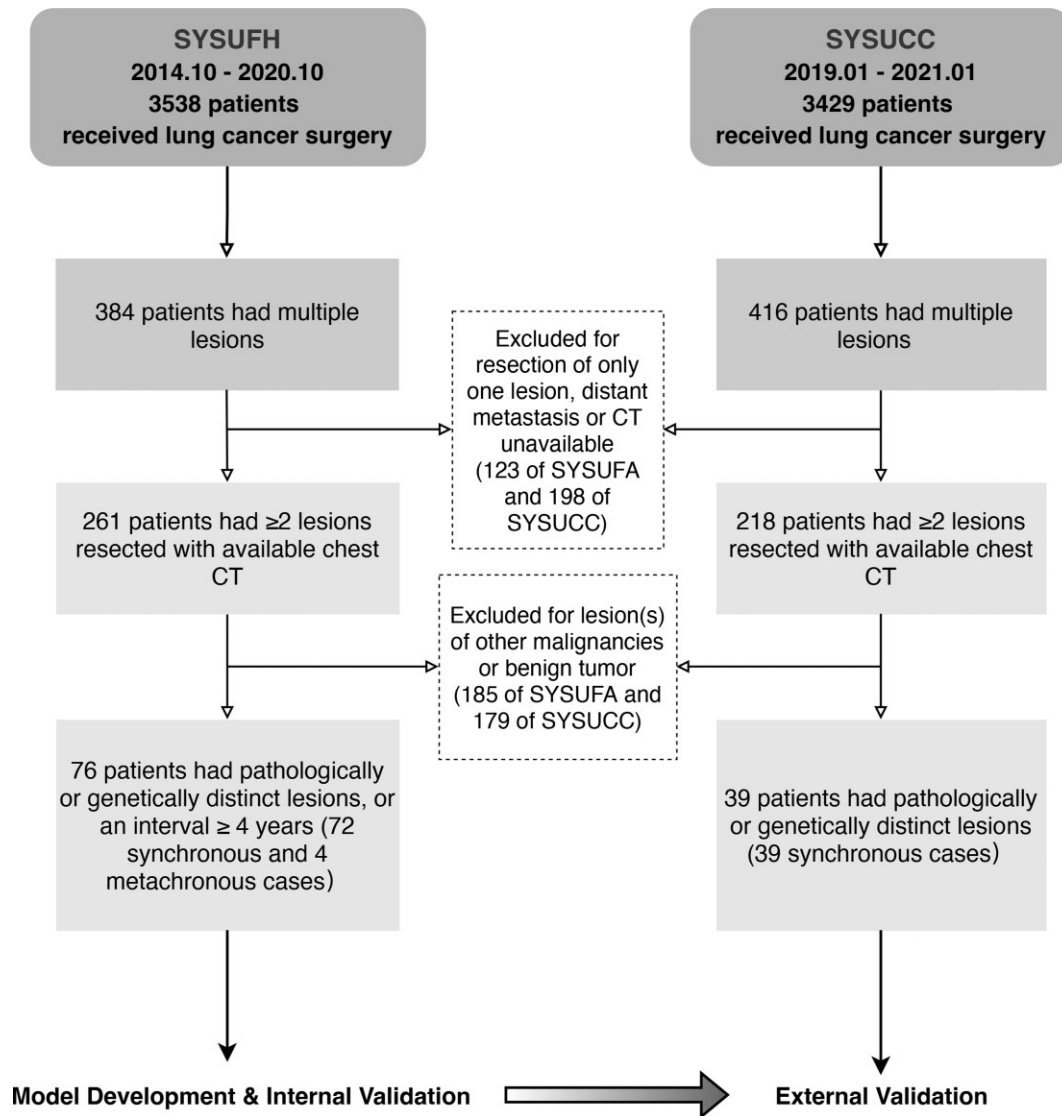
Figure 1. Flowchart of the enrollment process of MPLC.

training dataset (SYSUFH training cohort) and the remaining 30% as testing dataset for model evaluation (SYSUFH internal validation). Considering the small patient cohort, 107 features for training could be excessive and decrease the accuracy and stability of the model. To alleviate the over-fitting problem, feature selection was achieved by a two-step procedure, including the removal of redundant features of high correlation coefficients and those of lower variances. That is, for a feature pair, the correlation coefficient was larger than the threshold T_{cc} (empirically set as 0.7). Then 5-fold cross-validation and random forest model fitting methods were applied to obtain the optimal feature number and evaluate the importance of features, respectively. Each time, one feature was added according to its ranking in descending order to train the model on the intra-training dataset (4 of 5 folds), and the receiver operating characteristic (ROC) curve was calculated on the intra-validation dataset (1 of 5 folds). The area under the ROC curve (AUC)[28] was monitored and validated until the optimal model was achieved and its corresponding feature number obtained. In this way, the top k important features were selected with the best AUC in the 5-fold cross-validation.

PRE model construction and MPLC and IPM classification

On the training dataset with selected features, a new model based on random forest was developed for PRE to diagnose MPLC. Considering individuals with two or more pairs of lesions, a major voting strategy as shown in Fig. 2F was adopted where the predicted probability was the mean probability of majorities for such cases. For example, for a patient with three lung lesions (noted as a, b, and c), pair a with b and pair a with c were predicted similar while pair b with c different. IPM would be considered for this case.

Model evaluation

To evaluate the accuracy and robustness of the proposed method, the final PRE model training steps were repeated 100 times with datasets randomly split as training and testing with the ratio 7 : 3. AUC, accuracy sensitivity specificity negative predictive value (NPV), and positive predictive value (PPV) were used to qualitatively evaluate the model's performance in diagnosing MPLC.

For the external validation set of SYSUCC, two experienced clinicians (one experienced chest radiologist and one senior thoracic
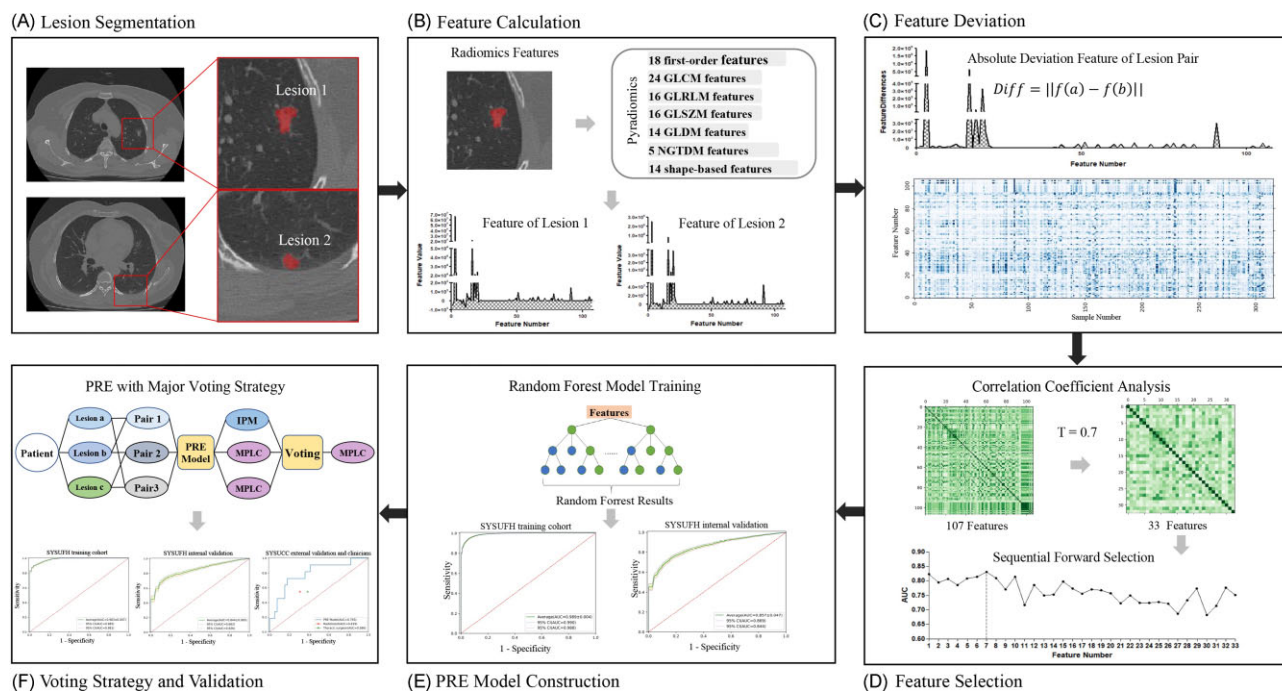
Figure 2. (**A**–**F**)The main steps to develop the pairwise machine-learning algorithm for MPLC and IPM differentiation.

surgeon, both with ≥5 years of experience) were given these images to make diagnostic decisions blinded of their groups. Their performance was then compared with that of the PRE model.

Statistical analysis

Analysis of clinicopathological statistics was performed with the chi-square test (for categorical variables) and Student's *t* test (for continuous variables of normal distribution). Statistical analysis was carried out with R version 3.6.2 (R Foundation for Statistical Computing, Vienna, Austria). $P < 0.05$ was considered statistically significant.

To analyze feature differences between lesion pairs, the Kolmogorov–Smirnov test was applied for normal distribution testing ($P > 0.05$). The Wilcoxon rank-sum test and unpaired *t*-test were used for statistical significance analysis ($P < 0.05$) on datasets of abnormal and normal distribution, respectively. The radiomic feature calculation was achieved based on an open-source radiomics package (Pyradiomics 3.0.1).[25] Model construction and evaluation were conducted on the platform Python 3.7 (Python Software Foundation) via Pycharm 2020 (JetBrains, Czech Republic).

Results
Demographic and histopathologic characteristics

Clinicopathological characteristics of patients from SYSUFH are shown in Table 1. In this study, we identified 76 MPLC patients with 137 lesion pairs and 42 IPM patients with 93 lesion pairs. Among them, if an individual harbored three lesions, the patient would be considered as having three pairs of lesions for pairwise comparison. There was no statistical difference between the two groups in terms of age, sex, smoking status, family history of primary lung cancer, performance status, tumor markers, and main pathologic type. The clinicopathological characteristics of the SYSUCC validation cohort are shown in supplementary Table S1, see online

supplementary material. Chest CT images of representative cases of MPLC and IPM are given in Fig. 3.

Radiomic feature selection

CT-derived radiomic features were acquired from the SYSUFH cohort. There were 33 features left after performing the correlation coefficient-based redundant feature reduction procedure,[29] which were put into the next feature selection procedure as illustrated in supplementary Fig. S1a, see online supplementary material. As the number of selected features increased from 1 to 7, performance of the model improved correspondingly. However, model performance deteriorated when more features were included, which could imply overfitting. Hence the top seven features (as given in supplementary Table S2, see online supplementary material) were selected for the final PRE model construction and evaluation. The correlation coefficients among these seven selected features were all <0.52 (supplementary Fig. S1b), suggesting weak correlations of these features. Besides, feature distribution between IPM and MPLC cases was distinct, as shown in supplementary Fig. S1c. Further analysis of the selected feature deviation (F^{AD}) between MPLC and IPM lesion pairs is detailed in supplementary Fig. S2, see online supplementary material.

Model performance

Performance of the PRE model for distinguishing MPLC from IPM on the SYSUFH training and internal validation datasets is shown in Fig. 4A. For the diagnosis of MPLC, the mean AUC, accuracy, sensitivity, specificity, NPV, and PPV on the training dataset were 0.989, 0.947, 0.946, 0.948, 0.947, and 0.949, respectively (Table 2, Fig. 4A i). The corresponding metrics on the internal validation dataset were 0.857, 0.794, 0.758, 0.850, 0.677, and 0.906, respectively (Table 2, Fig. 4Aa ii). With the application of a major voting strategy, performance of the PRE model for the diagnosis of MPLC is illustrated in Fig. 4B. The mean AUC, accuracy, sensitivity, specificity, NPV, and PPV of this model were 0.983, 0.942, 0.905, 0.962, 0.950, and 0.934 on the SYSUFH training dataset (Table 2, Fig. 4B i)

Table 1. Clinicopathological characteristics of patients with MPLC and IPM.

Patient characteristic	MPLC (n = 76)	IPM (n = 42)	P Value
Number of lesions (pairs)	180 (137)	108 (93)	NA
Mean age (range, years)	60 (32–79)	62 (30–86)	0.531
Sex, n (%)			0.415
Female	43 (56.58%)	27 (64.29%)	
Male	33 (43.42%)	15 (35.71%)	
Smoking status, n (%)			0.369
Non-smoker	50 (65.79%)	31 (73.81%)	
Current/former smoker	26 (34.21%)	11 (26.19%)	
Family history of primary lung cancer, n (%)	11 (14.47%)	4 (9.52%)	0.440
Performance status, n (%)			0.082
0	63 (82.89%)	29 (69.05%)	
1	13 (17.11%)	13 (30.95%)	
Tumour markers, n (%)			0.295
Abnormal	47 (61.84%)	30 (71.43%)	
Normal	29 (38.16%)	12 (28.57%)	
Distribution of lesions, n (%)			0.019*
Same lobe	25 (32.89%)	5 (11.9%)	
Ipsilateral	22 (28.95%)	21 (50%)	
Contralateral	29 (38.16%)	16 (38.1%)	
Synchronicity, n (%)			0.326
Synchronous	72 (94.74%)	42 (100%)	
Metachronous	4 (5.26%)	0 (0%)	
Same histologic subtype of the lesion pair	9 (11.84%)	42 (100%)	0.000*
Staged operation, n (%)	27 (35.53%)	NA	NA
Range of resection, n (%)			
Lobectomy	23 (30.26%)	NA	NA
Bilobectomy	11 (14.47%)	NA	NA
Lobectomy + sublobectomy(ies)	35 (46.05%)	NA	NA
Multiple sublobectomies	7 (9.21%)	NA	NA

NA, Not applicable; *P < 0.05 was considered as statistically significant.
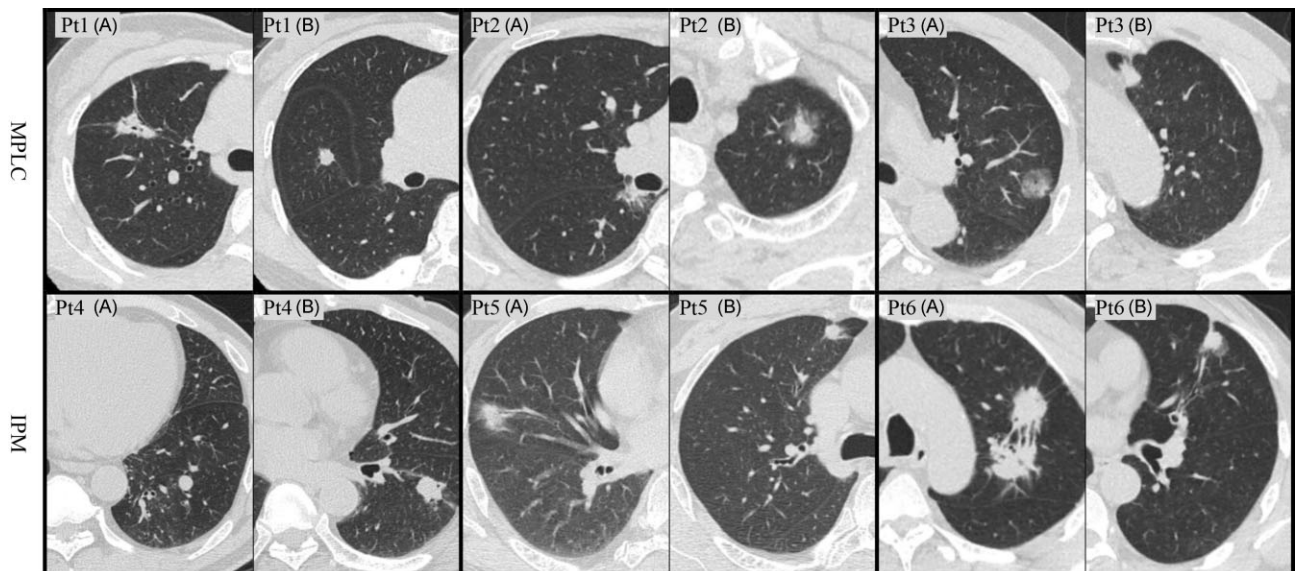

Figure 3. Chest CT images of representative cases of MPLC and IPM. Patient 1 (Pt1) showed two solid lesions located in the right upper lobe (Pt1 A, APA) and the right middle lobe (Pt1 B, APA) respectively. Patient 2 (Pt2) had two mixed GGNs situated contralaterally in the right lower lobe (Pt2 A, APA) and the left upper lobe (Pt2 B, APA). A mixed GGN (Pt3 A, LPA) and a solid nodule (Pt3 B, APA) were found in patient 3 (Pt3), both located in the left upper lobe. These three were confirmed MPLC cases while patients 4–6 (Pt4–Pt6) were IPM cases. Pt4 and Pt6 both showed the main lesion (Pt4 B and Pt6 A) located in the same lobe as the metastatic tumor (Pt4 A and Pt6 B). For Pt5, the main lesion was in the right middle lobe (Pt5 A) while the metastatic lesion was in the right upper lobe (Pt5 B). Based on preoperative chest CT images, it was difficult to decide whether these lesions were MPLC or IPM. With the radiomic PRE model built by our team, diagnostic accuracy was improved significantly and hence clinical decisions made with stronger evidence. APA, Acinar-predominant adenocarcinoma; GGN, ground-glass nodule; LPA, lepidic-predominant adenocarcinoma.

(A) ROCs of the PRE model trained and tested with lesion pairs from SYSUFH.
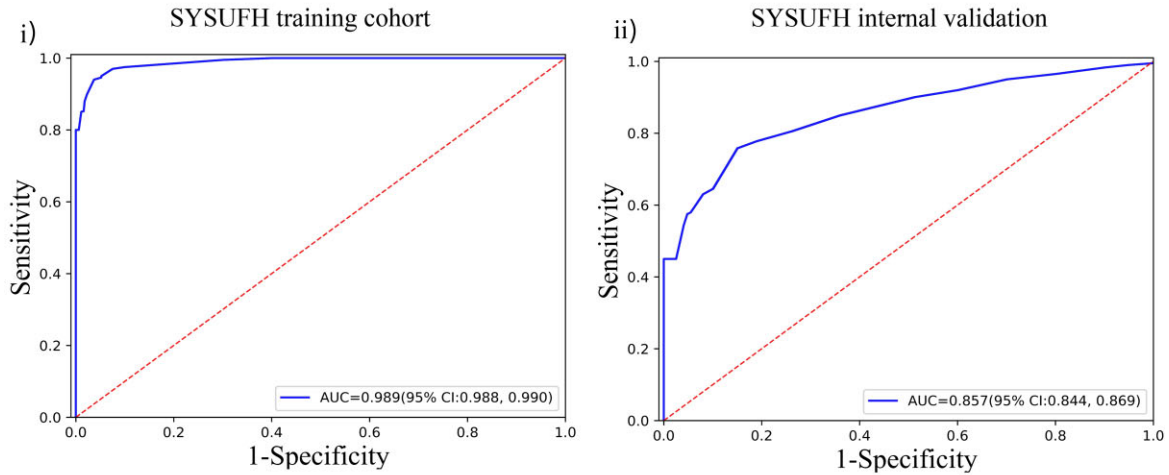
(B) ROCs of the PRE model with voting strategy (case-based prediction).
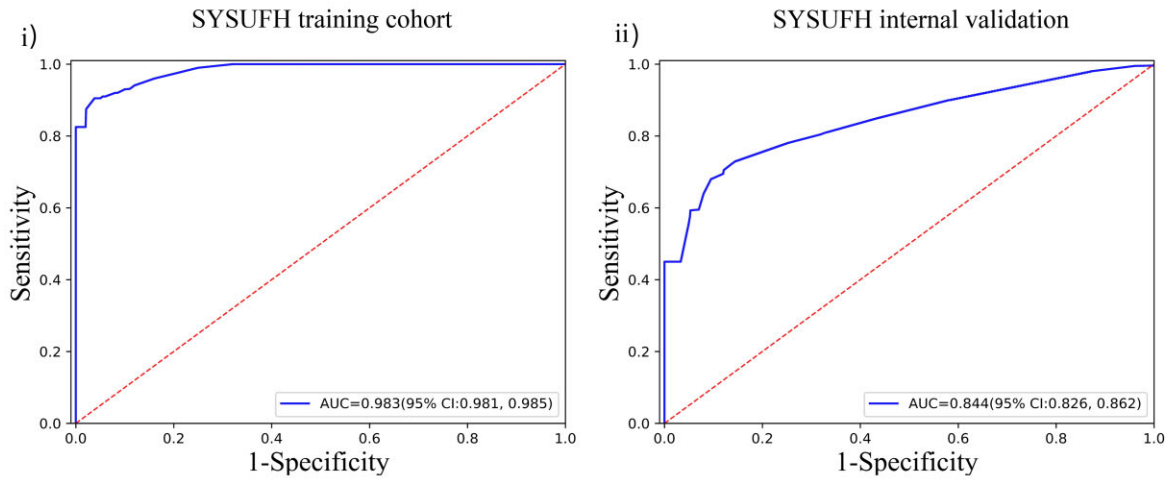
(C) ROCs of the PRE model on SYSUCC external validation cohort and AUCs of clinicians.
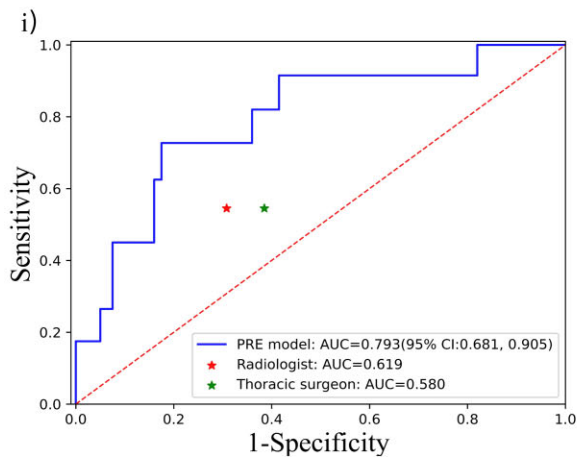

Figure 4. The quantitative results of the PRE model for MPLC and IPM differentiation. (**A**, i and ii) ROC curves of PRE model on SYSUFH training and internal validation datasets. (**B**, i and ii) ROC curves of PRE model for MPLC and IPM differentiation (with major voting strategy) on SYSUFH training and internal validation datasets. (**C**, i) ROC curve of PRE model (with major voting strategy) on SYSUCC external validation dataset and AUCs of clinicians.

Table 2. Performance evaluation of the PRE model and of clinicians.

	Accuracy	Sensitivity	Specificity	NPV	PPV	AUC
PRE model trianed and tested with lesion pairs from SYSUFH						
SYSUFH training cohort	0.947 (0.944–0.950)	0.946 (0.941–0.952)	0.948 (0.941–0.954)	0.947 (0.943–0.952)	0.949 (0.943–0.954)	0.989 (0.988–0.990)
SYSUFH internal validation	0.794 (0.784–0.804)	0.758 (0.738–0.778)	0.850 (0.829–0.870)	0.677 (0.656–0.696)	0.906 (0.895–0.916)	0.8597 (0.844–0.869)
PRE model with major voting strategy (case-based prediction)						
SYSUFH training cohort	0.942 (0.938–0.946)	0.905 (0.895–0.915)	0.962 (0.956–0.968)	0.950 (0.945–0.954)	0.934 (0.924–0.944)	0.983 (0.981- 0.985)
SYSUFH internal validation	0.846 (0.836–0.873)	0.728 (0.705–0.751)	0.910 (0.894–0.925)	0.861 (0.850–0.873)	0.828 (0.805–0.852)	0.844 (0.826–0.862)
Performance of the PRE model on cases from SYSUCC						
SYSUCC external validation	0.760 (0.693–0.811)	0.727 (0.634–0.775)	0.769 (0.710–0.804)	0.909 (0.869–0.941)	0.471 (0.417–0.521)	0.793 (0.758–0.835)
Performance of clinicians in differentiating MPLC and IPM cases from SYSUCC						
Radiologist	0.660	0.545	0.692	0.844	0.333	0.619
Thoracic surgeon	0.600	0.545	0.615	0.828	0.286	0.580

and 0.844, 0.846, 0.728, 0.910, 0.861, and 0.828 on the SYSUFH internal validation cohort (Table 2 and Fig. 4Bb ii). Despite the slight decline in performance on the testing dataset, it remains sensible to believe that the PRE model could be of great value to clinical practice for the differential diagnosis of MPLC.

Validation results of the SYSUCC cohort are shown in Table 2 and Fig. 4C i. The mean AUC, accuracy, sensitivity, specificity, NPV, and PPV of the established PRE model with major voting strategy were 0.793, 0.760, 0.727, 0.769, 0.909, and 0.471, respectively, to diagnose MPLC. Performance of the model in the SYSUCC external validation cohort was compared with that of experts, as illustrated in Fig. 4C i. AUCs of the chest radiologist and thoracic surgeon were 0.619 and 0.580 respectively, significantly lower than that of the model (0.793). These validation results suggest promising performance of the proposed model and its great value for clinical practice.

Discussion

In this study, we have successfully established a novel non-invasive diagnostic algorithm to differentiate MPLC from IPM. Our model has demonstrated promising diagnostic performance for MPLC with a mean AUC of 0.844 and high accuracy, sensitivity, specificity, NPV, and PPV, outperforming clinical diagnoses in the external validation cohort.

The escalation in CT screening's accuracy and popularity has led to increased detection of multiple pulmonary lesions. The incidence of MPLC reported ranged from 1.1% to 6.9%,[30–32] consistent with the findings in this study. The significance of preoperative diagnosis of these lesions is ever-growing, mainly to distinguish between MPLC and IPM. Since its proposal in 1975, the Martini and Melamed criteria[19] have garnered wide acceptance. Recent scientific advancements have further harnessed molecular genetic characteristics[20,22,33–39] to develop algorithms that consider comprehensive information.[13,21,23,40] These advances aim to aid decision-making in the differentiation between IPM and MPLC. However, such differentiation relied heavily on postoperative pathological analysis, highlighting an unmet need for preoperative diagnostic tools. Previously, some diagnostic indicators have been identified, such as multi-ground-glass nodule (GGN) and solid-GGN as potential MPLC flags[12] and multiple pure solid nodules suggestive of IPM.[12,41] CT imaging features including subsolid consistency, spiculated contour, size difference, and small-

est lesion being pure solid were identified as potential discriminators between MPLC and IPM.[15] Another study successfully developed an algorithm to discriminate between the two, achieving an AUC of 0.833.[10] However, the existing studies primarily provided somewhat subjective results without the development of predictive models. In contrast, the PRE model attained a mean AUC of 0.844 and demonstrated strong performance in the external validation cohort, with an AUC of 0.793, showcasing its stability and reproducibility.

Artificial intelligence has increasingly been utilized in this domain to predict malignancy[42,43] or invasiveness[16,44,45] of pulmonary nodules. But most existing models narrowly focus on singular lesions, ignoring the holistic view of a patient. Radiomics methods were used to predict histopathological results in previous studies, indicating the potential relevance of radiological features in the pathology of lung cancer.[46–50] Our work hypothesizes that variations among MPLC lesions could be more significant than those between a primary and its metastatic lesion. Thus, we aimed to craft a machine-learning model that observes each patient as a cohesive entity. What sets this method apart and makes it unique and novel is its utilization of the concept of lesion pairs. Therefore, the radiomic features selected (details shown in supplementary Table S2) through the comparison of lesion pairs proved effective in distinguishing MPLC from IPM cases. The resulting PRE model, grounded in this perspective, has demonstrated superior efficacy compared to existing methodologies and hence showcased promising potential for clinical use. The distinction between MPLC and IPM can pose a significant challenge, even for seasoned clinicians, as our results have indicated. By leveraging this model, clinical practitioners can gain increased confidence and enhance the accuracy of their MPLC diagnoses prior to surgical interventions.

There are a few limitations in this study. Even with MPLC's increasing detection rate, our sample size remained limited compared to solitary lung cancer. But, given our focus on lesion pairs and comparison to extant MPLC research, our sample is reasonably sufficient, with an external validation cohort from another institution adding credibility to the model. While we did not venture into various machine-learning methods, considering that our primary aim was not a comparative analysis of these techniques, future studies might explore this aspect further. Besides, the construction of the PRE was exclusively image-centric, sidelining potentially influential factors like age or smoking history. Yet, its

commendable AUC is a testament to its efficacy, even surpassing experienced clinicians in the external validation cohort. The efficacy of this model remains unknown in the real world and multicenter studies will be indispensable.

In summary, our novel machine-learning model offers a promising preoperative tool for distinguishing IPM from MPLC. The PRE model's excellent performance has the potential to significantly impact clinical practice.

Acknowledgements

Author contributions

T.-F.C.: Conceptualization, investigation, data curation, and writing original draft. L.Y.: Conceptualization, investigation, resources, data curation, and writing—review and editing. H.-B.C.: Methodology, software, validation, formal analysis, and visualization. Z.G.Z.: Methodology, and data curation. Z.-T.W.: Resources, and data curation. H.-H.L.: Resources, funding acquisition, supervision, and project administration. Q.L.: Resources, data curation, and supervision. Y.Z.: Conceptualization, methodology, formal analysis, data curation, writing—review and editing, and project administration.

Supplementary data

Supplementary data is available at *PCMEDI* Journal online.

Conflict of interest

The authors declare that they have no conflicts of interests.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020;**70**:7–30. doi:10.3322/caac.21590.

2. Han SS, Rivera GA, Tammemägi MC, *et al*. Risk stratification for second primary lung cancer. *J Clin Oncol* 2017;**35**:2893–9. doi:10.1200/JCO.2017.72.4203.

3. Vazquez M, Carter D, Brambilla E, *et al*. Solitary and multiple resected adenocarcinomas after CT screening for lung cancer: histopathologic features and their prognostic implications. *Lung Cancer* 2009;**64**:148–54. doi: 10.1016/j.lungcan.2008.08.009.

4. Tanvetyanon T, Boyle TA. Clinical implications of genetic heterogeneity in multifocal pulmonary adenocarcinomas. *J Thorac Dis* 2016;**8**:E1734–8. doi:10.21037/jtd.2016.12.06.

5. Beyreuther H. Multiplicität von Carcinomen bei einem Fall von sog. "Schneeberger" Lungenkrebs mit Tuberkulose. *Virchows Archiv für Pathologische Anatomie und Physiologie und für Klinische Medizin* 1924;**250**:230–43. doi:10.1007/bf01891568.

6. Adebonojo SA, Moritz DM, Danby CA. The results of modern surgical therapy for multiple primary lung cancers. *Chest* 1997;**112**:693–701. doi:10.1378/chest.112.3.693.

7. Finley DJ, Yoshizawa A, Travis W, *et al*. Predictors of outcomes after surgical treatment of synchronous primary lung cancers. *J Thorac Oncol* 2010;**5**:197–205. doi:10.1097/JTO.0b013e3181c814c5.

8. Fabian T, Bryant AS, Mouhlas AL, *et al*. Survival after resection of synchronous non-small cell lung cancer. *J Thorac Cardiovasc Surg* 2011;**142**:547–53. doi:10.1016/j.jtcvs.2011.03.035.

9. Tie H, Luo J, Shi R, *et al*. Characteristics and prognosis of synchronous multiple primary lung cancer after surgical treatment: A systematic review and meta-analysis of current evidence. *Cancer Med* 2021;**10**:507–20. doi:10.1002/cam4.3614.

10. Jiang L, He J, Shi X, *et al*. Prognosis of synchronous and metachronous multiple primary lung cancers: systematic review and meta-analysis. *Lung Cancer* 2015;**87**:303–10. doi:10.1016/j.lungcan.2014.12.013.

11. Castiglioni M, Louie BE, Wilshire CL, *et al*. Patients with multiple nodules and a dominant lung adenocarcinoma have similar outcomes and survival compared with patients who have a solitary adenocarcinoma. *Interact Cardiovasc Thorac Surg* 2015;**20**:229–35. doi:10.1093/icvts/ivu366.

12. Zhang Y, Li G, Li Y, *et al*. Imaging features suggestive of multiple primary lung adenocarcinomas. *Ann Surg Oncol* 2020;**27**:2061–70. doi:10.1245/s10434-019-08109-w.

13. Suh YJ, Lee HJ, Sung P, *et al*. A novel algorithm to differentiate between multiple primary lung cancers and intrapulmonary metastasis in multiple lung cancers with multiple pulmonary sites of involvement. *J Thorac Oncol* 2020;**15**:203–15. doi:10.1016/j.jtho.2019.09.221.

14. Liu Y, Tang Y, Xue Z, *et al*. SUVmax ratio on PET/CT may differentiate between lung metastases and synchronous multiple primary lung cancer. *Acad Radiol* 2020;**27**:618–23. doi:10.1016/j.acra.2019.07.001.

15. Araujo-Filho JAB, Chang J, Mayoral M, *et al*. Are there imaging characteristics that can distinguish separate primary lung carcinomas from intrapulmonary metastases using next-generation sequencing as a gold standard?. *Lung Cancer* 2021;**153**:158–64. doi:10.1016/j.lungcan.2021.01.019.

16. Zhao W, Yang J, Yang Y, *et al*. 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. *Cancer Res* 2018;**78**:6881–9. doi:10.1158/0008-5472.CAN-18-0696.

17. Li X, Hu B, Li H, *et al*. Application of artificial intelligence in the diagnosis of multiple primary lung cancer. *Thoracic Cancer* 2019;**10**:2168–74. doi:10.1111/1759-7714.13185.

18. Chen M, Copley SJ, Viola P, *et al*. Radiomics and artificial intelligence for precision medicine in lung cancer treatment. *Semin Cancer Biol* 2023;**93**:97–113, doi: 10.1016/j.semcancer.2023.05.004.

19. Martini N, Melamed MR. Multiple primary lung cancers. *J Thorac Cardiovasc Surg* 1975;**70**:606–12.

20. Kozower BD, Larner JM, Detterbeck FC, *et al*. Special treatment issues in non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013;**143**:e369S–99S. doi:10.1378/chest.12-2362.

21. Shao J, Wang C, Li J, *et al*. A comprehensive algorithm to distinguish between MPLC and IPM in multiple lung tumors patients. *Ann Transl Med* 2020;**8**:1137, doi:10.21037/atm-20-5505.

22. Murphy SJ, Harris FR, Kosari F, *et al*. Using genomics to differentiate multiple primaries from metastatic lung cancer. *J Thorac Oncol* 2019;**14**:1567–82. doi:10.1016/j.jtho.2019.05.008.

23. Mansuet-Lupo A, Barritault M, Alifano M, *et al*. Proposal for a combined histomolecular algorithm to distinguish multiple primary adenocarcinomas from intrapulmonary metastasis in patients with multiple lung tumors. *J Thorac Oncol* 2019;**14**:844–56. doi:10.1016/j.jtho.2019.01.017.

24. Travis WD, Brambilla E, Nicholson AG, *et al*. The 2015 World Health Organization classification of lung tumors:

Impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol* 2015;**10**:1243–60. doi:10.1097/JTO.0000000000000630.

25 Van Griethuysen JJM, Fedorov A, Parmar C, *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017;**77**:e104.doi:10.1158/0008-5472.CAN-17-0339.

26 Yushkevich PA, Piven J, Hazlett HC, *et al.* User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* 2006;**31**:1116–28. doi:10.1016/j.neuroimage.2006.01.015.

27 Breiman LRF. Random forests. *Machine Learning* 2001;**45**:5–32. doi:10.1023/A:1010933404324.

28 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36. doi:10.1148/radiology.143.1.7063747.

29 Jiang SY, Wang LX. Efficient feature selection based on correlation measure between continuous and discrete features. *Information Processing Letters* 2016;**116**:203–15. doi:10.1016/j.ipl.2015.07.005.

30 Trousse D, Barlesi F, Loundou A, *et al.* Synchronous multiple primary lung cancer: an increasing clinical occurrence requiring multidisciplinary management. *J Thorac Cardiovasc Surg* 2007;**133**:1193–200. doi:10.1016/j.jtcvs.2007.01.012.

31 Yu YC, Hsu PK, Yeh YC, *et al.* Surgical results of synchronous multiple primary lung cancers: similar to the stage-matched solitary primary lung cancers?. *Ann Thorac Surg* 2013;**96**:1966–74. doi:10.1016/j.athoracsur.2013.04.142.

32 Huang YQ, Liang CH, He L, *et al.* Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *J Clin Oncol* 2016;**34**:2157–64. doi:10.1200/JCO.2015.65.9128.

33 Liu Y, Zhang J, Li L, *et al.* Genomic heterogeneity of multiple synchronous lung cancer. *Nat Commun* 2016;**7**:13200. doi:10.1038/ncomms13200.

34 Wu C, Zhao C, Yang Y, *et al.* High discrepancy of driver mutations in patients with NSCLC and synchronous multiple lung ground-glass nodules. *J Thorac Oncol* 2015;**10**:778–83. doi:10.1097/JTO.0000000000000487.

35 Vignot S, Frampton GM, Soria JC, *et al.* Next-generation sequencing reveals high concordance of recurrent somatic alterations between primary tumor and metastases from patients with non-small-cell lung cancer. *J Clin Oncol* 2013;**31**:2167–72. doi:10.1200/JCO.2012.47.7737.

36 Eguren-Santamaria I, Sanchez-Bayona R, Patino-Garcia A, *et al.* Targeted DNA sequencing for assessing clonality in multiple lung tumors: A new approach to an old dilemma. *Lung Cancer* 2018;**122**:120–3. doi:10.1016/j.lungcan.2018.05.029.

37 Patel SB, Kadi W, Walts AE, *et al.* Next-generation sequencing: A novel approach to distinguish multifocal primary lung adenocarcinomas from intrapulmonary metastases. *J Mol Diagn* 2017;**19**:870–80. doi:10.1016/j.jmoldx.2017.07.006.

38 Murphy SJ, Aubry MC, Harris FR, *et al.* Identification of independent primary tumors and intrapulmonary metastases using DNA rearrangements in non-small-cell lung cancer. *J Clin Oncol* 2014;**32**:4050–8. doi:10.1200/JCO.2014.56.7644.

39 Ma P, Fu Y, Cai MC, *et al.* Simultaneous evolutionary expansion and constraint of genomic heterogeneity in multifocal lung cancer. *Nat Commun* 2017;**8**:823. doi:10.1038/s41467-017-00963-0.

40 Girard N, Deshpande C,Lau C, *et al.* Comprehensive histologic assessment helps to differentiate multiple lung primary nonsmall cell carcinomas from metastases. *Am J Surg Pathol* 2009;**33**:1752–64. doi:10.1097/PAS.0b013e3181b8cf03.

41 Hattori A, Matsunaga T, Takamochi K, *et al.* Radiological classification of multiple lung cancers and the prognostic impact based on the presence of a ground glass opacity component on thin-section computed tomography. *Lung Cancer* 2017;**113**:7–13. doi:10.1016/j.lungcan.2017.09.001.

42 Matsuki Y, Nakamura K, Watanabe H, *et al.* Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis. *AJR Am J Roentgenol* 2002;**178**:657–63. doi:10.2214/ajr.178.3.1780657.

43 Heuvelmans MA, van Ooijen PMA, Ather S, *et al.* Lung cancer prediction by deep learning to identify benign lung nodules. *Lung Cancer* 2021;**154**:1–4. doi:10.1016/j.lungcan.2021.01.027.

44 Ather S, Kadir T, Gleeson F. Artificial intelligence and radiomics in pulmonary nodule management: current status and future applications. *Clin Radiol* 2020;**75**:13–9. doi:10.1016/j.crad.2019.04.017.

45 Dzobo K, Adotey S, Thomford NE, *et al.* Integrating artificial and human intelligence: A partnership for responsible innovation in biomedical engineering and medicine. *OMICS* 2020;**24**:247–63. doi:10.1089/omi.2019.0038.

46 Ferreira Junior JR, Koenigkam-Santos M, Cipriano FEG, *et al.* Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. *Comput Methods Programs Biomed* 2018;**159**:23–30. doi:10.1016/j.cmpb.2018.02.015.

47 Qi J, Deng Z, Sun G, *et al.* One-step algorithm for fast-track localization and multi-category classification of histological subtypes in lung cancer. *Eur J Radiol* 2022;**154**:110443. doi:10.1016/j.ejrad.2022.110443.

48 Wu W, Parmar C, Grossmann P, *et al.* Exploratory study to identify radiomics classifiers for lung cancer histology. *Front Oncol* 2016;**6**:71. doi:10.3389/fonc.2016.00071.

49 Linning E, Lin L, Li L, *et al.* Radiomics for classifying histological subtypes of lung cancer based on multiphasic contrast-enhanced computed tomography. *J Comput Assist Tomogr* 2019;**43**:300–6. doi:10.1097/RCT.0000000000000836.

50 Wu G, Woodruff HC, Shen J, *et al.* Diagnosis of invasive lung adenocarcinoma based on chest CT radiomic features of part-solid pulmonary nodules: A multicenter study. *Radiology* 2020;**297**:451–8. doi:10.1148/radiol.2020192431.