A pairwise radiomics algorithm - lesion pair relation estimation (PRE) model for distinguishing multiple primary lung cancer (MPLC) from intrapulmonary metastasis (IPM)

Ting-Fei Chen  
The First Affiliated Hospital of Sun Yat-sen University

Lei Yang  
The First Affiliated Hospital of Sun Yat-sen University

Hai-Bin Chen  
PCAB Research Center of Breath and Metabolism

Hao-Shuai Yang  
The First Affiliated Hospital of Sun Yat-sen University

Zhen-Tian Wu  
The First Affiliated Hospital of Sun Yat-sen University

Hong-He Luo  
The First Affiliated Hospital of Sun Yat-sen University

Qiong Li  
Sun Yat-sen University Cancer Center

Ying Zhu  
zhuy45@mail.sysu.edu.cn  
The First Affiliated Hospital of Sun Yat-sen University

Research Article

Keywords: multiple primary lung cancer, radiomics, X-ray computed, intrapulmonary metastasis

Posted Date: November 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2216255/v1

License: ☒  This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

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 five-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 vs. internal validation vs. external validation cohort to distinguish MPLC were 0.983 vs. 0.844 vs. 0.793, 0.942 vs. 0.846 vs. 0.760, 0.905 vs. 0.728 vs. 0.727, and 0.962 vs. 0.910 vs. 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.

Background

As one of the most common cancer 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. Along with this trend is the increasing incidence of those finding more than one lesions in their lungs.
Multiple primary lung cancer (MPLC) was first reported in 1924[2] and has shown to be increasingly common since that time[3, 4]. 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[5–8]. 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 could be applied.

Preoperatively, the diagnosis of MPLC depends on imaging examinations. Recent studies[9, 10] 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 seemed inconsistent among studies[10–12]. Other emerging techniques such as machine learning tools have been applied in the diagnosis of pulmonary nodules, yet mostly focusing on single-lesion cases[13]. One study[14] tried to apply an artificial intelligence method for single lung nodule diagnosis in the condition of MPLC without attention to the differentiation between pairs of lesions.

With the application of radiomics, we hypothesize that by comparing the intrinsic features between paired lesions, the discrimination between MPLC and IPM could 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 preoperative differential diagnosis of MPLC and the subsequent treatment decision-making.

**Methods**

**Patients**

Patients with two or more suspicious lesions receiving surgery for lung cancer in the First Affiliated Hospital of Sun Yat-sen University (SYSUFH) from Oct. 2014 to Oct. 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[15] and the 2013 ACCP (American College of Chest Physicians) criteria[16], patient inclusion criteria for MPLC in this study were as follows:

1. The time gap between incidence of two lesions was more than four years (metachronous),

2. Histopathological results of lesions were obtained by lung resection, bronchoscopy, biopsy or aspiration:
   2.1 Lesions of the same patient were of different pathological types or subtypes, or with in-situ histology,

   2.2 Lesions of proven atypical adenomatous hyperplasia (AAH) were ruled out,
2.3 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[12].

Cases were excluded for:

1. the lack of pre-operative chest CT images,
2. distant metastases,
3. both lesions displayed as the same rare histological type such as neuroendocrine tumors or adenosquamous carcinoma[17],
4. gene testing showed lesions share the same uncommon driver mutation or fusion of the same breakpoint[18].

For the last two situations, IPM would be considered.

Flowchart of patient enrollment process was showed in Fig. 1. The same criteria were used for the validation cohort from Sun Yat-sen University Cancer Center (SYSUCC) between Jan. 2020 and Mar. 2021. This retrospective study was approved by respective Institutional Ethics Commission of SYSUFA (No.[2020]371) and SYSUCC (No. YB2018-13) with waiver of informed consent. The information of CT scan protocols were detailed in Supplementary Text.

**Histopathological Assessment and Molecular Genetic Testing**

Samples of resected lesions were sent for histopathological examination and assessed by at least two experienced pathologists. The diagnosis was based on the 2015 World Health Organization (WHO) classification[19] with immunohistochemistry carried out whenever necessary. Among them, 31 patients had at least one lesion sent for further molecular genetic testing.

**Machine Learning Based Model Development**

With the collected patient dataset, a machine learning method based on radiomic features[20] 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 PRE model training, and MPLC and IPM classification, as illustrated in Fig. 2.

**Lesion Segmentation**

All lung cancer lesions were manually delineated using an open-source software ITK-Snap (Penn Image Computing and Science Laboratory, Philadelphia, USA)[21] 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 interest (ROI) was extracted from the normalized image by clipping the image intensities in the range of [-1024, 1024] and then
mapping them to the range of \([-1, 1]\). Each lesion was processed via this method to get cropped ROI 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, property, and material construction, etc. On the contrary, 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 \( F_{AD} = [F_{AD}^1, \ldots, F_{AD}^i, \ldots, F_{AD}^{107}] \) for each lesion pair of each patient, where \( F_{AD}^i = \| F_a^i - F_b^i \| \), \( F_a^i \), and \( F_b^i \) represent the \( i \)th radiomic feature of lesion \( a \) and \( b \) from the same patient. Thus, \( F_{AD}^i \) was defined as the representative similarity measure of the lesion pair.

**Feature Selection**

Based on the similarity feature vectors calculated above, a lesion pair relation estimation (PRE) model was built via the random forest classifier\(^22\) for pairwise lesions. In this study, 70% of patients were randomly split as training dataset (SYSUFH training cohort) and the rest 30% as testing dataset for model evaluation (SYSUFH internal validation). Considering the small patient cohort, 107 features for training could be excessive to 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, of a feature pair, correlation coefficient was larger than threshold \( T_{cc} \) (empirically set as 0.7). Then the five-fold cross-validation and the random forest model fitting methods were applied to obtain the optimal feature number and evaluate importance of features respectively. Each time one feature would be added according to its ranking in a descending order to train the model on the intra-training dataset (4 of 5 folds), and the receiver operating characteristic (ROC) curve calculated on the intra-validation dataset (1 of 5 folds). The area under ROC curve (AUC)\(^23\) was monitored and validated until the optimal model achieved and its corresponding feature number obtained. In this way, the top \( k \) important features were selected with the best AUC in the five-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 lesion pair relation estimation 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 of 7/3. AUC, accuracy (acc), sensitivity (sen), specificity (spe), Negative Predictive Value (NPV) and Positive Predictive Value (PPV) were used to qualitatively evaluate the model performance in diagnosing MPLC.

For the external validation set of SYSUCC, two experienced clinicians (one experienced chest radiologist and one senior thoracic surgeon, both of at least five 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 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 \( (p > 0.05) \). Wilcoxon rank-sum test and unpaired t-test were used for statistical significance analysis \( (p < 0.05) \) on datasets of abnormal and normal distributions respectively. The radiomic feature calculation was achieved based on an open-source radiomics package (Pyradiomics 3.0.1)[20]. Model construction and evaluation were conducted on the platform of Python 3.7 (Python Software Foundation) via Pycharm 2020 (JetBrains, Czech Republic).

Results

Demographic and Histopathologic Characteristics

Clinicopathological characteristics of patients from SYSUFH were 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 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. Clinicopathological characteristics of the SYSUCC validation cohort were shown in Supplementary Table 1. Chest CT images of representative cases of MPLC and IPM were given in Fig. 3.
# Table 1
Clinicopathological characteristics of patients with MPLC and IPM.

<table>
<thead>
<tr>
<th>Patient characteristics</th>
<th>MPLC</th>
<th>IPM</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n = 76)</td>
<td>(n = 42)</td>
<td></td>
</tr>
<tr>
<td>Number of Lesions (pairs)</td>
<td>180 (137)</td>
<td>108 (93)</td>
<td>NA</td>
</tr>
<tr>
<td>Mean Age (range, years)</td>
<td>60 (32–79)</td>
<td>62 (30–86)</td>
<td>0.531</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td>0.415</td>
</tr>
<tr>
<td>Female</td>
<td>43 (56.58%)</td>
<td>27 (64.29%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>33 (43.42%)</td>
<td>15 (35.71%)</td>
<td></td>
</tr>
<tr>
<td>Smoking status, n (%)</td>
<td></td>
<td></td>
<td>0.369</td>
</tr>
<tr>
<td>Non-smokers</td>
<td>50 (65.79%)</td>
<td>31 (73.81%)</td>
<td></td>
</tr>
<tr>
<td>Current/Former smokers</td>
<td>26 (34.21%)</td>
<td>11 (26.19%)</td>
<td></td>
</tr>
<tr>
<td>Family History of Primary Lung Cancer, n (%)</td>
<td>11 (14.47%)</td>
<td>4 (9.52%)</td>
<td>0.440</td>
</tr>
<tr>
<td>Performance status, n (%)</td>
<td></td>
<td></td>
<td>0.082</td>
</tr>
<tr>
<td>0</td>
<td>63 (82.89%)</td>
<td>29 (69.05%)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13 (17.11%)</td>
<td>13 (30.95%)</td>
<td></td>
</tr>
<tr>
<td>Tumour markers, n (%)</td>
<td></td>
<td></td>
<td>0.295</td>
</tr>
<tr>
<td>Abnormal</td>
<td>47 (61.84%)</td>
<td>30 (71.43%)</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>29 (38.16%)</td>
<td>12 (28.57%)</td>
<td></td>
</tr>
<tr>
<td>Distribution of lesions, n (%)</td>
<td>25 (32.89%)</td>
<td>5 (11.9%)</td>
<td>0.019*</td>
</tr>
<tr>
<td>Same Lobe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ipsilateral</td>
<td>22 (28.95%)</td>
<td>21 (50%)</td>
<td></td>
</tr>
<tr>
<td>Contralateral</td>
<td>29 (38.16%)</td>
<td>16 (38.1%)</td>
<td></td>
</tr>
<tr>
<td>Synchronicity, n (%)</td>
<td></td>
<td></td>
<td>0.326</td>
</tr>
<tr>
<td>Synchronous</td>
<td>72 (94.74%)</td>
<td>42 (100%)</td>
<td></td>
</tr>
<tr>
<td>Metachronous</td>
<td>4 (5.26%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>Same histologic subtype of the lesion pair</td>
<td>9 (11.84%)</td>
<td>42 (100%)</td>
<td>0.000*</td>
</tr>
<tr>
<td>Staged Operation, n (%)</td>
<td>27 (35.53%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Range of Resection, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lobectomy</td>
<td>23 (30.26%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Patient characteristics</td>
<td>MPLC (n = 76)</td>
<td>IPM (n = 42)</td>
<td>p Value</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
</tr>
<tr>
<td>Bilobectomy</td>
<td>11 (14.47%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Lobectomy + Sublobectomy(ies)</td>
<td>35 (46.05%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Multiple Sublobectomies</td>
<td>7 (9.21%)</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

MPLC, multiple primary lung cancer; IPM, intrapulmonary metastasis; NA, Not Applicable; * p < 0.05 was considered as statistically significant.

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[24], which were put into the next feature selection procedure as illustrated in Supplementary Fig. 1a. 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 2) were selected for the final PRE model construction and evaluation. The correlation coefficients among these seven selected features were all smaller than 0.52 (Supplementary Fig. 1b), suggesting weak correlations of these features. Besides, feature distribution between IPM and MPLC cases was distinct, as shown in Supplementary Fig. 1c. Further analysis of the selected feature deviation (F^AD) between MPLC and IPM lesion pairs was detailed in Supplementary Fig. 2.

Model Performance

Performance of the PRE model for distinguishing MPLC from IPM on the SYSUFH training and internal validation datasets was 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. 4a1). The corresponding metrics on internal validation dataset were 0.857, 0.794, 0.758, 0.850, 0.677 and 0.906, respectively (Table 2, Fig. 4a2). With the application of major voting strategy, performance of the PRE model for the diagnosis of MPLC was 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. 4b1) and 0.844, 0.846, 0.728, 0.910, 0.861 and 0.828 on the SYSUFH internal validation cohort (Table 2 and Fig. 4b2). 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.
<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>NPV</th>
<th>PPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The PRE model trianed and tested with lesion pairs from SYSUFH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYSUFH training cohort</td>
<td>0.947</td>
<td>0.946</td>
<td>0.948</td>
<td>0.947</td>
<td>0.949</td>
<td>0.989</td>
</tr>
<tr>
<td></td>
<td>(0.944–</td>
<td>(0.941–</td>
<td>(0.941–</td>
<td>(0.943–</td>
<td>(0.943–</td>
<td>(0.988–</td>
</tr>
<tr>
<td></td>
<td>0.950)</td>
<td>0.952)</td>
<td>0.954)</td>
<td>0.952)</td>
<td>0.954)</td>
<td>0.990)</td>
</tr>
<tr>
<td>SYSUFH internal validation</td>
<td>0.794</td>
<td>0.758</td>
<td>0.850</td>
<td>0.677</td>
<td>0.906</td>
<td>0.8597</td>
</tr>
<tr>
<td></td>
<td>(0.784–</td>
<td>(0.738–</td>
<td>(0.829–</td>
<td>(0.656–</td>
<td>(0.895–</td>
<td>(0.844–</td>
</tr>
<tr>
<td></td>
<td>0.804)</td>
<td>0.778)</td>
<td>0.870)</td>
<td>0.696)</td>
<td>0.916)</td>
<td>0.869)</td>
</tr>
<tr>
<td><strong>The PRE model with major voting strategy (case-based prediction)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYSUFH training cohort</td>
<td>0.942</td>
<td>0.905</td>
<td>0.962</td>
<td>0.950</td>
<td>0.934</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>(0.938–</td>
<td>(0.895–</td>
<td>(0.956–</td>
<td>(0.945–</td>
<td>(0.924–</td>
<td>(0.981–</td>
</tr>
<tr>
<td></td>
<td>0.946)</td>
<td>0.915)</td>
<td>0.968)</td>
<td>0.954)</td>
<td>0.944)</td>
<td>0.985)</td>
</tr>
<tr>
<td>SYSUFH internal validation</td>
<td>0.846</td>
<td>0.728</td>
<td>0.910</td>
<td>0.861</td>
<td>0.828</td>
<td>0.844</td>
</tr>
<tr>
<td></td>
<td>(0.836–</td>
<td>(0.705–</td>
<td>(0.894–</td>
<td>(0.850–</td>
<td>(0.805–</td>
<td>(0.826–</td>
</tr>
<tr>
<td></td>
<td>0.873)</td>
<td>0.751)</td>
<td>0.925)</td>
<td>0.873)</td>
<td>0.852)</td>
<td>0.862)</td>
</tr>
<tr>
<td><strong>Performance of the PRE model on cases from SYSUCC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYSUCC external validation</td>
<td>0.760</td>
<td>0.727</td>
<td>0.769</td>
<td>0.909</td>
<td>0.471</td>
<td>0.793</td>
</tr>
<tr>
<td><strong>Performance of clinicians in differentiating MPLC and IPM cases from SYSUCC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radiologist</td>
<td>0.660</td>
<td>0.545</td>
<td>0.692</td>
<td>0.844</td>
<td>0.333</td>
<td>0.619</td>
</tr>
<tr>
<td>Thoracic surgeon</td>
<td>0.600</td>
<td>0.545</td>
<td>0.615</td>
<td>0.828</td>
<td>0.286</td>
<td>0.580</td>
</tr>
</tbody>
</table>

PRE model, pair relation estimation model; AUC, area under the receiver-operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value; SYSUFH, First Affiliated hospital of Sun Yat-sen University; SYSUCC, Sun Yat-sen University Cancer Center.

Validation results of the SYSUCC cohort were shown in Table 2 and Fig. 4c. 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. 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 suggested promising performance of the proposed model and its great value for clinical practice.

Discussion
In this study, we established a novel non-invasive diagnostic algorithm to differentiate MPLC from IPM. The model has shown promising diagnostic performance for MPLC with mean AUC, accuracy, sensitivity, specificity, NPV and PPV of 0.844, 0.846, 0.728, 0.910, 0.861 and 0.828. The corresponding numbers for the external validation cohort were 0.793, 0.760, 0.727, 0.769, 0.909 and 0.471 respectively, outperforming that of clinicians.

With the popularization and improved accuracy of CT screening, the occurrence of multiple pulmonary lesions increases. The incidence of MPLC reported ranged from 1.1–6.9%[25–27], consistent with the findings in this study. The preoperative diagnosis of multiple lesions has become a crucial issue for decision-making, especially the differential diagnosis between MPLC and IPM.

Since its proposal in 1975, the Martini and Melamed criteria[15] has been widely accepted for its simplicity and feasibility. The latest version of ACCP Special Treatment Issues[16] proposed that molecular genetic characteristics should be taken into account while considering MPLC. Researchers have demonstrated that the genetic profiles of MPLC from the same individual are distinct[28, 29] and techniques such as NGS could be critical for the differentiation of MPLC and IPM[30–32]. One promising tool to solve this conundrum is the whole-genome-based mate-pair sequencing technique that tells lineage through shared and unique somatic breakpoint junctions[18, 33]. Researchers also found both inter-focal heterogeneity and functional convergence phenomenon in MPLC lesions [34]. Algorithms considering comprehensive information were to help decision-making in terms of differentiation between IPM and MPLC[10, 17, 35, 36].

Still, most studies relied on pathological analysis for differentiation. It is of great importance to develop a diagnostic tool that can solve this issue before surgery. Researchers noticed that multi-GGN and solid-GGN tumors were indicative of multiple primaries[9] and multiple pure solid nodules suggestive of IPM[9, 37]. In a previous publication, J.A.B. Araujo-Filho et al.[12] found that CT imaging features including subsolid consistency, spiculated contour, size difference and smallest lesion being pure solid were significant to differentiate MPLC from IPM. Yet no model was built upon these characteristics and the results tended to be subjective. In recent years, there were studies differentiating benign from malignant pulmonary nodules[38, 39] or predicting the invasiveness of a lesion[13, 40, 41] with the help of artificial intelligence (AI). Machine learning-based models were built for multiple nodules to predict lung malignancy[42], or borrowed from solitary nodules to diagnose MPLC[14]. These methods could be inappropriate in that each patient should be considered as a whole rather than targeting each lesion separately. Besides, this model could not differentiate MPLC from IPM, a more crucial issue for patients suffering from multiple lesions. In this study, we proposed a hypothesis that the distinctions between MPLC lesions could be greater than distinctions between a primary and IPM lesion, genetically as well as radiologically. The inherent heterogeneity of MPLC versus the homogeneity of IPM lesions could be reflected in CT images. Therefore, we set out to establish an innovative model with machine learning methods comparing lesions of the same individual objectively. This PRE model has achieved better efficacy compared with previously published methods. We believe that this model is of great application value for clinical diagnosis of MPLC and treatment decision-making.
There are a few limitations in this study. Despite the increasing occurrence of MPLC in recent years, the number of included cases was rather small compared with solitary lung cancer. However, among published articles covering the issue of MPLC, 76 cases could be adequate, especially considering that each case had at least two lesions matched for imaging investigation. Besides, 39 more cases were collected from another institution as an external validation cohort, adding credibility to the value of the model. Since the performance difference of various machine learning methods was not the main concern of this study, other methods such as support vector machine (SVM), logistical regression and convolutional neural network were not employed or evaluated in this study. The application of more elaborate methods for multiple-omics tasks in the future is warranted. Besides, the building of the PRE include images of suspicious lesions without considering other factors such as age, smoking status, family history of lung cancer, etc. Even so, the performance of the PRE model is satisfactory as its AUC reached 0.846, outperforming other models reported. Moreover, in the SYSUCC external validation cohort, it showed better performance for discrimination than experienced clinicians. The efficacy of this model remains unknown in the real world and multicenter results including even more centers are intended for further application.

Conclusions

In summary, a novel machine-learning model was developed for distinguishing IPM from MPLC preoperatively. The PRE model showed an excellent performance, which could be of great value for clinical practice.

abbreviations

PRE, pair relation estimation, MPLC, multiple primary lung cancer, IPM, intrapulmonary metastasis, AUC, area under receiver operating characteristic curve, PET-CT, positron emission tomography, SUV, standard uptake value, AAH, atypical adenomatous hyperplasia, GGN, ground-glass nodule, IEC, Institutional Ethics Committee, ROI, minimal region of interest, AD, absolute deviation, ROC, receiver operating characteristic, ACCP, American College of Chest Physicians, AI, artificial intelligence, SVM, support vector machine

Declarations

Ethics approval and consent to participate

This retrospective study was approved by Institutional Ethics Commission of The First Affiliated Hospital of Sun Yat-sen University (No.[2020]371) and Institutional Ethics Commission of Sun Yat-sen University Cancer Center (No. YB2018-13) with waiver of informed consent. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by Grants from the National Natural Science Foundation of China (No. 82102109), by Grants from Development Center for Medical Science & Technology National Health Commission of China (No. WA2020RW10) and by Grants from Shanghai Municipal Commission of Health and Family Planning Program (No. 20184Y0037).

Authors' contributions

TC participated in the conceptualization and investigation of the study, and wrote the original draft. LY took part in the conceptualization, investigation, reviewing and editing the manuscript. HC was responsible for the methodology, and formal analysis. HY and ZW worked on data curation and visualization. HL QL and YZ provided resources, fundings and supervision of the project. YZ was also a major contributor in reviewing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References


thoracic oncology : official publication of the International Association for the Study of Lung Cancer 2015, 10(5):778-783.


**Figures**
Figure 1

Flowchart of the enrollment process of MPLC. CT, computed tomography, MPLC, multiple primary lung cancer, AIS, adenocarcinoma in situ
Figure 2

The main steps to develop the pairwise machine learning algorithm for MPLC and IPM differentiation. 

MPLC, multiple primary lung cancer, IPM, intrapulmonary metastasis, PRE model, lesion pair relation estimation model.

Figure 3

Pt1 (a)  
Pt1 (b)  
Pt2 (a)  
Pt2 (b)  
Pt3 (a)  
Pt3 (b)  
Pt4 (a)  
Pt4 (b)  
Pt5 (a)  
Pt5 (b)  
Pt6 (a)  
Pt6 (b)
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 patient 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 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 lesion pair relation estimation (PRE) model built by our team, diagnostic accuracy was improved significantly and hence clinical decisions made with stronger evidence. CT, computed tomography, MPLC, multiple primary lung cancer, IPM, intrapulmonary metastasis, 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. ROC of the PRE model on SYSUCC external validation cohort and AUCs of clinicians.

**Figure 4**

The quantitative results of PRE model for lesion pair relation estimation and MPLC and IPM differentiation. a) and a), ROC curves of PRE model on SYSUFH training and internal validation datasets. b) and b), ROC curves of PRE model for MPLC and IPM differentiation (with major voting strategy) on SYSUFH training and internal validation datasets. c) ROC curve of PRE model (with major voting strategy) on SYSUCC external validation dataset and AUCs of clinicians. PRE model, lesion pair
relation estimation model, MPLC, multiple primary lung cancer, IPM, intrapulmonary metastasis, ROC curve, receiver operating characteristic curve, AUC, area under ROC curve.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigure1.pdf](#)
- [SupplementaryFigure2.pdf](#)
- [SupplementaryTable1.docx](#)
- [SupplementaryTable2.docx](#)
- [SupplementaryfigurelegendsandText.docx](#)
- [SupplementaryText.docx](#)