Preoperative Prediction of Lymph Node Metastasis in Patients With Early-T-Stage Non-small Cell Lung Cancer by Machine Learning Algorithms

Yijun Wu1,2†, Jianghao Liu1,2†, Chang Han1,2, Xinyu Liu2,3, Yuming Chong1,2, Zhile Wang1,2, Liang Gong1,2, Jiaqi Zhang1, Xuehan Gao1, Chao Guo1, Naixin Liang* and Shanqing Li*†

1 Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, 2 Peking Union Medical College, Eight-year MD Program, Chinese Academy of Medical Sciences, Beijing, China, 3 Department of Radiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Background: Lymph node metastasis (LNM) is difficult to precisely predict before surgery in patients with early-T-stage non-small cell lung cancer (NSCLC). This study aimed to develop machine learning (ML)-based predictive models for LNM.

Methods: Clinical characteristics and imaging features were retrospectively collected from 1,102 NSCLC ≤ 2 cm patients. A total of 23 variables were included to develop predictive models for LNM by multiple ML algorithms. The models were evaluated by the receiver operating characteristic (ROC) curve for predictive performance and decision curve analysis (DCA) for clinical values. A feature selection approach was used to identify optimal predictive factors.

Results: The areas under the ROC curve (AUCs) of the 8 models ranged from 0.784 to 0.899. Some ML-based models performed better than models using conventional statistical methods in both ROC curves and decision curves. The random forest classifier (RFC) model with 9 variables introduced was identified as the best predictive model. The feature selection indicated the top five predictors were tumor size, imaging density, carcinoembryonic antigen (CEA), maximal standardized uptake value (SUVmax), and age.

Conclusions: By incorporating clinical characteristics and radiographical features, it is feasible to develop ML-based models for the preoperative prediction of LNM in early-T-stage NSCLC, and the RFC model performed best.

Keywords: non-small cell lung cancer, machine learning, lymph node metastasis, predictive model, cross-validation

INTRODUCTION

Lung cancer remains the leading cause of global cancer death (1). Early-T-stage non-small cell lung cancer (NSCLC) has been detected more frequently following the rapid development and employment of radiographical technology (2). An accurate nodal stage is critical for treatment decision-making (3). Currently, there are several evaluation methods, such as computed tomography (CT), positron emission tomography/CT (PET/CT), mediastinoscopy...
and endobronchial ultrasound transbronchial needle aspiration (EBUS-TBNA), that can be used to classify the nodal stage before operation. However, performing mediastinoscopy or EBUS-TBNA is not cost-effective for patients with early-stage NSCLC. Furthermore, although CT and PET/CT have been widely used for the preoperative evaluation of lung cancer, the incidence of occult lymph node metastasis (LNM) in early-T-stage NSCLC remains high and cannot be ignored (4, 5). Therefore, new reliable methods for the preoperative prediction of LNM are highly required.

Machine learning (ML) is an emerging computer-based method that has been widely used for data analysis in medicine during the past decade (6, 7). It learns from data and finds the dataset pattern to identify the outcome (7, 8). Supervised ML is a process in which the model is trained with fully labeled and classified data. Compared with conventional statistical methods such as logistic regression (LR), which relies on predetermined models, ML can deeply detect the interactions among variations and iteratively learn from data to update algorithms (9).

A number of predictive models have been made based on ML algorithms. Several studies have reported effective ML-based models for the prediction of LNM in other carcinomas, such as breast cancer (10, 11). It was reported that radiomics could be used to predict LNM by analyzing radiological images in NSCLC (12). However, few reports have incorporated clinical characteristics and radiographical features as in our study. This study aimed to develop and validate effective ML-based models for the prediction of LNM in patients with early-T-stage NSCLC.

MATERIALS AND METHODS

Study Population

Between January 2013 and June 2019, 1,102 patients who underwent surgical resection for NSCLC at Peking Union Medical College Hospital were included in this study. The inclusion criteria were as follows: (1) single NSCLC lesion; (2) tumor maximum diameter ≤ 2 cm on CT; and (3) receiving lung resection with systematic lymph node dissection. The exclusion criteria were as follows: (1) small cell lung cancer (SCLC); (2) multiple lung cancer; (3) receiving radiotherapy or chemotherapy before surgery; (4) distant metastasis; and (5) incomplete clinical records. The pathological classification of carcinomas was based on the 2015 World Health Organization (WHO) classification (13). The clinical and pathological staging was performed according to the 8th edition of the TNM staging system (14). This study was approved by the Ethics Committee of Peking Union Medical College Hospital. All patients signed informed consent before operation.

Clinical Characteristics and Radiographical Features

A total of 23 variables were analyzed in this study. The patients’ clinical characteristics included age, sex, smoking status and serum tumor biomarkers. All preoperative serum tumor biomarkers were measured within 3 months before surgery, including carbohydrate antigen 24-2 (CA242), squamous cell carcinoma antigen (SCCAg), carcinoembryonic antigen (CEA), carbohydrate antigen 19-9 (CA199), carbohydrate antigen 12-5 (CA125), carbohydrate antigen 72-4 (CA724), carbohydrate antigen 15-3 (CA153), neuron-specific enolase (NSE), tissue polypeptide-specific antigen (TPS), cytokeratin 19-fragments (Cyfra211) and pro-gastrin-releasing peptide (proGRP). CT features were reviewed by one radiologist and two thoracic surgeons independently, including tumor location side, tumor maximum size, spiculation, vessel convergence, lobulation, pleural indentation, calcification, and imaging density. If disagreement occurred, the final result was reached by consensus. Based on imaging density on CT, the cancer lesions were divided into pure ground-glass opacity (pGGO), mixed GGO (mGGO) and solid nodules. The mGGO was further divided into two groups according to different percentages of solid components, whose cut-off value was 50% (the ratio between the maximal diameter of the solid component at the mediastinal window and the maximal tumor diameter at the lung window). In addition, the maximal standardized uptake value (SUV\text{max}) on PET/CT was also included. However, PET scan was not routinely performed in early-T-stage NSCLC. All patients underwent CT or PET scan within 60 days at our hospital before the operation.

Construction of ML-Based Models

All patients were randomly divided into training and testing groups at a ratio of 8:2, keeping the distribution of node-positive and node-negative data in both groups consistent. To construct more reliable ML-based predictive models, all continuous variables were preprocessed by z-score normalization except for multinomial naïve Bayes (MNB) in which min-max normalization is preferred (15). Some continuous variables with missing data (Table S1), such as SUV\text{max} and tumor biomarkers, were imputed by median value (16, 17).

Eight algorithms were applied to predict LNM, including adaptive boosting (AdaBoost), artificial neural network (ANN), decision tree (DT), gradient boosting decision tree (GBDT), logistic regression (LR), MNB, random forest classifier (RFC), and extreme gradient boosting (XGBoost) (18–23). Among all 8 algorithms, LR and MNB are considered conventional methods, and the others are representative supervised ML-based algorithms. Only DT, LR, and MNB were interpretable, in which users were able to recognize function between variable and predictive outcome.

The prediction ability of the 8 models was first evaluated by the receiver operating characteristic (ROC) curve, which is a conventional diagnostic test method that only pays attention to the sensitivity and specificity but ignores the clinical utility of predictive information. Decision curve analysis (DCA) was performed to calculate the clinical values of these models, which is a novel method to assess the information value between diagnostic models by considering the possible range of a patient's risk and benefit preferences without actually measuring these preferences for one particular patient (24).

Validation Strategy and Feature Selection

Overfitting is a common problem in ML, especially with high dimensions (number of variables). To minimize the negative influence of overfitting, some strategies, such as the preselection of variables and cross-validation, were feasible (25,
### Predictive Performance and Clinical Utility of ML-Based Models

A total of 23 preoperative variables were used to develop predictive models for LNM based on 8 algorithms. The predictive performance of all models is shown in **Figure 1** and **Table 2**. The best performance was observed in the GBDT model (AUC = 0.899, SD = 0.048), which performed similarly to RFC (AUC = 0.890, SD = 0.045, P = 0.773), XGBoost (AUC = 0.883, SD = 0.047, P = 0.627), AdaBoost (AUC = 0.873, SD = 0.048, P = 0.432), and ANN (AUC = 0.868, SD = 0.049, P = 0.341). All
ML-based models except DT (AUC = 0.802, SD = 0.057) were better than the two conventional methods, LR (AUC = 0.867, SD = 0.049, P = 0.338) and MNB (AUC = 0.784, SD = 0.058, P = 0.002). Moreover, all models performed significantly better than using only tumor size (AUC = 0.753, SD = 0.023, P < 0.001; the cut-off value was 1.5 cm), SUV$_{\text{max}}$ (AUC = 0.734, SD = 0.024, P < 0.001; the cut-off value was 2.8) or CEA (AUC = 0.720, SD = 0.026, P < 0.001; the cut-off value was 2.98 ng/ml).

Furthermore, the decision curve showed the clinical values of these models (Figure 2). The net benefits of 8 models at each threshold probability are shown in Table S2. Most of these models presented better net benefits than two control models that were represented by positive and negative line, respectively. The negative line represents the net benefit is zero when none of patients receive lobectomy with systematic lymph node dissection (SND), assuming that all patients have no positive nodes. On the contrary, the positive line represents the net benefits at the time when all patients have positive nodes and receive lobectomy with SND. Four models (RFC, XGBoost, GBDT, and LR) performed significantly better than the others at most of threshold points. At the range of 0.2–0.5, the LR model was less beneficial than RFC, XGBoost and GBDT on most occasions. The RFC model with 9 variables introduced, which achieved a very high AUC (0.890) and had the highest net benefits almost across the entire range of threshold probabilities, was regarded as the best predictive model in this study, although its AUC was slightly lower than that of GBDT (P = 0.773).

Variable Importance

By feature selection, the 23 variables for each algorithm were ranked by their predictive importance (Table S3). The top 10 variables are shown in Figure 3. The five top-ranked predictors were tumor size, imaging density, CEA, SUV$_{\text{max}}$, and age. The relationship between the AUCs of models and the number of variables were evaluated in Figure 4. The AUCs of most models reached a plateau when 7 variables were introduced, while those of ANN, DT, and MNB started to drop down when they reached the highest points. The AUCs of RFC for each number of variables are shown in Figure 5. Its AUC value reached a plateau when 9 variables were introduced and reached the highest value when 13 variables were introduced, but it did not increase significantly with the change from 9 variables (AUC = 0.886) to 13 variables (AUC = 0.890) introduced. Considering the clinical utility, the 9 top-ranked variables were identified to construct the optimal predictive model, which included tumor size, SUV$_{\text{max}}$, imaging density, vessel convergence sign, CEA, CA125, sex, age, and spiculation sign.
DISCUSSION

Lobectomy with systematic lymph node dissection remains the standard treatment for patients with early-T-stage NSCLC (≤ 2 cm) (27). However, sublobar resection, including segmentectomy and wedge resection, has been proposed to achieve more precise intervention with the advancement of imaging techniques in recent years. In addition, the reasonable extent of lymph node dissection remains controversial. An exact nodal status is critical for treatment selection and prognosis.

In this study, using ML algorithms, we developed 8 models to predict LNM in 1,102 patients with NSCLC ≤ 2 cm, incorporating their clinical characteristics and radiographical features. ROC analysis and DCA were used to evaluate the predictive performance and clinical values of the models, respectively. Most of 8 models maintained high AUCs and All ML-based models (with AUCs ranging from 0.868 to 0.899) except DT performed better than two models using conventional statistical methods (LR and MNB) in the prediction of LNM (Figure 1 and Table 2).

DCA has been used for many medical studies and has shown great clinical utility (28, 29). In the decision curve, most of these models performed better than positive line and negative line, indicating that the overall net benefit of giving lobectomy with SND to patients identified by the models to have high risk of LNM was higher than that of giving the same surgical procedures to all patients or no patient. Four models (RFC, XGBoost, GBDT, and LR) performed better than the others at most of threshold points (Figure 2). Thus, these four potential models were used to identify variable importance by feature selection (Figure 3). The other four models, AdaBoost, MNB, DT, and ANN, had lower net benefits in the decision curve (Figure 2), although they possessed high AUCs in the ROC curve. This indicated that models with high predictive accuracy might not be clinically practical and require further evaluation by other methods, such as DCA.

Using conventional univariate analysis, previous studies reported the risk factors associated with LNM in NSCLC ≤ 2 cm, including tumor size, serum CEA and imaging density (30, 31). In addition, $SUV_{\text{max}}$ was also thought to be a risk factor in patients with cT1 NSCLC (32). Thus, the AUCs when using tumor size (AUC = 0.753), $SUV_{\text{max}}$ (AUC = 0.734), or CEA (AUC = 0.720) alone were also calculated, which were significantly lower than those of ML-based models (Table 2). Thus, previous studies might not provide precise predictive information for LNM. Reliable predictive models for LNM in patients with NSCLC are needed. To our knowledge, our study was the first to provide potential models for the prediction of LNM in patients with NSCLC by incorporating clinical characteristics and radiographical features.
Although most of the ML-based models in our study cannot demonstrate the connection between the predictive variables and the outcomes, the contribution of each variable to the models could be inferred by feature selection. Tumor size, imaging density, serum CEA, \( \text{SUV}_{\text{max}} \), and age were indicated to be the most contributive risk factors of LNM (Figure 3), which was similar to the results of univariate analysis (Table 1). Since none of the patients with pGGO NSCLC had positive nodes in our and previous studies (30, 31), it could be inferred that pGGO might be predictive of node-negative status in early-T-stage NSCLC. It was also reported that a higher serum CEA level was significantly associated with a higher incidence of LNM (31, 33). Although only 611 patients’ \( \text{SUV}_{\text{max}} \) values (pN+: \( n = 62 \), pN0: \( n = 549; p > 0.05 \)) were available because some patients did not undergo PET scans, \( \text{SUV}_{\text{max}} \) was ranked at 4 among the four potential models (Figure 3) and was ranked at 2 in the RFC model (Figure 4). Meanwhile, a high AUC (0.734) for \( \text{SUV}_{\text{max}} \) was also obtained. Above all, \( \text{SUV}_{\text{max}} \) might be one of the most important predictive factors, which was consistent with previous studies (32, 34). Surprisingly, age showed no significance in univariate analysis (\( p = 0.382 \)) but was ranked at the top 5 (Figure 3). This might be attributed to the surprising superiority of ML-based models in data mining, which could find more relations between the variables and the outcomes than conventional methods.

According to the ROC curve (Figure 1) and decision curve (Figure 2), the RFC model with 9 variables introduced (AUC = 0.890) was identified as the optimal model. By considering the clinical utility, an application based on the RFC algorithm with 9 variables (AUC = 0.886) should be developed in the future. These 9 variables were tumor size, \( \text{SUV}_{\text{max}} \), imaging density, vessel convergence sign, CEA, CA125, sex, age, and spiculation sign. Thus, clinicians from other hospitals could benefit from our study.

In addition to the clinical values, there were several methodological indications in our study. First, although there were several studies of machine learning involving NSCLC, few of them have reported predictive models for LNM using ML algorithms by incorporating clinical characteristics and radiographical features. Most of them performed image analysis by radiographical data (12) or histological slides (35). This is the first study to predict LNM in NSCLC \( \leq 2 \text{ cm} \) by incorporating clinical and radiographical characteristics. Based on ROC analysis and DCA, some ML-based models performed better than models using conventional methods, and the RFC model performed best. The feature selection approach identified that tumor size, imaging density, CEA, \( \text{SUV}_{\text{max}} \), and age were the most important predictive risk factors for LNM.

**CONCLUSIONS**

ML-based models are effective in the prediction of LNM in NSCLC \( \leq 2 \text{ cm} \) by incorporating clinical and radiographical characteristics. Based on ROC analysis and DCA, some ML-based models performed better than models using conventional methods, and the RFC model performed best. The feature selection approach identified that tumor size, imaging density, CEA, \( \text{SUV}_{\text{max}} \), and age were the most important predictive risk factors for LNM.

**DATA AVAILABILITY STATEMENT**

All datasets generated for this study are included in the article/Supplementary Material.

**ETHICS STATEMENT**

The studies involving human participants were reviewed and approved by Ethics Committee of Peking Union Medical College Hospital. Written informed consent to participate in this study was provided by the participants’ legal guardian/next of kin.

**AUTHOR CONTRIBUTIONS**

SL, NL, YW, and JL: conceptualization. YW and JL: methodology. YW, JL, CH, XL, and YC: formal analysis. YW, ZW, LG, JZ, XG, and CG: investigation. YW and JL: writing—original draft preparation. YW, SL, and NL: writing—review and editing.

SL: supervision.

**FUNDING**

This research was funded by (1) Foundation for Key Program of Ministry of Education, China (Grant No. 311037); (2) CAMS Innovation Fund for Medical Sciences (CIFMS), (2017-12M-1-009; 2019-12M-1-001); (3) Beijing Natural Science Foundation (7182132); (4) Special Data Service for Oncology, The National Population and Health Scientific Data Sharing Platform (NCMI-ABD02-201809; NCMI-YF02N-201906), supported by Ministry of Science and
ACKNOWLEDGMENTS

We would like to give our sincere thanks to Professor Hongsheng Liu, Yushang Cui, Zhijun Han and Zhili Cao for their contributions to the clinical works and we also would like to thank American Journal Experts (www.aje.com) for its linguistic assistance during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.00743/full#supplementary-material
34. Li L, Ren S, Zhang Y, Guan Y, Zhao J, Liu J, et al. Risk factors for predicting the occult nodal metastasis in T1-2N0M0 NSCLC patients staged by PET/CT: potential value in the clinic. Lung Cancer. (2013) 81:213–7. doi: 10.1016/j.lungcan.2013.04.012

35. Takamatsu M, Yamamoto N, Kawachi H, Chino A, Saito S, Ueno M, et al. Prediction of early colorectal cancer metastasis by machine learning using digital slide images. Comput Methods Programs Biomed. (2019) 178:155–61. doi: 10.1016/j.cmpb.2019.06.022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wu, Liu, Han, Chong, Wang, Gong, Zhang, Gao, Liang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.