frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fonc.2022.1054300

Prediction of bone metastasis in non-small cell lung cancer based on machine learning

Meng-Pan Li[1,2†], Wen-Cai Liu[1,2,3†], Bo-Lin Sun[1,4], Nan-Shan Zhong[1,4], Zhi-Li Liu[1,4], Shan-Hu Huang[1,4], Zhi-Hong Zhang[1,4*] and Jia-Ming Liu[1,4*]

[1]Department of Orthopedic Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, China, [2]The First Clinical Medical College of Nanchang University, Nanchang, China, [3]Department of Orthopaedics, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China, [4]Institute of Spine and Spinal Cord, Nanchang University, Nanchang, China

Objective: The purpose of this paper was to develop a machine learning algorithm with good performance in predicting bone metastasis (BM) in non-small cell lung cancer (NSCLC) and establish a simple web predictor based on the algorithm.

Methods: Patients who diagnosed with NSCLC between 2010 and 2018 in the Surveillance, Epidemiology and End Results (SEER) database were involved. To increase the extensibility of the research, data of patients who first diagnosed with NSCLC at the First Affiliated Hospital of Nanchang University between January 2007 and December 2016 were also included in this study. Independent risk factors for BM in NSCLC were screened by univariate and multivariate logistic regression. At this basis, we chose six commonly machine learning algorithms to build predictive models, including Logistic Regression (LR), Decision tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Naive Bayes classifiers (NBC) and eXtreme gradient boosting (XGB). Then, the best model was identified to build the web-predictor for predicting BM of NSCLC patients. Finally, area under receiver operating characteristic curve (AUC), accuracy, sensitivity and specificity were used to evaluate the performance of these models.

Results: A total of 50581 NSCLC patients were included in this study, and 5087 (10.06%) of them developed BM. The sex, grade, laterality, histology, T stage, N stage, and chemotherapy were independent risk factors for NSCLC. Of these six models, the machine learning model built by the XGB algorithm performed best in both internal and external data setting validation, with AUC scores of 0.808 and 0.841, respectively. Then, the XGB algorithm was used to build a web predictor of BM from NSCLC.

Conclusion: This study developed a web predictor based XGB algorithm for predicting the risk of BM in NSCLC patients, which may assist doctors for clinical decision making

KEYWORDS

bone metastasis, NSCLC, machine learning, XGB algorithm, SEER

EDITED BY
Shyamala Guruvare,
Manipal Academy of Higher Education, India

REVIEWED BY
Shobitha Rao,
Srinivas University, India
Geetha M.,
Manipal Institute of Technology, India

*CORRESPONDENCE
Jia-Ming Liu
liujiamingdr@hotmail.com
Zhi-Hong Zhang
13803505665@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cancer Epidemiology and Prevention, a section of the journal Frontiers in Oncology

RECEIVED 26 September 2022
ACCEPTED 21 November 2022
PUBLISHED 09 January 2023

CITATION
Li M-P, Liu W-C, Sun B-L, Zhong N-S, Liu Z-L, Huang S-H, Zhang Z-H and Liu J-M (2023) Prediction of bone metastasis in non-small cell lung cancer based on machine learning. *Front. Oncol.* 12:1054300. doi: 10.3389/fonc.2022.1054300

Introduction

Lung cancer, as one of the most common malignant tumors worldwide, has an annual incidence of 2 million and causes 1.76 million deaths each year (1). Non-small cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer cases, which has an improving overall survival rate due to better therapy (2). However, bone metastasis (BM) is a negative prognostic factor in NSCLC patients. Studies have reported that the incidence of BM in patients with NSCLC is 26-36%, and the 2-year survival rate of patients with BM is 3% (3). Also, bone metastases often result in a range of complications, such as pain, hypercalcemia, spinal cord compression, pathological fractures and neurological defect, which will decrease the patient's quality of life (4).

Early diagnosis and intervention in patients with BM could significantly improve prognosis of patients. At present, bone scan is the most classic way to detect bone related diseases (5). But due to its low sensitivity to bone metastases, early BM from cancer is often not detected. Although studies had shown that PET-CT can improve the detection rate of small bone lesions (6), it is limited as a screening tool due to its high cost and high radiation. Therefore, bone scans and PET-CT are recommended only when there is a suspicious bone-related event, which usually occurs 5 months after BM (7). By then, many NSCLC patients may have developed multiple metastases, and the prognosis of patients is poor.

Previous studies (8–10) reported that histopathological type, gender, histological differentiation, serum CA-125, ALP, and multiple lymph node metastasis are independent risk factors for BM of lung cancer, which lays the foundation for prediction model construction. Zhang Chao et al (10). constructed a nomogram to predict the BM in different histological types of lung cancer based on the traditional logistic model. However, the limitations of this method in prediction accuracy and processing big data have made it difficult to make great breakthroughs in precision medicine (11, 12). Therefore, advanced machine learning models were used in this study.

Compared with traditional logistical model, machine learning (ML) technology can unlock more information in large datasets to achieve the purpose of outcome prediction and have higher accuracy (12). There already are many applications of this technology throughout science and society ranging from driverless cars to Board games to decision-making (12, 13). In biomedicine, the development of big data in healthcare (14, 15) offers great potential for ML to understand disease and health and ML has been used in clinical diagnostics, precision therapeutics, and health monitoring (16).

Therefore, in this study, we aimed to found a machine learning algorithm with good prediction performance, and establish it as a web-based calculator that can be easily used to predict the risk of BM in NSCLC patients.

Methods

Study population

In the study, we used SEER*stat 8.4.0 software to download the patients' data from the SEER-Medicare database submitted in November 2020. Patients diagnosed with lung cancer between 2010 and 2018 were involved in this study. Exclusion criteria were detailed as follows: (1) The histological type of lung cancer is small cell cancer and unknown; (2) The information of T stage, N stage, race, grade, marital status, and bone metastatic status missed or unknown; (3) LC is not the first tumor. A study flow chart of case screening was presented in Figure 1. Additionally, 507 patients newly diagnosed with NSCLC were included in this study between January 2007 and December 2016 at the First Affiliated Hospital of Nanchang University.

Data selection

In this study, 11 variables related to the clinicopathology and demographics of patients were selected for analysis. Demographic variables included race, sex, marital and age. Clinicopathological variables included primary site, grade, histology, T stage, N stage, laterality and chemotherapy. According to the ICD-O-3 codes, histological types of NSCLC were divided into 5 categories, including adenocarcinoma (814-838), squamous cell carcinoma (805-808), adenosquamous carcinoma (856), large cell carcinoma (8012-8014) and others. All NSCLC patients were staged according the AJCC 7th edition guidelines and SEER staging information. In addition, we divided patients into two groups at 60 years to analyze the effect of age on outcome events by referring to the study of Zhou et al (8).

Data pre-processing and feature engineering

All statistical analyses were conducted with Python3.8, SPSS 23 and R 4.2.0. In this study We performed a logistic regression analysis on data collected in the SEER database to identify suitable variables for machine learning model by using SPSS 23 software. Significant variables between BM and non-BM patients were identified by univariate logistic regression analysis (P<0.05). Then, these variables were enclosed within multivariate logistic regression analysis, and variables with a P < 0.05 in multivariate logistic regression analysis were subjected for further analysis of ML model. Correlation analysis was used to analyze the correlation among the selected features. Since this data set is an unbalanced data set, the over-sampling method were adopted for data processing (17). The key of this method is to oversampling the data samples of small classes to increase the

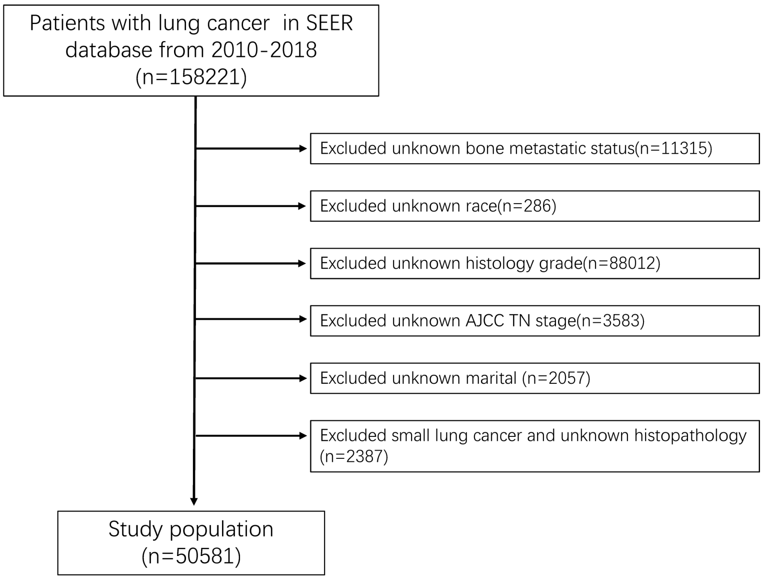

Patients with lung cancer in SEER database from 2010-2018 (n=158221)
Excluded unknown bone metastatic status(n=11315)
Excluded unknown race(n=286)
Excluded unknown histology grade(n=88012)
Excluded unknown AJCC TN stage(n=3583)
Excluded unknown marital (n=2057)
Excluded small lung cancer and unknown histopathology (n=2387)
Study population (n=50581)

FIGURE 1
The study flow chart of case screening.

number of data samples of small classes to improve the accuracy of the model.

For the 507 external samples, the cancer stage and grade were unified according to the AJCC 7th edition criteria, so that the parameters of the two data sets could be matched. Missing values were imputed by R mouse package using classification and regression tree principle (18). Meanwhile, to compare the importance of each feature, we extract the feature importance of each variable in the machine learning model according to the Permutation Importance principle (19).

Model establishment and evaluation

During the modeling establishment, SEER data set was used as internal data to build and validate the models, while hospital data were used as external validation data to validate and evaluate the predictive ability of the machine learning models. The risk stratification threshold of the model was set to 0.5(50%) (20).

Six commonly used classifier algorithms were chosen to this study, including three ensemble algorithms (12) (Random Forest (RF), Gradient Boosting Machine (GBM), eXtreme gradient boosting (XGB)) and three simple classification algorithms (Logistic Regression (LR), Decision tree (DT), Naive Bayes classifiers (NBC)). The ML models were trained using Python software. In the internal test, all SEER data were divided into 10 parts for 10-fold cross-validation (21). For external testing, external data was imported directly into the built model for verification. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, accuracy and F-score were evaluated indicators of ML algorithms. Comparing the evaluation indexes of each model, the best model was selected to build a network predictor.

Results

Demographic and pathological characteristics

In this study, 50581 patients diagnosed with NSCLC in the SEER database were included. Of whom, 5087 (10.06%) developed BM and 45494 (89.94%) had no BM. Meanwhile, 507 NSCLC patients at the First Affiliated Hospital of Nanchang University were selected for external validation of the models. The Characteristics of all data is showed in Table 1.

Logistic regression analysis

The univariate analysis based on the training set is presented in Table 2, and the results showed that marital was not significantly associated with the BM in NSCLC patients ($P >$ 0.05). The remaining ten significant variables were selected for multivariate logistic regression analysis. Multivariate logistic regression analysis (Table 2) indicated that seven variables, including sex, grade, laterality, histology, T stage, N stage, and chemotherapy, were independent risk factors for BM of NSCLC.

TABLE 1 Clinical and pathological characteristics of training set and external test set.

Variables	Train Set (n=50581)		External Test Set (n=507)	
	NBM (%) *N*=45494	BM (%) *N*=5087	NBM (%) *N*=441	BM (%) *N*=66
Race:				
Black	4853 (10.7%)	628 (12.3%)	0	0
White	36818 (80.9%)	3924 (77.1%)	0	0
Others	3823 (8.40%)	535 (10.5%)	441 (100%)	66 (100%)
Sex:				
Female	22809 (50.1%)	2189 (43.0%)	206 (46.7%)	28 (42.4%)
Male	22685 (49.9%)	2898 (57.0%)	235 (53.3%)	38 (57.6%)
Primary.Site:				
Main bronchus	1097 (2.41%)	200 (3.93%)	13 (2.95%)	5 (7.58%)
Upper lobe	25860 (56.8%)	2760 (54.3%)	247 (56.0%)	42 (63.6%)
Middle lobe	2372 (5.21%)	246 (4.84%)	32 (7.26%)	3 (4.55%)
Lower lobe	14160 (31.1%)	1434 (28.2%)	137 (31.1%)	14 (21.2%)
Others	2005 (4.41%)	447 (8.79%)	12 (2.72%)	2 (3.03%)
Grade:				
Grade I	7090 (15.6%)	302 (5.94%)	113 (25.6%)	5 (7.58%)
Grade II	17875 (39.3%)	1408 (27.7%)	177 (40.1%)	24 (36.4%)
Grade III	19567 (43.0%)	3217 (63.2%)	148 (33.6%)	35 (53.0%)
Grade IV	962 (2.11%)	160 (3.15%)	3 (0.68%)	2 (3.03%)
Laterality:				
Left	18798 (41.3%)	2129 (41.9%)	164 (37.2%)	29 (43.9%)
Right	26393 (58.0%)	2855 (56.1%)	275 (62.4%)	36 (54.5%)
Others	303 (0.67%)	103 (2.02%)	2 (0.45%)	1 (1.52%)
Histology:				
Adenocarcinoma	24098 (53.0%)	3113 (61.2%)	312 (70.7%)	42 (63.6%)
squamous cell carcinoma	14106 (31.0%)	1033 (20.3%)	75 (17.0%)	9 (13.6%)
Adenosquamous carcinoma	696 (1.53%)	65 (1.28%)	9 (2.04%)	1 (1.52%)
Large cell carcinoma	668 (1.47%)	89 (1.75%)	2 (0.45%)	1 (1.52%)
Others	5926 (13.0%)	787 (15.5%)	43 (9.75%)	13 (19.7%)
T stage:				
T1	15875 (34.9%)	483 (9.49%)	156 (35.4%)	3 (4.55%)
T2	14885 (32.7%)	1359 (26.7%)	146 (33.1%)	11 (16.7%)
T3	7984 (17.5%)	1365 (26.8%)	82 (18.6%)	13 (19.7%)
T4	6750 (14.8%)	1880 (37.0%)	57 (12.9%)	39 (59.1%)
N stage:				
N0	27907 (61.3%)	1027 (20.2%)	269 (61.0%)	13 (19.7%)
N1	4431 (9.74%)	482 (9.48%)	36 (8.16%)	5 (7.58%)
N2	10121 (22.2%)	2389 (47.0%)	106 (24.0%)	35 (53.0%)
N3	3035 (6.67%)	1189 (23.4%)	30 (6.80%)	13 (19.7%)
Chemotherapy:				
NO	29077 (63.9%)	2110 (41.5%)	274 (62.1%)	23 (34.8%)
YES	16417 (36.1%)	2977 (58.5%)	167 (37.9%)	43 (65.2%)
Marital				
Unmarride	12987 (28.5%)	1504 (29.6%)	–	–
Married	32507 (71.5%)	3583 (70.4%)	–	–
Age:				
<=60	8944 (19.7%)	1260 (24.8%)	79 (17.9%)	20 (30.3%)
>60	36550 (80.3%)	3827 (75.2%)	362 (82.1%)	46 (69.7%)

BM, bone metastasis; NBM, no bone metastasis.

TABLE 2 Univariate analysis and multivariate logistic regression analysis of variables.

Variables	Univariate Analysis		Multivariate Logistic Analysis	
	OR (95%Cl)	P-value	OR (95%Cl)	P-value
Race:				
Black	Reference		Reference	
White	0.824 (0.753-0.901)	<0.001	0.976 (0.887-1.074)	0.617
Others	1.081 (0.956-1.223)	0.212	1.082 (0.948-1.234)	0.242
Sex:				
Female	Reference		Reference	
Male	1.331 (1.256-1.411)	<0.001	1.221 (1.147-1.301)	<0.001
Primary.Site:				
Main bronchus	Reference		Reference	
Upper lobe	0.585 (0.501-0.684)	<0.001	1.036 (0.877-1.222)	0.679
Middle lobe	0.569 (0.466-0.695)	<0.001	1.160 (0.936-1.439)	0.175
Lower lobe	0.555 (0.473-0.652)	<0.001	0.148 (0.967-1.362)	0.115
Others	1.223 (1.019-1.467)	0.031	1.206 (0.990-1.469)	0.062
Grade:				
Grade I	Reference		Reference	
Grade II	1.849 (1.628-2.100)	<0.001	1.434 (1.254-1.640)	<0.001
Grade III	3.860 (3.420-4.356)	<0.001	1.888 (1.658-2.150)	<0.001
Grade IV	3.905 (3.187-4.784)	<0.001	1.666 (1.333-2.084)	<0.001
Laterality:				
Left	reference		Reference	
Right	0.955 (0.900-1.013)	0.128	0.912 (0.855-0.972)	0.005
Others	3.001 (2.390-3.770)	<0.001	1.227 (0.944-1.595)	0.126
Histology:				
Adenocarcinoma	Reference		Reference	
squamous cell carcinoma	0.567 (0.527-0.610)	<0.001	0.428 (0.396-0.463)	<0.001
Adenosquamous carcinoma	0.723 (0.559-0.935)	0.013	0.587 (0.449-0.769)	<0.001
Large cell carcinoma	1.031 (0.824-1.291)	0.787	0.729 (0.569-0.932)	0.012
Others	1.028 (0.946-1.117)	0.515	0.875 (0.799-0.957)	0.004
T stage:				
T1	Reference		Reference	
T2	3.001 (2.698-3.337)	<0.001	1.986 (1.778-2.219)	<0.001
T3	5.619 (5.048-6.255)	<0.001	3.007 (2.682-3.371)	<0.001
T4	9.154 (8.250-10.157)	<0.001	4.029 (3.596-4.513)	<0.001
N stage:				
N0	Reference		Reference	
N1	2.956 (2.641-3.309)	<0.001	2.163 (1.923-2.433)	<0.001
N2	6.414 (5.941-6.925)	<0.001	3.988 (3.664-4.341)	<0.001
N3	10.645 (9.715-11.666)	<0.001	5.700 (5.152-6.307)	<0.001
Chemotherapy:				
NO	Reference		Reference	
YES	2.499 (2.356-2.651)	<0.001	1.108 (1.037-1.184)	<0.001
Marital				
Unmarried	Reference			
Married	0.952 (0.893-1.014)	0.127		
Age:				
<=60	Reference		Reference	
>60	0.743 (0.695-0.795)	<0.001	0.992 (0.922-1.068)	0.837

OR, odds ratio.

The seven variables were used further machine learning model study.

Correlation analysis of features

Correlation analysis between data features is often used to measure the degree of correlation between factors. To identify the independence between features, we obtained a correlation heat map by Spearman correlation analysis. The figure showed that there was no strong correlation among these 11 features (Figure 2).

Importance of features on prediction

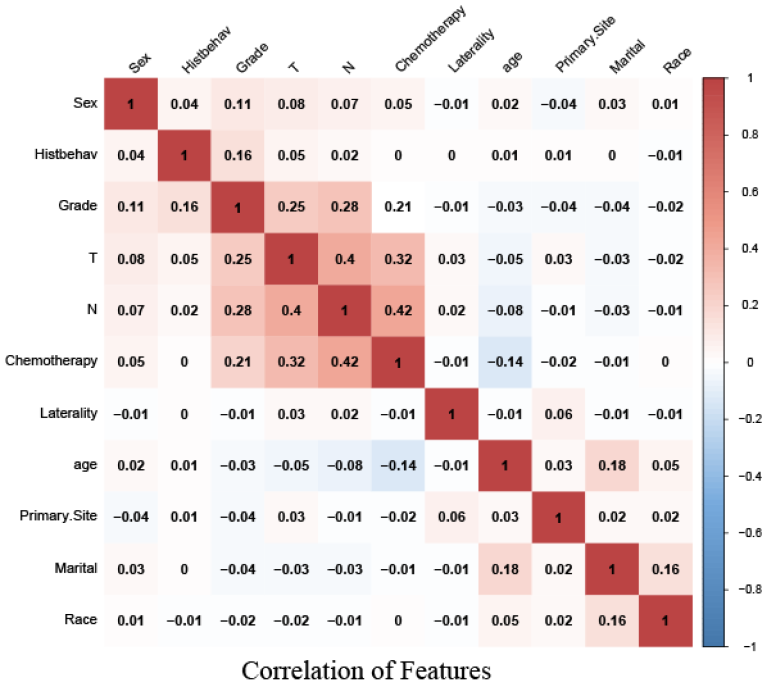
The importance of features extracted from each machine learning algorithm are shown in Figure 3. Variables screened by univariate and multivariate logistic analysis all have made extraordinary contributions to prediction in the six models. N-stage ranked top one in feature importance of all prediction models, indicating that N-stage has a great influence on BM of NSCLC, followed by T-stage. In most algorithms, grading, histology, laterality, gender and chemotherapy ranked the last five, with no significant difference in their contributions to the model. N-stage, T-stage, chemotherapy, histology, grade, sex and laterality are arranged in descending order in XGB model.

Model performance

The performance of the six predictive models is described in Figures 4, 5 and Table 3. Internal ten-fold cross-validation (Figure 4) showed that XGB model performed best among the six models with an average AUC of 0.808, followed by the GBM model (AUC=0.804). External test validation was shown in Table 3 and Figure 5. Interestingly, the XGB model also achieves the best AUC score (0.841) in the external test validation and the score of accuracy, sensitivity (recall rate) and specificity were 0.744, 0.735 and 0.803, respectively. The confusion matrix (Figure 6) of the XGB model in the training set and the test set indicated its high accuracy.

Web predictor

In this study, a web predictor based on the XGB model, which has the best predictive performance on BM of NSCLC patients, was developed to assist doctors to make more accurate clinical decisions. The odds of BM from NSCLC patients could be easily calculated by simply setting the variables associated with BM given on the web predictor. (https://share.streamlit.io/liuwencai6/lung_bone/main/lung.py) (Figure 7). When the predicted value of the patient is greater than 0.5, the network predictor shows high risk.



	Sex	Histbehav	Grade	T	N	Chemotherapy	Laterality	age	Primary.Site	Marital	Race
Sex	1	0.04	0.11	0.08	0.07	0.05	−0.01	0.02	−0.04	0.03	0.01
Histbehav	0.04	1	0.16	0.05	0.02	0	0	0.01	0.01	0	−0.01
Grade	0.11	0.16	1	0.25	0.28	0.21	−0.01	−0.03	−0.04	−0.04	−0.02
T	0.08	0.05	0.25	1	0.4	0.32	0.03	−0.05	0.03	−0.03	−0.02
N	0.07	0.02	0.28	0.4	1	0.42	0.02	−0.08	−0.01	−0.03	−0.01
Chemotherapy	0.05	0	0.21	0.32	0.42	1	−0.01	−0.14	−0.02	−0.01	0
Laterality	−0.01	0	−0.01	0.03	0.02	−0.01	1	−0.01	0.06	−0.01	−0.01
age	0.02	0.01	−0.03	−0.05	−0.08	−0.14	−0.01	1	0.03	0.18	0.05
Primary.Site	−0.04	0.01	−0.04	0.03	−0.01	−0.02	0.06	0.03	1	0.02	0.02
Marital	0.03	0	−0.04	−0.03	−0.03	−0.01	−0.01	0.18	0.02	1	0.16
Race	0.01	−0.01	−0.02	−0.02	−0.01	0	−0.01	0.05	0.02	0.16	1

FIGURE 2
Heat map of the correlation of features.

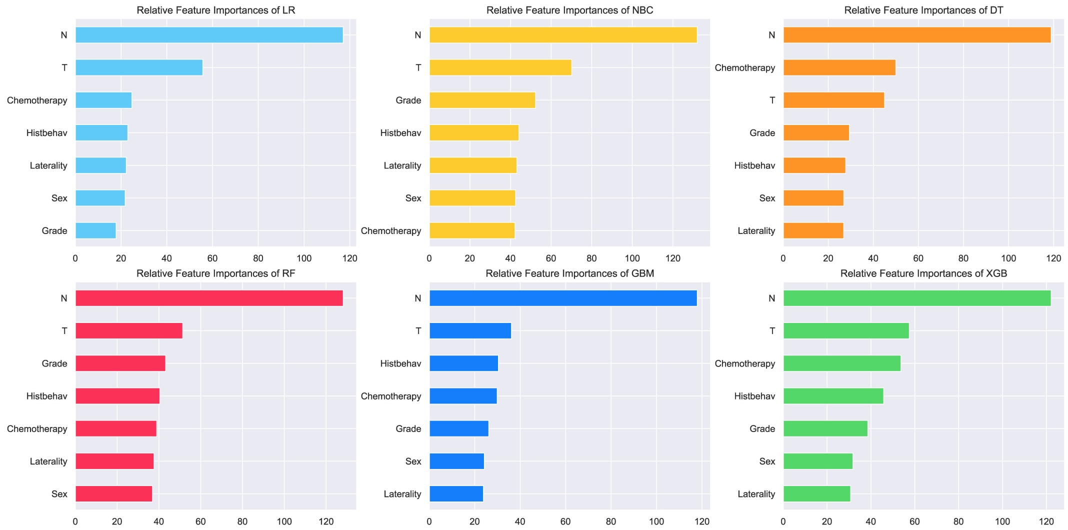


FIGURE 3
Feature importance of different models.

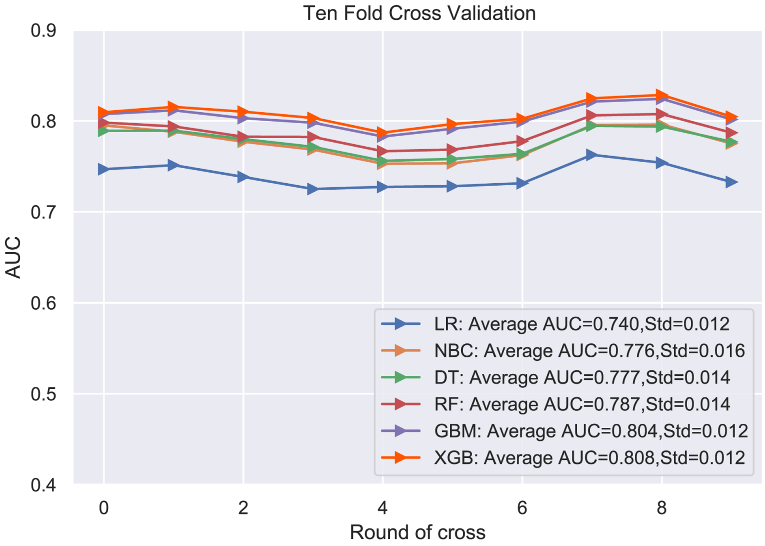


FIGURE 4
Ten-fold cross-validation results of different machine learning models.

Discussion

Lung cancer is one of the most common malignant tumors, which has an annual incidence of 2 million (1). Traditionally, lung cancer is divided into two types, including small cell lung cancer and NSCLC, with NSCLC accounting for 85% of cases (22). Compared with small cell carcinoma, NSCLC patients often has a better prognosis because of its slow growth and division (23). However, BM is considered an essential risk factor for the prognosis of NSCLC (24). Studies have reported that the incidence of BM in patients with NSCLC is 26-36% (3). However, the Lung Cancer National Comprehensive Cancer Network (NCCN) screening guidelines do not recommend routine bone imaging evaluation in asymptomatic patients (10). In order to identify lung cancer patients with high risk of BM, we innovatively constructed a clinical predictor based on an

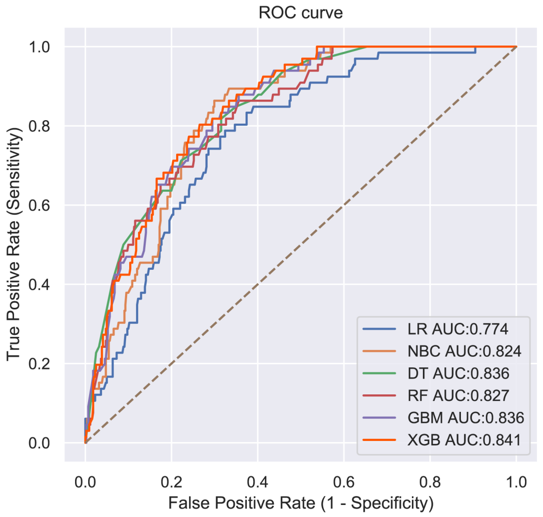


FIGURE 5
The roc curves of different machine learning models in external test set.

advanced ML algorithm (XGB). Figure 8 shows the construction process of the clinical predictor and the outcomes of the NSCLC with and without the predictor.

Artificial intelligence (AI) is a field of research in which computers are applied to mimic human intelligence which has been successfully applied in various fields, including driverless driving, face recognition and music creation (25–27). ML is a subfield of AI that focuses on developing algorithms to learn from data (11). Therefore, the emergence of electronic medical records (28) (EMR) has created a huge amount of analyzable data in the medical field, which provides the potential for the development of ML in the medical field. In biomedicine, there have been many studies using ML algorithms to guide clinical diagnosis and treatment, Including COVID-19 and cancer metastases field (19, 29). Statistical review of ML in medical field by Kaustubh Arun Bhavsar et al. suggested that ML techniques can help clinicians make better clinical decisions and improve patient care and overall health (30).

In this study, we compared the performance of six different algorithms and found that XGB algorithm perform best. XGB algorithm is an efficient, flexible and scalable machine learning algorithm classifier that has been widely used in the medical filed, such as COVID-19, Chronic Kidney Disease Diagnosis and BM of prostate cancer (PCa) (19, 31, 32). Liu et al (19) compared the diagnostic ability of six algorithm models, including XGB, to predict BM of PCa, and found that XGB model performed best. The consistency of the results indicates that the XGB algorithm has great potential in medical applications. Compared with the traditional logistic regression model, XGB model can process big data efficiently and has higher accuracy (33). Table 4 shows the strengths and weaknesses of the previous model as well as the proposed model (33).

In this study, logistic regression analysis helped us screen out seven independent risk factors for BM in NSCLC. And, through various ML algorithm verification, it is found that all features have essential contributions in the process of predicting BM, which is in high agreement with the logistic regression analysis.

In previous studies (34–36,) T stage, N stage and grade are considered risk factors for BM of cancer. Jie et al. found that (37) patients are more likely to have metastases, if they had higher N stage at diagnoses. In this research, we also found that lymphatic metastasis promotes BM in NSCLC patients and with increasing N stage, the risk of BM in NSCLC patients also increases. Fan et al. (35) found that T stage and tumor grade affected BM of renal cell carcinoma. In this research, we also found that T stage and tumor grade play an important role in predicting the development of BM in NSCLC patients. And the BM of NSCLC was related to the higher tumor grade and advanced T stage.

Interestingly, BM in NSCLC are more likely to occur in patients who have been treated with chemotherapy drugs in this study, which may be due to the misuse of chemotherapy drugs, resulting in their toxic effects on normal cells and promoting the proliferation of tumor cells (38, 39). Hui et al. (40) found that adenocarcinoma was related to a high risk of BM. In this study, we found that tumor histological type is an important character that affects BM. And adenocarcinoma, the most common tumor type of NSCLC, is more prone to BM than others. In addition, sex and the laterality of the primary tumor site can also affect BM of NSCLC patients. Many studies have shown that gender affects the development of tumors (41, 42). In this study, we found that male at a higher risk of bone metastases than females. Moreover,

TABLE 3 Prediction performance of different models.

	LR	NBC	DT	RF	GBM	XGB
Auc	0.774	0.824	0.836	0.827	0.836	0.841
Accuracy	0.694	0.738	0.698	0.734	0.720	0.744
Sensitivity	0.687	0.728	0.685	0.732	0.705	0.735
Specificity	0.742	0.803	0.788	0.742	0.818	0.803
F-score	0.387	0.444	0.405	0.420	0.432	0.449

LR, logistic regression; NBC, Naive Bayes classifiers; DT, decision tree; RF, random forest; GBM, Gradient Boosting Machine; XGB, eXtreme gradient boosting.

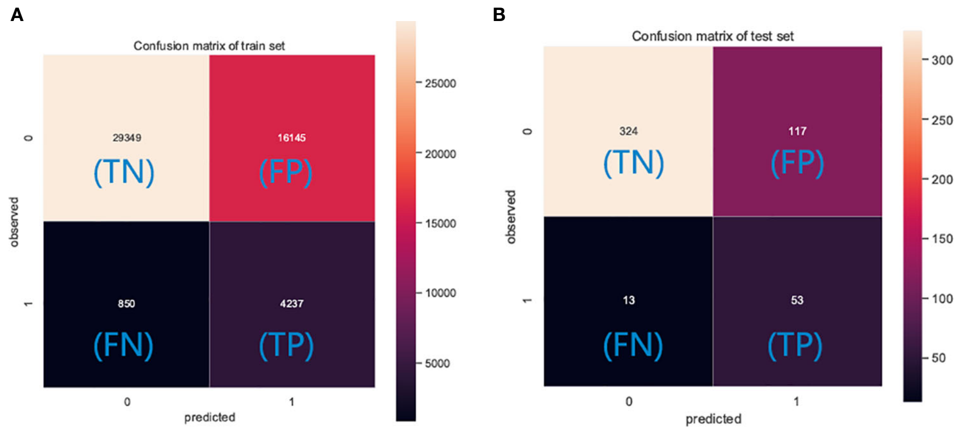
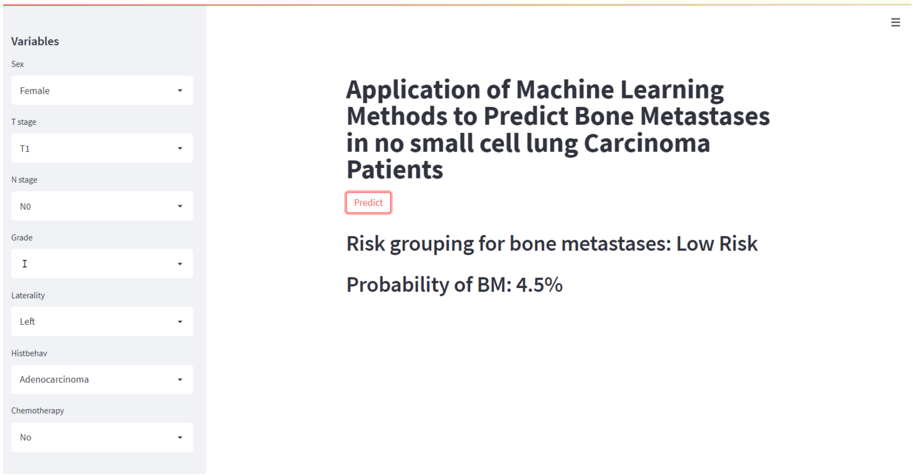


FIGURE 6
The confusion matrix of the XGB model in the **(A)** training set and the **(B)** test set. TP, true positive; TN, true negative; FP, false positive; FN, false negative.



Variables

Sex

Female

T stage

T1

N stage

N0

Grade

I

Laterality

Left

Histbehav

Adenocarcinoma

Chemotherapy

No

Application of Machine Learning Methods to Predict Bone Metastases in no small cell lung Carcinoma Patients

Predict

Risk grouping for bone metastases: Low Risk

Probability of BM: 4.5%

FIGURE 7
A web predictor for predicting bone metastases in no small cell lung carcinoma patients.

we found that left primary lung cancer is more likely to have bone metastases than right, which may be associated with the left lung being close to the heart, leading to more hematogenous metastasis of the left lung tumor.

In this study, we constructed a predictor based XGB algorithm with SEER data to predict BM in NSCLC. This research can help clinicians make better clinical decisions and promote the integration of medicine and machine learning. Meanwhile, this study has some limitations. First, Current machine learning is almost entirely statistical or black-box, bring severe theoretical limitations to its performance (43). Second, we cannot comment on which chemotherapeutic agents affect NSCLC BM because the SEER database does not record the treatment regimen and dosage of chemotherapy patients. Third, most of the variables in SEER database are clinical, which limits the accuracy of model prediction to some extent. Fourth, important parameters closely related to lung cancer such as dust, smoking, passive smokers, Tobacco chewing, and alcohol are missing in the SEER database, resulting in the failure of these parameters to be included in the predictor. In the future, with the continuous improvement of the database, we will incorporate more correlation parameters associated with the BM of NSCLC into the web predictor to improve its adaptability.

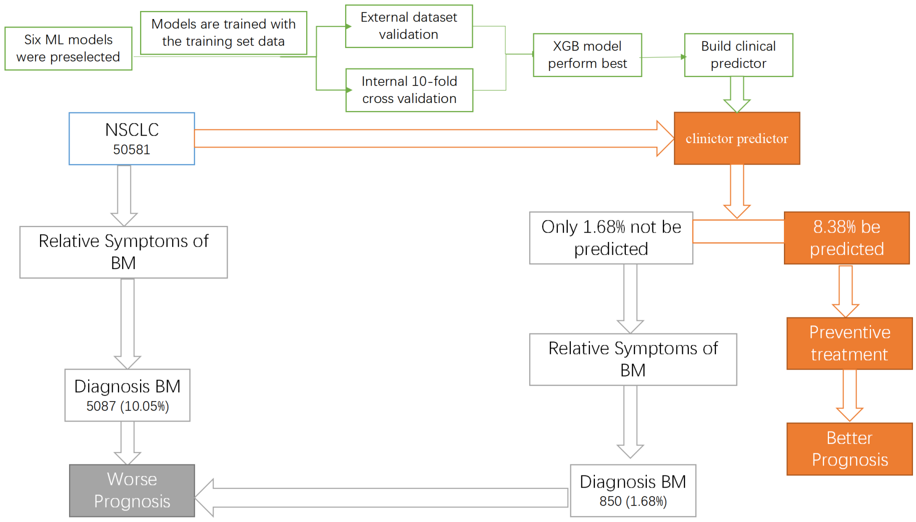


FIGURE 8
The construction process of the clinical predictor and the outcomes of the NSCLC with and without the predictor.

TABLE 4 The strength and weakness of the previous model and the proposed model.

	Strength	Weakness
XGB model	• Regularization prevents overfitting • High precision • Evaluating feature importance • Automatically handle missing values • used to solve classification problems or regression problems	• High space complexity • High storage resources for classification
Logistic regression model	• simple • Classification has very small computation and low storage resources • highly interpretable	• It is easy to underfit, and the general accuracy is not too high • Does not handle a large number of multi-class features or variables well • For nonlinear features, a transformation is required • Logistic regression does not perform very well when the feature space is large

In conclusion, in this study, we found that XGB algorithm performed best in six different algorithms and then as a tool build a web predictor for predicting BM of NSCLS which was accurate, simple and convenient to operate. This web predictor can predict BM of NSCLS easily and assist clinicians in diagnosis and making better clinical decisions for NSCLS patients.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Ethics statement

We received permission to access the research data file in the SEER program from the National Cancer Institute, US. Approval was waived by the local ethics committee, as SEER data is publicly available and de-identified. This study was approved by the Ethics Committee of the First Affiliated Hospital of Nanchang University, and cases from the First Affiliated Hospital of Nanchang University signed a written informed consent form. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

MPL and WCL conceived of and designed the study. MPL, WCL, BLS, and NSZ performed analysis and generated the figures and tables. MPL and WCL wrote the manuscript, and ZLL, SHH, ZHZ and JML critically reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work is supported by the "Double Thousand Plan" Talent Project of Jiangxi Province and the central government guides local funds for scientific and technological development, China (No. 20222ZDH04095).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS. Lung cancer. *Lancet* (2021) 398:535–54. doi: 10.1016/s0140-6736(21)00312-3

2. Xie X, Zhan C, Wang J, Zeng F, Wu S. An activatable nano-prodrug for treating tyrosine-Kinase-Inhibitor-Resistant non-small cell lung cancer and for optoacoustic and fluorescent imaging. *Small* (2020) 16:e2003451. doi: 10.1002/smll.202003451

3. Jung JM, Hyun SJ, Kim KJ. Surgical impacts of metastatic non-small cell lung cancer to the thoracic and lumbar spine. *J Korean Med Sci* (2021) 36:e52. doi: 10.3346/jkms.2021.36.e52

4. Fornetti J, Welm AL, Stewart SA. Understanding the bone in cancer metastasis. *J Bone Miner Res* (2018) 33:2099–113. doi: 10.1002/jbmr.3618

5. Wood SL, Brown JE. Skeletal metastasis in renal cell carcinoma: Current and future management options. *Cancer Treat Rev* (2012) 38:284–91. doi: 10.1016/j.ctrv.2011.06.011

6. Peterson LM, O'Sullivan J, Wu QV, Novakova-Jiresova A, Jenkins I, Lee JH, et al. Prospective study of serial (18)F-FDG PET and (18)F-fluoride PET to predict time to skeletal-related events, time to progression, and survival in patients with bone-dominant metastatic breast cancer. *J Nucl Med* (2018) 59:1823–30. doi: 10.2967/jnumed.118.211102

7. Farooki A, Leung V, Tala H, Tuttle RM. Skeletal-related events due to bone metastases from differentiated thyroid cancer. *J Clin Endocrinol Metab* (2012) 97:2433–9. doi: 10.1210/jc.2012-1169

8. Zhou Y, Yu Q-F, Peng A-F, Tong W-L, Liu J-M, Liu Z-L. The risk factors of bone metastases in patients with lung cancer. *Sci Rep* (2017) 7:8970. doi: 10.1038/s41598-017-09650-y

9. Niu Y, Lin Y, Pang H, Shen W, Liu L, Zhang H. Risk factors for bone metastasis in patients with primary lung cancer: A systematic review. *Med (Baltimore)* (2019) 98:e14084. doi: 10.1097/md.0000000000014084

10. Zhang C, Mao M, Guo X, Cui P, Zhang L, Xu Y, et al. Nomogram based on homogeneous and heterogeneous associated factors for predicting bone metastases in patients with different histological types of lung cancer. *BMC Cancer* (2019) 19:238. doi: 10.1186/s12885-019-5445-3

11. Deo RC. Machine learning in medicine. *Circulation* (2015) 132:1920–30. doi: 10.1161/circulationaha.115.001593

12. Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. *Cell* (2020) 181:92–101. doi: 10.1016/j.cell.2020.03.022

13. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* (2018) 362:1140–4. doi: 10.1126/science.aar6404

14. Aarestrup FM, Albeyatti A, Armitage WJ, Auffray C, Augello L, Balling R, et al. Towards a European health research and innovation cloud (HRIC). *Genome Med* (2020) 12:18. doi: 10.1186/s13073-020-0713-z

15. Zhuang Y, Chen YW, Shae ZY, Shyu CR. Generalizable layered blockchain architecture for health care applications: Development, case studies, and evaluation. *J Med Internet Res* (2020) 22:e19029. doi: 10.2196/19029

16. Shilo S, Rossman H, Segal E. Axes of a revolution: Challenges and promises of big data in healthcare. *Nat Med* (2020) 26:29–38. doi: 10.1038/s41591-019-0727-5

17. Solihah B, Azhari A, Musdholifah A. Enhancement of conformational b-cell epitope prediction using CluSMOTE. *PeerJ Comput Sci* (2020) 6:e275. doi: 10.7717/peerj-cs.275

18. Tian H, Ning Z, Zong Z, Liu J, Hu C, Ying H, et al. Application of machine learning algorithms to predict lymph node metastasis in early gastric cancer. *Front Med (Lausanne)* (2021) 8:759013. doi: 10.3389/fmed.2021.759013

19. Liu W-C, Li M-X, Qian W-X, Luo Z-W, Liao W-J, Liu Z-L, et al. Application of machine learning techniques to predict bone metastasis in patients with prostate cancer. *Cancer Manag Res* (2021) 13:8723–36. doi: 10.2147/cmar.S330591

20. Millman A, Dar D, Shamir M, Sorek R. Computational prediction of regulatory, premature transcription termination in bacteria. *Nucleic Acids Res* (2017) 45:886–93. doi: 10.1093/nar/gkw749

21. Metwaly A, Dunkel A, Waldschmitt N, Raj ACD, Lagkouvardos I, Corraliza AM, et al. Integrated microbiota and metabolite profiles link crohn's disease to sulfur metabolism. *Nat Commun* (2020) 11:4322. doi: 10.1038/s41467-020-17956-1

22. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin* (2017) 67:7–30. doi: 10.3322/caac.21387

23. Rice SJ, Belani CP. Diversity and heterogeneity of immune states in non-small cell lung cancer and small cell lung cancer. *PloS One* (2021) 16:e0260988. doi: 10.1371/journal.pone.0260988

24. Hong S, Youk T, Lee SJ, Kim KM, Vajdic CM. Bone metastasis and skeletal-related events in patients with solid cancer: A Korean nationwide health insurance database study. *PloS One* (2020) 15:e0234927. doi: 10.1371/journal.pone.0234927

25. Khatua A, Khatua A, Chi X, Cambria A. Artificial intelligence, social media and supply chain management: The way forward. *Electronics* (2021) 10(19):2348. doi: 10.3390/electronics10192348

26. Molas G, Nowak E. Advances in emerging memory technologies: From data storage to artificial intelligence. *Appl Sci* (2021) 11(23):11254. doi: 10.3390/app112311254

27. Kikon A, Deka PC. Artificial intelligence application in drought assessment, monitoring and forecasting: A review. *Stochastic Environ Res Risk Assess* (2021), 1–18. doi: 10.1007/s00477-021-02129-3

28. Christodouleas DC, Kaur B, Chorti P. From point-of-Care testing to eHealth diagnostic devices (eDiagnostics). *ACS Cent Sci* (2018) 4:1600–16. doi: 10.1021/acscentsci.8b00625

29. Shrivastav LK, Jha SK. A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. *Appl Intell (Dordr)* (2021) 51:2727–39. doi: 10.1007/s10489-020-01997-6

30. Bhavsar KA, Singla J., Al-Otaibi Y. D., Song O. Y., Zikria Y. B., Bashir A. K. Medical diagnosis using machine learning: A statistical review. *Comput Mater Contin* (2021) 67:107–25. doi: 10.32604/cmc.2021.014604

31. Ogunleye A, Wang QG. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans Comput Biol Bioinform* (2020) 17:2131–40. doi: 10.1109/tcbb.2019.2911071

32. Guan X, Zhang B, Fu M, Li M, Yuan X, Zhu Y, et al. Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: Results from a retrospective cohort study. *Ann Med* (2021) 53:257–66. doi: 10.1080/07853890.2020.1868564

33. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), 785–94. doi: 10.1145/2939672.2939785

34. Hu C, Yang J, Huang Z, Liu C, Lin Y, Tong Y, et al. Diagnostic and prognostic nomograms for bone metastasis in hepatocellular carcinoma. *BMC Cancer* (2020) 20:494. doi: 10.1186/s12885-020-06995-y

35. Fan Z, Huang Z, Huang X. Bone metastasis in renal cell carcinoma patients: Risk and prognostic factors and nomograms. *J Oncol* (2021) 2021:5575295. doi: 10.1155/2021/5575295

36. Ma X, Fan Y, Chen Z, Zhang Y, Wang S, Yu J. Blood biomarkers of bone metastasis in digestive tract malignant tumors. *Future Oncol* (2021) 17:1507–18. doi: 10.2217/fon-2020-0509

37. Yang J, Peng A, Wang B, Gusdon AM, Sun X, Jiang G, et al. The prognostic impact of lymph node metastasis in patients with non-small cell lung cancer and distant organ metastasis. *Clin Exp Metastasis* (2019) 36:457–66. doi: 10.1007/s10585-019-09985-y

38. Gadgeel SM. Role of chemotherapy and targeted therapy in early-stage non-small cell lung cancer. *Am Soc Clin Oncol Educ Book* (2017) 37:630–9. doi: 10.1200/edbk_175188

39. Kong P, Yan J, Liu D, Ji Y, Wang Y, Zhuang J, et al. Skeletal-related events and overall survival of patients with bone metastasis from nonsmall cell lung cancer - a retrospective analysis. *Med (Baltimore)* (2017) 96:e9327. doi: 10.1097/md.0000000000009327

40. Wang H, Zhang Y, Zhu H, Yu J. Risk factors for bone metastasis in completely resected non-small-cell lung cancer. *Future Oncol* (2017) 13:695–704. doi: 10.2217/fon-2016-0237

41. Willenbrink TJ, Ruiz ES, Cornejo CM, Schmults CD, Arron ST, Jambusaria-Pahlajani A. Field cancerization: Definition, epidemiology, risk factors, and outcomes. *J Am Acad Dermatol* (2020) 83:709–17. doi: 10.1016/j.jaad.2020.03.126

42. Haupt S, Caramia F, Klein SL, Rubin JB, Haupt Y. Sex disparities matter in cancer development and therapy. *Nat Rev Cancer* (2021) 21:393–407. doi: 10.1038/s41568-021-00348-y

43. Azodi CB, Tang J, Shiu SH. Opening the black box: Interpretable machine learning for geneticists. *Trends Genet* (2020) 36:442–55. doi: 10.1016/j.tig.2020.03.005