scientific reports

Check for updates

OPEN

Machine learning-based prediction model for brain metastasis in patients with extensive-stage small cell lung cancer

Erha Munai[1,3], Siwei Zeng[1,3], Ze Yuan[2], Dingyi Yang[2], Yong Jiang[2], Qiang Wang[2], Yongzhong Wu[2]✉, Yunyun Zhang[2]✉ & Dan Tao[2]✉

Brain metastases (BMs) in extensive-stage small cell lung cancer (ES-SCLC) are often associated with poor survival rates and quality of life, making the timely identification of high-risk patients for BMs in ES-SCLC crucial. Patients diagnosed with ES-SCLC between 2010 and 2018 were screened from the Surveillance, Epidemiology, and End Results (SEER) database. Four different machine learning (ML) algorithms were used to create prediction models for BMs in ES-SCLC patients. The accuracy, sensitivity, specificity, AUROC, and AUPRC were compared among these models and traditional logistic regression (LR). The random forest (RF) model demonstrated the best performance and was chosen for further analysis. The AUROC and AUPRC were calculated and compared. The findings from the RF model were utilized to identify the risk factors linked to BMs in patients diagnosed with ES-SCLC. Examining 4,716 instances of ES-SCLC, the research conducted an analysis, with brain metastases arising in 1,900 cases. Through evaluation of the ROC curve and PRC concerning the RF Model, results depicted an AUROC of 0.896 (95% CI: 0.889–0.899) and AUPRC of 0.900 (95% CI: 0.895–0.904). Test accuracy measured at 0.810 (95% CI: 0.784–0.833), sensitivity at 0.797 (95% CI: 0.756–0.841), and specificity at 0.819 (95% CI: 0.754–0.879). Based on the SHAP analysis of the RF predictive model, the top 10 most relevant features were identified and ranked in order of relative importance: bone metastasis, liver metastasis, radiation, age, tumor size, primary tumor location, N-stage, race, T-stage, and chemotherapy. The research developed and validated a predictive RF model using clinical and pathological data to predict the risk of BMs in patients with ES-SCLC. This model may assist physicians in making clinical decisions that could delay the onset of BMs and improve patient survival rates.

Keywords Machine learning, Random Forest, Extensive-stage small cell lung cancer, Brain metastasis

Small-cell lung cancer (SCLC) accounts for approximately 14% of lung cancer cases, and nearly two-thirds of SCLC patients develop extensive-stage disease (ES-SCLC)[1,2]. More than half of ES-SCLC patients experience brain metastasis(BMs)[3]. Four to six cycles of chemotherapy containing cisplatin can achieve a clinical response rate of 60–70% in ES-SCLC patients, but the median survival period is only around 9 months[4–6]. Studies have shown that the development of BMs, especially when symptomatic, may severely impair the quality of life of ES-SCLC patients, resulting in a median survival of only 4 to 6 months[7]. Therefore, early assessment and identification of patients at high risk for BMs in ES-SCLC, followed by a comprehensive evaluation considering the patient's overall condition, such as physical health and economic status, and timely implementation of appropriate interventions, can contribute to improving patient outcomes.

A clinical predictive model, which evaluates the risk of disease and the effectiveness of treatment, is an essential component of contemporary clinical practice[8]. Machine learning (ML) is an application of artificial intelligence that relies on data to automatically learn and improve without the need for explicit programming. Compared to traditional independent risk factor assessments for predicting cancer metastasis in patients, machine learning offers higher accuracy in predicting and diagnosing cancer metastasis. ML can autonomously identify new variables and their complex relationships within a dataset. Its application in healthcare is experiencing rapid

[1]School of Medicine, Chongqing University, Chongqing, China. [2]Department of Radiation Oncology, Chongqing University Cancer Hospital, Chongqing, China. [3]Erha Munai and Siwei Zeng contributed equally as the first authors. ✉email: cqmdwyz@163.com; 1120303669@qq.com; taodan@cqu.edu.cn

growth and is increasingly utilized to develop new prognostic models for various diseases[9]. A study has indicated that the healthcare burden is skyrocketing, which is associated with the lack of early prevention and treatment for patients at risk of developing BMs in SCLC[10]. Machine learning is particularly well-suited for utilizing the growing big data and constantly enhancing computational capacities. This allows for the possibility of conducting large-scale analyses in a more manageable and simplified manner[11].

This study aimed to comprehensively investigate the risk factors for BMs in ES-SCLC patients using population-based surveillance, epidemiology, and outcomes (SEER) databases. We have employed innovative ML techniques to build a predictive model for assessing the risk of brain metastasis in ES-SCLC patients. By predicting the probability of BMs in ES-SCLC patients, clinicians could make more informed and effective decisions, to minimize or delay the occurrence of BMs to the greatest extent possible.

Methods

Data source and study population

The SEER database, comprising 18 cancer registries in the United States, provided data on cancer incidence, therapy, and survival for around 30% of the American population. Demographic factors like income, location, age at diagnosis, ethnicity, and sex, as well as cancer-specific factors including tumor location, histological subtypes, and sites of metastasis, were included in the SEER database. For this study, we extracted information on patients histologically diagnosed with SCLC between 2010 and 2018 from the SEER database. The inclusion criteria for this study were as follows: (1) Patients who had a confirmed histological diagnosis of ES-SCLC; (2) Patients who had complete information on TNM stage and other important characteristics; (3) Patients who had complete follow-up. The exclusion criteria were as follows: (1)Patients diagnosed at autopsy or death certificate only; (2) Not the first primary malignancy; (3) TN stage was not available; (4) Patients with limited-stage small cell lung cancer; (5) Incomplete records of the patient's other information. This study was conducted using publicly available de-identified data from the SEER database. No specific ethical approval or informed consent was required for this analysis.

Data preprocessing

Based on our research objectives, we extracted 36 indicators from the SEER database related to the metastasis of ES-SCLC. We filtered the necessary data, eliminating any missing or abnormal entries. Text data underwent encoding as part of the data cleaning process. Additionally, the data was standardized, normalized, and discretized to enhance analysis[12]. Various processing techniques were implemented on various factors, including tumor dimensions and age groups. Discriminating against non-significant variables or highly duplicative ones led to the identification of the most informative features. To equalize the imbalanced dataset, the resampling method was employed to ensure uniformity between the two categories. Ultimately, a total of 15 features were chosen for the model.

Model development

Various input variables were used to establish machine learning models in this research, such as age, gender, race, primary site of tumor, surgical procedures, radiation therapy, chemotherapy, marital status, bone metastases, liver metastases, lung metastases, laterality, tumor size, T stage, and N stage. A 7:3 split of the SEER cohort was determined using Python and the scikit-learn package[13]. In addition, random forests (RF)[14], AdaBoosts[15], extreme gradient boostings (XGB)[16], logistic regression (LR)[17], and support vector machine (SVM)[18] were used in the prediction process. Based on a development dataset, all models are built using 10-fold cross-validation and 50 iterations, with each iteration's samples drawn randomly from observational data[19].

Python was used to export predictive models based on machine learning, resulting in the creation of five models. These models varied in discrimination and accuracy across datasets. To determine the final model, the predictive performance of each model was assessed and compared on the testing dataset. Various metrics such as accuracy, sensitivity, specificity, as well as ROC and PRC curves were utilized to evaluate model performance. The significance of input variables in predicting BMs in ES-SCLC was determined by ranking the feature importance of input variables.

Evaluation methods

In order to assess the effectiveness of ML models, the performance of five different models was analyzed in terms of accuracy, sensitivity, specificity, AUROC, and AUPRC, with a confidence interval of 95%. The RF model was chosen as the optimal model for predicting BMs in ES-SCLC research after evaluating these metrics. Each model's AUROC was computed and compared to determine their generalization capabilities and clinical utility. In cases of imbalanced datasets, the AUPRC is a more reliable metric than the AUC for evaluating model performance. Thus, the Precision-Recall curve was plotted and the AUPRC was calculated to supplement the AUC value[20]. Additionally, SHAP, a cooperative game-theoretic-based technique, was utilized to provide explanations for predictions made by the best-performing ML model[21]. The statistical software employed for these analyses included R (v4.2.2) and Python (v3.9.13).

Result

Patient clinical characteristics

This study analyzed a total of 4,716 patients diagnosed with ES-SCLC between 2010 and 2018, using data from the SEER database. Among the 79,769 patients who did not meet the inclusion criteria, they were excluded from the study. The 4,716 ES-SCLC patients included in the analysis were divided into a training group ($n=3,301$)

and a testing group ($n = 1,415$) at a ratio of 7:3. Out of these patients, 2,816 had no brain metastases (non-BMs), while 1,900 patients had brain metastases(BMs) (Fig. 1).

Table 1 displays the clinical features of individuals with ES-SCLC in the BMs group and non-BMs group. The mean ages of the two groups were similar (mean (SD), 67 vs. 67, $P < 0.001$). Both cohorts predominantly consisted of patients aged ≥ 65 ($P < 0.001$). The prevalence of the White race was notably higher in both groups than other ethnicities ($P < 0.001$). In comparison to the non-BMs group, a greater percentage of patients in the BMs group underwent radiation therapy (75.7% vs. 35.2%, $P < 0.001$). Most patients were diagnosed while married ($P < 0.001$). The non-BMs group had a higher rate of metastases at DX-liver, DX-lung, and DX-bone compared to the BMs group. The proportion of T4 and N2 stages was relatively elevated in both groups ($P < 0.001$). No significant disparities were noted in sex, tumor location, surgery, or chemotherapy between the two cohorts.

Feature analysis of common clinical variables

Based on the SEER database, a statistical analysis was conducted on six common clinical variables of ES-SCLC patients included in the study (Fig. 2). The research findings indicate that BMs in ES-SCLC patients predominantly occurred between the ages of 50 and 80 ($P < 0.001$). Among patients within this age range, those around 65 and 70 years old were more prone to developing BMs, with a relatively higher risk (Fig. 2A). In the diagnosis of ES-SCLC, we have observed that when patients have marital statuses such as divorce or widowhood, the likelihood of BMs occurrence is higher ($P < 0.001$) (Fig. 2B). The risk of BMs in ES-SCLC patients was not statistically significantly correlated with primary tumor site ($P = 0.764$) or laterality ($P = 0.461$) (Fig. 2C–D). The risk of BMs was found to be similar between stage T3 and T4 stage ($P < 0.001$) ES-SCLC patients (Fig. 2E). Additionally, when considering the stage of lymph node metastasis, it was observed that the risk of BMs was slightly higher in stage N2 and N3 stage ES-SCLC patients compared to those in stage N2 ($P < 0.001$) (Fig. 2F).

Model performance

Based on the statistical analysis results, a total of 15 clinical variables were identified. Figure 3 illustrates the correlation analysis conducted among these clinical variables. The results indicated that there appears to be a correlation between BMs in ES-SCLC and factors such as race, surgery, radiation, and laterality. Clinical

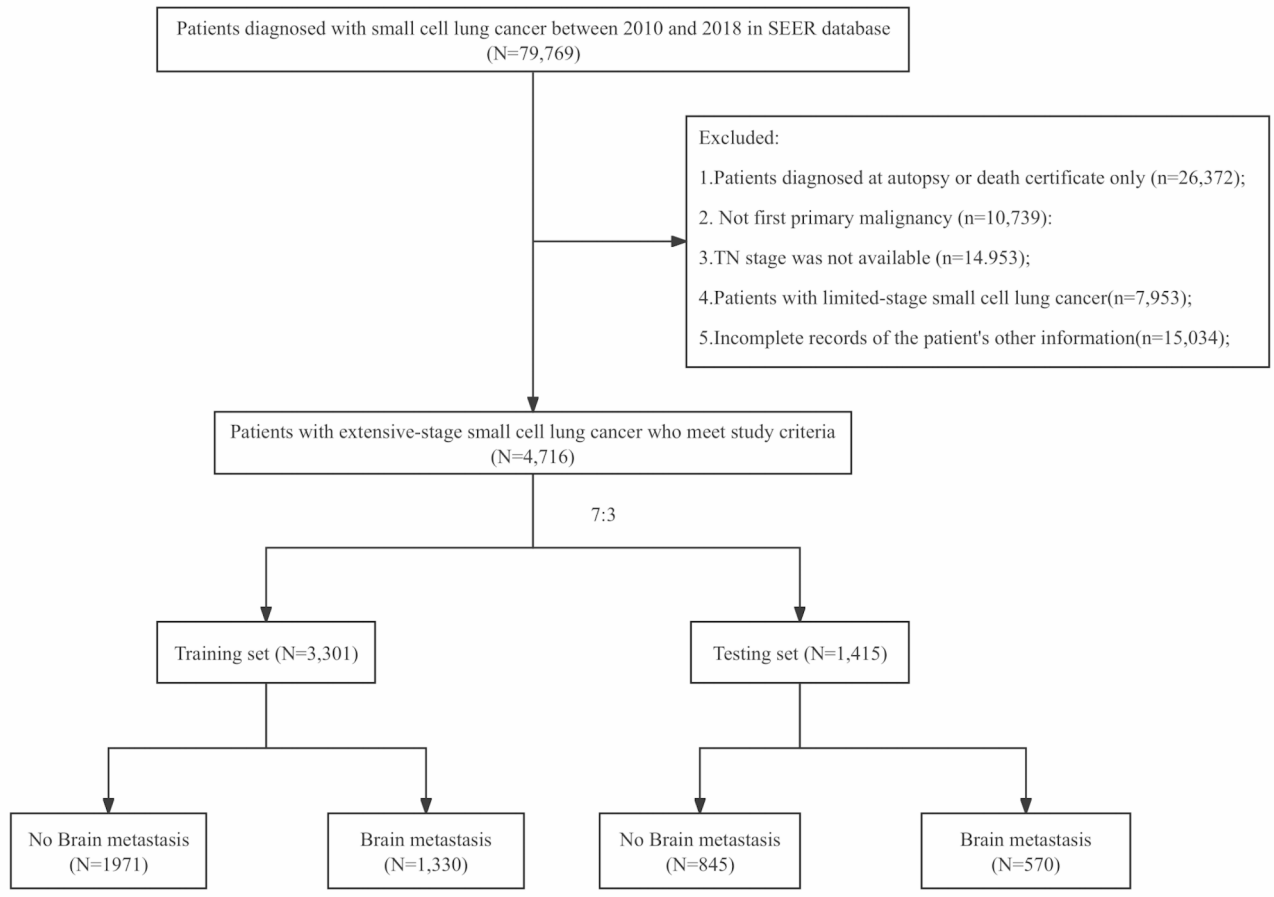

Fig. 1. The flowchart of patient selection from the SEER database. *SEER* Surveillance, Epidemiology, and End Results.

Characteristics, *n* (%)	No Brain metastasis (*N* = 2816)	Brain metastasis (*N* = 1900)	*P* Value
Age (years), median (range)	67(17–85)	67(37–85)	< 0.001
Age			< 0.001
< 65	993 (35.3%)	877 (46.2%)	
≥ 65	1823 (64.7%)	1023 (53.8%)	
Sex			0.545
Male	1555 (55.2%)	1067 (56.2%)	
Female	1261 (44.8%)	833 (43.8%)	
Race			< 0.001
White	2486 (88.3%)	1539 (81.0%)	
Black	234 (8.3%)	246 (12.9%)	
Asian or Pacific Islander	87 (3.1%)	102 (5.4%)	
American Indian/Alaska Native	9 (0.3%)	13 (0.7%)	
Primary.Site			0.764
Main bronchus	241 (8.6%)	149 (7.8%)	
Upper lobe, lung	1432 (50.9%)	1007 (53.0%)	
Middle lobe, lung	116 (4.1%)	79 (4.2%)	
Lower lobe, lung	675 (24.0%)	434 (22.8%)	
Overlapping lesion of lung	42 (1.5%)	30 (1.6%)	
Lung, NOS	310 (11.0%)	201 (10.6%)	
Surgery			0.872
No	178 (6.3%)	114 (6.0%)	
Yes	2631 (93.4%)	1782 (93.8%)	
Others	7 (0.2%)	4 (0.2%)	
Radiation			< 0.001
No	1808 (64.2%)	447 (23.5%)	
Yes	990 (35.2%)	1439 (75.7%)	
Others	18 (0.6%)	14 (0.7%)	
Chemotherapy			0.005
No	1051 (37.3%))	787 (41.4%)	
Yes	1765 (62.7%)	1113 (58.6%)	
Marital.status			< 0.001
Married	1465 (52.0%)	988 (52.0%)	
Single and Unmarried	370 (13.1%)	326 (17.2%)	
Others	981 (34.8%)	586 (30.8%)	
Mets.at.DX.bone			< 0.001
No	951 (33.8%)	1418 (74.6%)	
Yes	1865 (66.2%)	482 (25.4%)	
Mets.at.DX.liver			< 0.001
No	1140 (40.5%)	1453 (76.5%)	
Yes	1676 (59.5%)	447 (23.5%)	
Mets.at.DX.lung			0.123
No	2147 (76.2%)	1486 (78.2%)	
Yes	669 (23.8%)	414 (21.8%)	
Laterality			0.461
Left	1181 (41.9%)	793 (41.7%)	
Right	1576 (56.0%)	1062 (55.9%)	
Paired site	18 (0.6%)	20 (1.1%)	
Others	41 (1.5%)	25 (1.3%)	
T.stage			< 0.001
T1	220 (7.8%)	165 (8.7%)	
T2	811 (28.8%)	666 (35.1%)	
T3	150 (5.3%)	107 (5.6%)	
T4	1635 (58.1%)	962 (50.6%)	
N.stage			< 0.001
N0	405 (14.4%)	370 (19.5%)	
N1	265 (9.4%)	212 (11.2%)	

Continued

Characteristics, *n* (%)	No Brain metastasis (*N* = 2816)	Brain metastasis (*N* = 1900)	*P* Value
N2	1531 (54.4%)	951 (50.1%)	
N3	615 (21.8%)	367 (19.3%)	
Tumor.Size			< 0.001
Median(IQR)	50(33–71)	58(35–73)	

Table 1. Patient demographics and clinical characteristics.

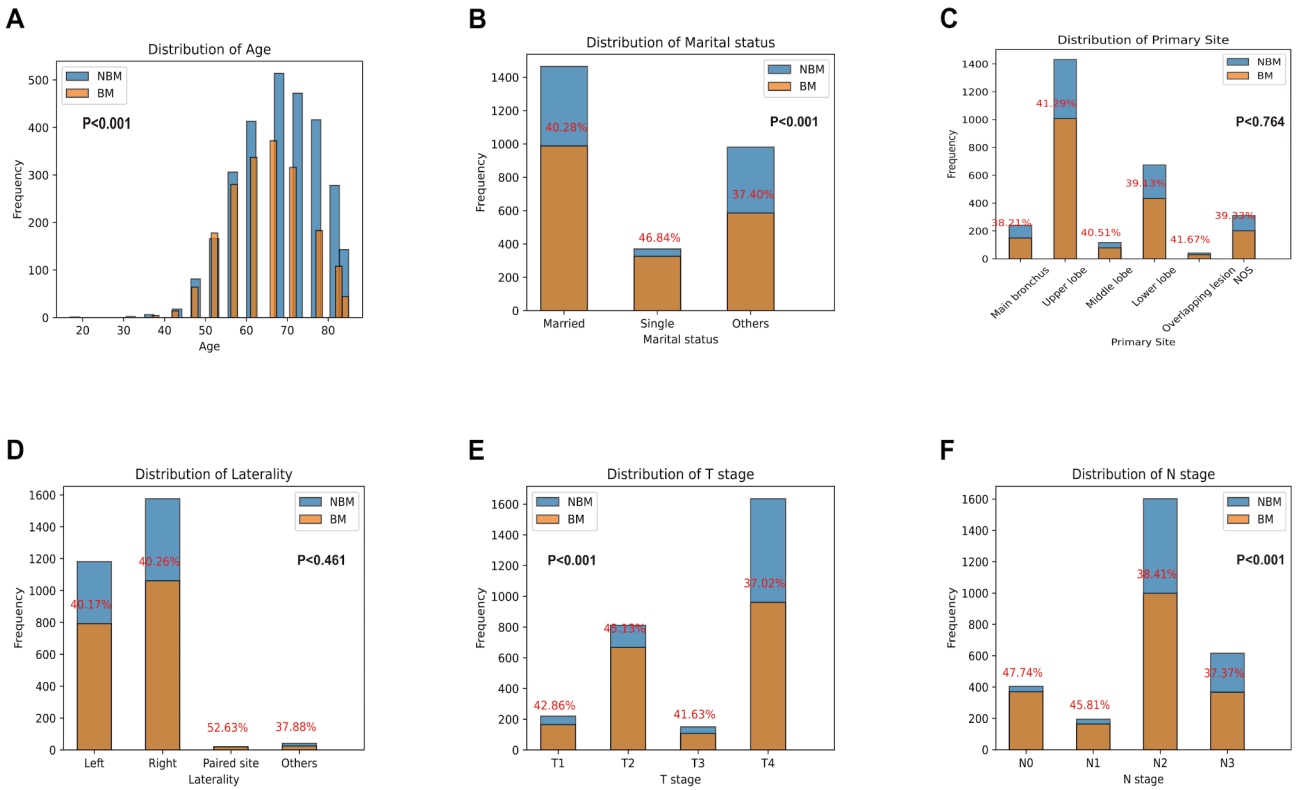

Fig. 2. Common clinical variable characteristics analysis of ES-SCLC is based on the SEER database.

variables associated with non-BMs cases might include age, sex, primary tumor site, chemotherapy, marital status, metastasis at DX-liver, metastasis at DX-lung, metastasis at DX-bone, tumor size, T stage, and N stage.

In order to evaluate the predictive abilities of the five models, a ten-fold cross-validation was conducted using the training dataset. Figure 4 illustrates the comparison of model performances, revealing that the RF model demonstrates superior results compared to other models in AUROC, AUPRC, sensitivity, accuracy, and specificity. Consequently, we suggest employing the RF model as the optimal classifier for predicting ES-SCLC brain metastasis.

The ROC curve and precision-recall curve for the RF model are illustrated in Fig. 5, displaying an AUROC of 0.896 (95% CI: 0.889–0.899) and an AUPRC of 0.900 (95% CI: 0.895–0.904). The RF model achieved a test accuracy score of 0.810 (95% CI: 0.784–0.833) along with a sensitivity of 0.797 (95% CI: 0.756–0.841) and a specificity of 0.819 (95% CI: 0.754–0.879). Moreover, the feasibility of the AdaBoost model in predicting BMs in ES-SCLC was also demonstrated through confusion matrix analysis (Supplemental Fig. 1).

The significance of factors in ML algorithms

SHAP was utilized in the analysis of the RF model in this study. Generally, higher SHAP values associated with features suggest a higher likelihood of the target event taking place. During the SHAP analysis, red is used to signify feature values that have a positive effect on the model, while blue is reserved for feature values that have a negative impact[22]. The results of the study revealed that out of the top ten feature variables, bone metastasis held the most significance, followed closely by liver metastasis, radiation, patient age, tumor size, primary tumor site, N-stage, ethnicity, T-stage, and the use of chemotherapy (Fig. 6).

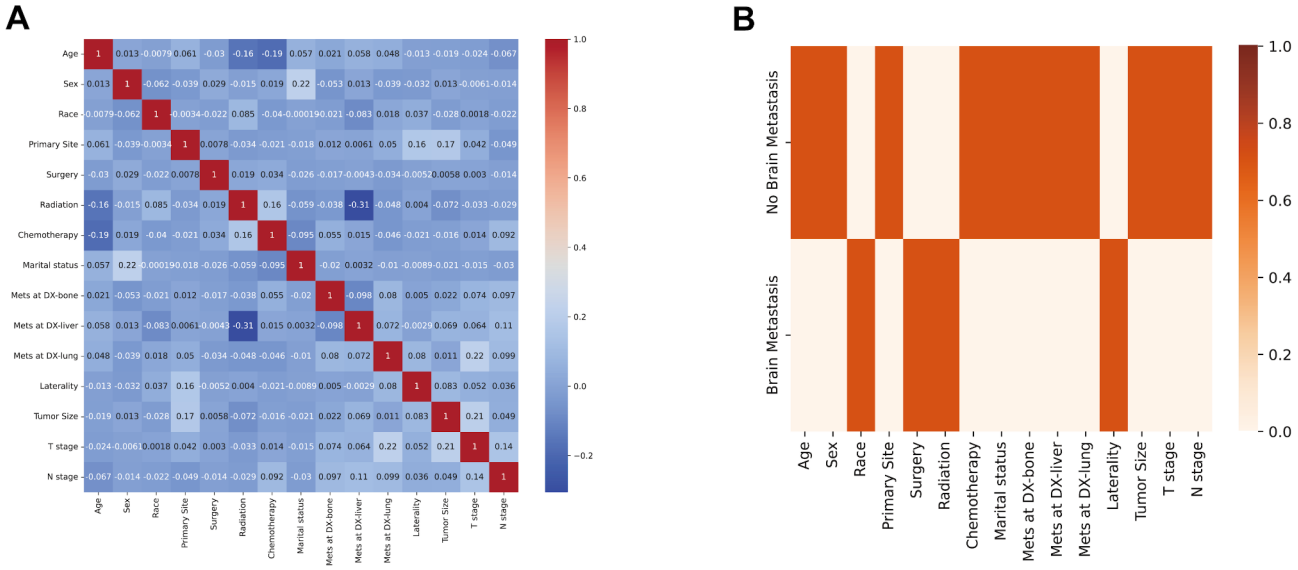

Fig. 3. Correlation analysis of clinical variables. (**A**) The correlation heatmap between clinical variables. (**B**) The correlation heatmap between ES-SCLC BMs or non-BMs and 15 clinical variables.

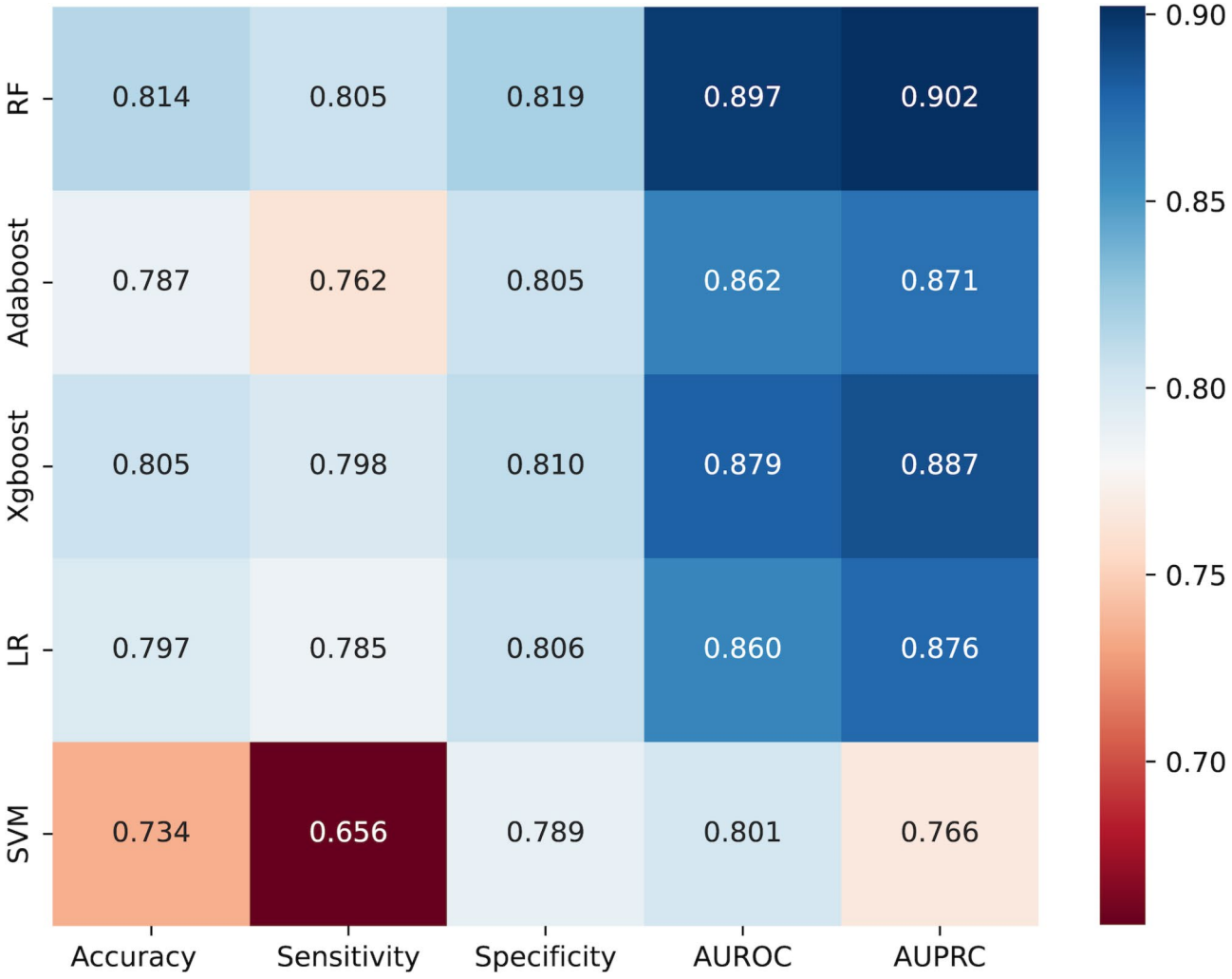

Fig. 4. Performance comparison of different models.

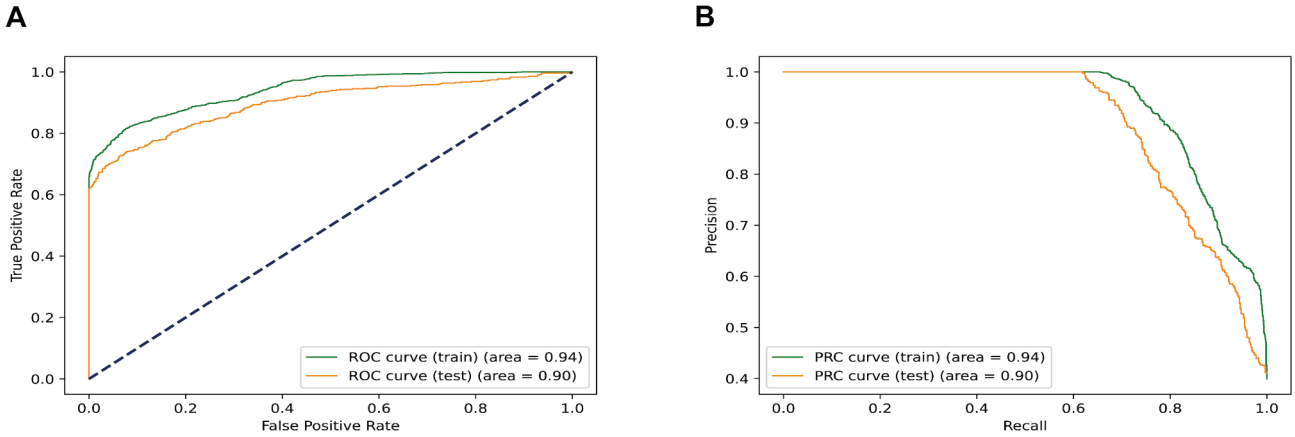

Fig. 5. The receiver operating characteristic curve and the precision-recall curve on the training and test dataset. (**A**) AUROC. (**B**) AUPRC.

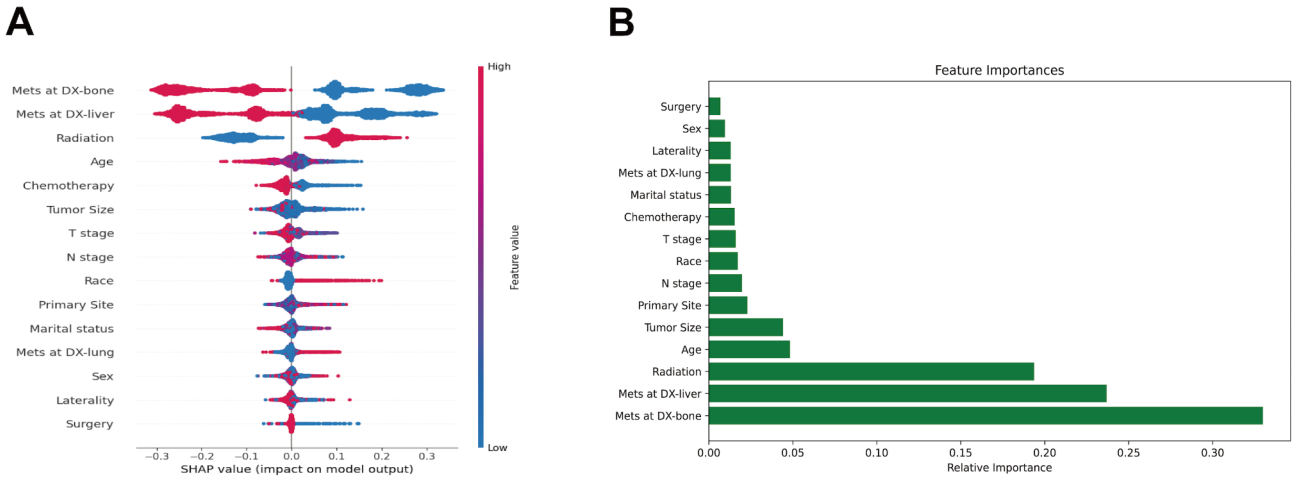

Fig. 6. Relative importance of variables based on SHAP for RF prediction model. SHAP, Shapley's Additive explanations.

Discussion

This study utilized data from the SEER database to gather information on extensive-stage small cell lung cancer(ES-SCLC) patients with or without brain metastasis (BMs). The 15 feature variables were carefully selected and determined, and four simple and practical machine learning (ML) models and traditional logistic regression (LR) were developed and validated. The performance of these models was compared in terms of accuracy, sensitivity, specificity, AUROC, and AUPRC. These research findings indicated that the Random forest (RF) model exhibited the highest predictive ability for BMs in ES-SCLC patients, as observed in both the training and testing models. The application of this model assisted clinical practitioners in accurately assessing the risk of BMs in ES-SCLC patients, enabling timely prophylactic cranial irradiation (PCI) treatment, thereby enhancing survival rates and delaying the reduction of BMs risk. Based on current understanding, this research was groundbreaking in utilizing machine learning algorithms to forecast the likelihood of BMs in ES-SCLC.

The RF is an ensemble learning method that employs bagging technology to combine multiple decision trees for tasks such as classification, regression, and others. During its training phase, it constructs and aggregates the outputs of numerous decision trees—producing a majority vote of classes for classification problems and an average of predicted values for regression problems[23]. This technique's primary benefit lies in its exceptional stability, effectively preventing overfitting in training[24], particularly beneficial for smaller datasets. Consequently, it enhances accuracy not only during testing but also when implementing the model on samples. While the prevailing notion suggested that random forest classifiers could yield satisfactory outcomes without systematic hyperparameter adjustments, we conducted hyperparameter optimization to both mitigate overfitting and enhance positive prediction accuracy.

Utilizing SHAP values, this research evaluated the influence of individual factors. In the visualization of SHAP for variable importance, it was observed that each of the 15 variables analyzed made unique contributions to the model. Feature importance analysis and ranking of the variables identified the top ten features associated

with BMs in ES-SCLC: bone metastasis, liver metastasis, radiation, age, tumor size, primary tumor location, N staging, race, T staging, and chemotherapy.

Relevant studies indicated that bone metastasis and liver metastasis are significant risk factors for SCLC brain metastasis[25–27]. This finding further corroborates our research results. In our study, bone metastasis and liver metastasis emerged as the two most notable variables for BMs in ES-SCLC. It was noteworthy that past research did not conduct distinctive analyses between limited-stage and extensive-stage patients, and we were the first to systematically explore the role of bone metastasis and liver metastasis as risk factors for BMs in ES-SCLC.

There is currently controversy over whether age was a risk factor for ES-SCLC-induced BMs. Multiple studies indicated that age is not a key risk factor for BMs in ES-SCLC patients[28–30]. However, in the systematic review and meta-analysis conducted by Zeng Haiyan et al.[31] regarding the risk factors for BMs in patients with SCLC (induced ES-SCLC), a comprehensive analysis of the conclusions from 14 studies revealed that age (< 65 years) is a risk factor for BMs. The explanation for this finding may be that younger SCLC patients typically have a longer lifespan[32,33], allowing for more time to experience BMs. A research led by et al.[34] delved into the predictive value of clinical features in relation to the occurrence and consequences of simultaneous brain metastases (BMs) in patients with small cell lung cancer (SCLC), revealing that older age was established as a notable risk factor for synchronous BMs in individuals with SCLC. This is consistent with our results. Age was the fourth factor associated with BMs in our study of ES-SCLC.

The size of the tumor was identified as another significant risk factor in the development of distant metastasis from malignant tumors. An analysis conducted by Zheng et al.[35] revealed that the size of the primary tumor can predict the occurrence of SCLC-BM. Research conducted by Chen et al.[36] regarding the factors associated with brain metastasis following PCI in limited-stage SCLC indicated that individuals with tumors ≥ 5 cm are at a higher risk of developing brain metastasis (Hazard Ratio: 1.781, 95% Confidence Interval: 1.044–3.039, $P = 0.034$). Similarly, a retrospective study by Farooqi AS et al.[37] showed that tumors ≥ 5 cm increase the risk of brain metastasis (HR 1.77, 95% CI 1.22–2.55, $P = 0.002$). Our research results are essentially consistent with theirs.

It was noteworthy that in our study, radiotherapy ranks third among factors associated with BMs, which appeared contradictory to general clinical knowledge. Considering that the SEER database did not comprehensively record patients' radiotherapy details, survival curve analysis of patients who received or did not receive radiotherapy reveals a significantly higher survival period for those who underwent radiotherapy ($P < 0.001$)(Supplemental Fig. 2). Radiotherapy has also been shown to provide benefits to patients with ES-SCLC in recent clinical studies. Several literature reports suggested that thoracic radiotherapy is an independent prognostic factor for improving the survival of ES-SCLC patients[38,39], indicating that thoracic radiotherapy may enhance overall survival. This aligns with our analyzed survival curve results, possibly because radiotherapy extended the survival time of ES-SCLC patients, allowing for more time to experience BMs.

Although there were previous reports on predictive models of BMs in SCLC based on the SEER database, these articles limited scope and only focused on the occurrence of BMs in SCLC. Therefore, we were the first to develop and validate a prediction model for BMs in ES-SCLC using the SEER database. Our study included up to 15 clinically available variables, which were used in both the development and validation of the model. Additionally, all predictors related to BMs in ES-SCLC were obtained through sequencing.

While our study had its strengths, there were also some limitations. Firstly, it was a retrospective study, which came with the inherent data biases typical of retrospective research. Secondly, due to a significant amount of missing data for some variables in multiple imputations within the SEER dataset, we choose to exclude these missing data in the manuscript, potentially introducing bias to the results. Lastly, the absence of key variables in the SEER database, such as KPS, ECOG scores, blood biochemical indicators, and detailed survival times, limits further optimization of our model. In future research, we plan to address these limitations by incorporating additional clinical factors to enhance the predictive power of the model, thereby providing more comprehensive support to healthcare professionals in their decision-making.

In summary, five ML models were created to forecast BMs in ES-SCLC. Among these models, it was observed that the RF model showcased the most reliable predictive capacity, showcasing outstanding discriminative performance not solely in the evaluation and validation sets, but also attaining the maximum levels of AUROC, AUPRC, sensitivity, specificity, and accuracy. We hope the RF model-based web calculator can help clinicians identify patients at high risk of BMs from ES-SCLC, enabling early and effective interventions to prevent and delay the onset of BMs and improve patient survival. We hope that the RF model-based online calculator will help clinicians identify high-risk ES-SCLC patients for BMs, facilitating the selection of early and appropriate interventions, such as Prophylactic Cranial Irradiation or Brain MRI surveillance. These early interventions aim to prevent or delay BMs and improve patient survival.

Data availability

The dataset analysed during the current study are available in the Surveillance, Epidemiology, and End Results (SEER) database (https://seer.cancer.gov/). All data analysed during this study are included in the supplementa ry file.

Received: 16 April 2024; Accepted: 19 November 2024
Published online: 20 November 2024

References

1. Chen, W. et al. Cancer statistics in China, 2015. *CA Cancer J. Clin.* **66**(2), 115–132. https://doi.org/10.3322/caac.21338 (2016).

2. Jett, J. R. et al. *Treatment of small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines.* Chest, 143(5 Suppl): p. e400S-e419S. DOI: https://doi.org/10.1378/chest.12-2363 (2013).
3. Seute, T. et al. Neurologic disorders in 432 consecutive patients with small cell lung carcinoma. *Cancer* **100**(4), 801–806. https://doi.org/10.1002/cncr.20043 (2004).
4. Chen, Y., Chen, L. & Zhong, D. Comparing the adverse effects of platinum in combination with etoposide or irinotecan in previously untreated small-cell lung cancer patients with extensive disease: a network meta-analyses. *Thorac. Cancer* **8**(3), 170–180. https://doi.org/10.1111/1759-7714.12420 (2017).
5. Aupérin, A. et al. Prophylactic cranial irradiation for patients with small-cell lung cancer in complete remission. Prophylactic cranial irradiation overview Collaborative Group. *N Engl. J. Med.* **341**(7), 476–484. https://doi.org/10.1056/nejm199908123410703 (1999).
6. Guo, S., Liang, Y. & Zhou, Q. Complement and correction for meta-analysis of patients with extensive-stage small cell lung cancer managed with irinotecan/cisplatin versus etoposide/cisplatin as first-line chemotherapy. *J. Thorac. Oncol.* **6**(2), 406–408. https://doi.org/10.1097/JTO.0b013e3182061d8c (2011). Author reply 408.
7. Greenspoon, J. N. et al. Selecting patients with extensive-stage small cell lung cancer for prophylactic cranial irradiation by predicting brain metastases. *J. Thorac. Oncol.* **6**(4), 808–812. https://doi.org/10.1097/JTO.0b013e31820d782d (2011).
8. Moons, K. G. et al. *Prognosis and prognostic research: what, why, and how?* Bmj, 338: p. b375. DOI: https://doi.org/10.1136/bmj.b375 (2009).
9. Alaa, A. M. et al. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One* **14**(5), e0213653. https://doi.org/10.1371/journal.pone.0213653 (2019).
10. Shan, Q. et al. A new nomogram and risk classification system for predicting survival in small cell lung cancer patients diagnosed with brain metastasis: a large population-based study. *BMC Cancer* **21**(1), 640. https://doi.org/10.1186/s12885-021-08384-5 (2021).
11. Shek, A. et al. Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing. *Eur. J. Neurol.* **28**(12), 4090–4097. https://doi.org/10.1111/ene.15071 (2021).
12. Hulsen, T. et al. From Big Data to Precision Medicine. *Front. Med. (Lausanne)* **6**, 34. https://doi.org/10.3389/fmed.2019.00034 (2019).
13. Kang, D. & Oh, S. Balanced training/test set sampling for proper evaluation of classification models. *Intell. Data Anal.* **24**, 5–18. https://doi.org/10.3233/IDA-194477 (2020).
14. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
15. Xiong, G. et al. Myocardial perfusion analysis in cardiac computed tomography angiographic images at rest. *Med. Image Anal.* **24**(1), 77–89. https://doi.org/10.1016/j.media.2015.05.010 (2015).
16. Chen, T. & Guestrin, C. *XGBoost*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 785–794. (2016).
17. Stoltzfus, J. C. Logistic regression: a brief primer. *Acad. Emerg. Med.* **18**(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x (2011).
18. Lapin, M., Hein, M. & Schiele, B. Learning using privileged information: SVM + and weighted SVM. *Neural Netw.* **53**, 95–108. https://doi.org/10.1016/j.neunet.2014.02.002 (2014).
19. Yadaw, A. S. et al. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit. Health* **2**(10), e516–e. https://doi.org/10.1016/s2589-7500(20)30217-x (2020).
20. Muschelli, J. ROC and AUC with a binary predictor: a potentially misleading Metric. *J. Classif.* **37**(3), 696–708. https://doi.org/10.1007/s00357-019-09345-1 (2020).
21. Gramegna, A. & Giudici, P. SHAP and LIME: an evaluation of discriminative power in Credit Risk. *Front. Artif. Intell.* **4**, 752558. https://doi.org/10.3389/frai.2021.752558 (2021).
22. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0 (2018).
23. Pellegrino, E. et al. Machine learning random forest for predicting oncosomatic variant NGS analysis. *Sci. Rep.* **11** (1), 21820. https://doi.org/10.1038/s41598-021-01253-y (2021).
24. Li, J. et al. ForestQC: quality control on genetic variants from next-generation sequencing data using random forest. *PLoS Comput. Biol.* **15**(12), e1007556. https://doi.org/10.1371/journal.pcbi.1007556 (2019).
25. Hao, Y. & Li, G. Risk and prognostic factors of brain metastasis in lung cancer patients: a Surveillance, Epidemiology, and end results population–based cohort study. *Eur. J. Cancer Prev.* **32**(5), 498–511. https://doi.org/10.1097/cej.0000000000000790 (2023).
26. Rong, Y. T., Zhu, Y. C. & Wu, Y. A novel nomogram predicting cancer-specific survival in small cell lung cancer patients with brain metastasis. *Transl Cancer Res.* **11**(12), 4289–4302. https://doi.org/10.21037/tcr-22-1561 (2022).
27. Zhang, G. H., Liu, Y. J. & De Ji, M. Risk factors, prognosis, and a New Nomogram for Predicting Cancer-Specific Survival among Lung Cancer patients with Brain Metastasis: a retrospective study based on SEER. *Lung* **200** (1), 83–93. https://doi.org/10.1007/s00408-021-00503-0 (2022).
28. Crockett, C. et al. Prophylactic cranial irradiation (PCI), hippocampal avoidance (HA) whole brain radiotherapy (WBRT) and stereotactic radiosurgery (SRS) in small cell lung cancer (SCLC): where do we stand? *Lung Cancer* **162**, 96–105. https://doi.org/10.1016/j.lungcan.2021.10.016 (2021).
29. Takahashi, T. et al. Prophylactic cranial irradiation versus observation in patients with extensive-disease small-cell lung cancer: a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol.* **18**(5), 663–671. https://doi.org/10.1016/s1470-2045(17)30230-9 (2017).
30. Moher, D. et al. *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS Med., 6(7): p. e1000097. DOI: https://doi.org/10.1371/journal.pmed.1000097 (2009).
31. Zeng, H. et al. Risk factors for brain metastases in patients with small cell Lung Cancer: a systematic review and Meta-analysis. *Front. Oncol.* **12**, 889161. https://doi.org/10.3389/fonc.2022.889161 (2022).
32. Zhu, H. et al. Risk factors for brain metastases in completely resected small cell lung cancer: a retrospective study to identify patients most likely to benefit from prophylactic cranial irradiation. *Radiat. Oncol.* **9**, 216. https://doi.org/10.1186/1748-717x-9-216 (2014).
33. Sahmoun, A. E. et al. Anatomical distribution of small cell lung cancer: effects of lobe and gender on brain metastasis and survival. *Anticancer Res.* **25**(2a), 1101–8DOI (2005).
34. Zhou, G. et al. Predictive value of clinical characteristics on risk and prognosis of synchronous brain metastases in small-cell lung cancer patients: a population-based study. *Cancer Med.* **12**(2), 1195–1203. https://doi.org/10.1002/cam4.4978 (2023).
35. Zheng, Y. et al. Risk factors for brain metastasis in patients with small cell lung cancer without prophylactic cranial irradiation. *Strahlenther Onkol* **194**(12), 1152–1162. https://doi.org/10.1007/s00066-018-1362-7 (2018).
36. Chen, M. Y. et al. Factors affecting the risk of Brain Metastasis in Limited-Stage Small Cell Lung Cancer after prophylactic cranial irradiation. *Cancer Manag Res.* **14**, 1807–1814. https://doi.org/10.2147/cmar.S347449 (2022).
37. Farooqi, A. S. et al. Prophylactic cranial irradiation after definitive chemoradiotherapy for limited-stage small cell lung cancer: do all patients benefit? *Radiother Oncol.* **122**(2), 307–312. https://doi.org/10.1016/j.radonc.2016.11.012 (2017).
38. Qi, J. et al. Thoracic Radiotherapy benefits Elderly extensive-stage small cell Lung Cancer patients with distant metastasis. *Cancer Manag Res.* **11**, 10767–10775. https://doi.org/10.2147/cmar.S221225 (2019).

39. Puglisi, M. et al. Treatment options for small cell lung cancer - do we have more choice? *Br. J. Cancer* **102**(4), 629–638. https://doi.org/10.1038/sj.bjc.6605527 (2010).

Author contributions

Yongzhong Wu and Yunyun Zhang designed the study.Siwei Zeng, Ze Yuan, Dingyi Yang, Yong Jiang, and Qiang Wang conducted data analysis.Erha Munai conceived the project and wrote the manuscript.Dan Tao revised and approved the paper.All authors contributed to the article and approved the submitted version.

Funding

The current study was supported by grants from the Chongqing Science and Health Joint Medical Research Project (No. 2023GGXM002 to YZ Wu), National Natural Science Foundation Project (No. 82073347 to YZ Wu) and Chongqing Talent Plan (No. cstc2022ycjh-bgzxm0208 to YZ Wu).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80425-y.

Correspondence and requests for materials should be addressed to Y.W., Y.Z. or D.T.