



Predicting survival after radiosurgery in patients with lung cancer brain metastases using deep learning of radiomics and EGFR status

Chien-Yi Liao¹ · Cheng-Chia Lee^{2,3,4} · Huai-Che Yang^{2,3} · Ching-Jen Chen⁵ · Wen-Yuh Chung⁶ · Hsiu-Mei Wu^{3,7} · Wan-Yuo Guo^{3,7} · Ren-Shyan Liu^{8,9} · Chia-Feng Lu¹

Received: 11 October 2022 / Accepted: 13 February 2023 / Published online: 1 March 2023
© Australasian College of Physical Scientists and Engineers in Medicine 2023

Abstract

The early prediction of overall survival (OS) in patients with lung cancer brain metastases (BMs) after Gamma Knife radiosurgery (GKRS) can facilitate patient management and outcome improvement. However, the disease progression is influenced by multiple factors, such as patient characteristics and treatment strategies, and hence satisfactory performance of OS prediction remains challenging. Accordingly, we proposed a deep learning approach based on comprehensive predictors, including clinical, imaging, and genetic information, to accomplish reliable and personalized OS prediction in patients with BMs after receiving GKRS. Overall 1793 radiomic features extracted from pre-GKRS magnetic resonance images (MRI), clinical information, and epidermal growth factor receptor (EGFR) mutation status were retrospectively collected from 237 BM patients who underwent GKRS. DeepSurv, a multi-layer perceptron model, with 4 different aggregation methods of radiomics was applied to predict personalized survival curves and survival status at 3, 6, 12, and 24 months. The model combining clinical features, EGFR status, and radiomics from the largest BM showed the best prediction performance with concordance index of 0.75 and achieved areas under the curve of 0.82, 0.80, 0.84, and 0.92 for predicting survival status at 3, 6, 12, and 24 months, respectively. The DeepSurv model showed a significant improvement ($p < 0.001$) in concordance index compared to the validated lung cancer BM prognostic molecular markers. Furthermore, the model provided a novel estimate of the risk-of-death period for patients. The personalized survival curves generated by the DeepSurv model effectively predicted the risk-of-death period which could facilitate personalized management of patients with lung cancer BMs.

Keywords Epidermal growth factor receptor · Brain metastases · Radiosurgery · Deep learning · MRI radiomics · Survival prediction

✉ Chia-Feng Lu
alvin4016@nycu.edu.tw

¹ Department of Biomedical Imaging and Radiological Sciences, National Yang Ming Chiao Tung University, No. 155, Sec. 2, Linong Street, Beitou District, Taipei 112, Taiwan

² Department of Neurosurgery, Neurological Institute, Taipei Veteran General Hospital, Taipei, Taiwan

³ School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

⁴ Brain Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

⁵ Department of Neurological Surgery, University of Virginia Health System, Charlottesville, VG, USA

⁶ Department of Neurosurgery, Kaohsiung Veteran General Hospital, Kaohsiung, Taiwan

⁷ Department of Radiology, Taipei Veteran General Hospital, Taipei, Taiwan

⁸ Department of Nuclear Medicine, Cheng Hsin General Hospital, Taipei, Taiwan

⁹ Molecular and Genetic Imaging Core, Taiwan Animal Consortium, Taipei, Taiwan

Introduction

Over 40% of patients with non-small cell lung cancer (NSCLC) develop brain metastases (BMs) during the course of the disease [1]. Patients with BMs often present with severe neurological symptoms and poor prognosis. Median overall survival (OS) of patients with NSCLC-BMs without additional therapy is approximately one month [2, 3]. To alleviate neurological symptoms and improve OS, several treatment options are available for BM patients, including neurosurgery, chemotherapy, whole brain radiotherapy (WBRT), Gamma Knife radiosurgery (GKRS), tyrosine kinase inhibitors (TKIs), and combinations of these. In particular, GKRS has become the most preferred treatment modality for BMs because of its minimal impact on cognitive impairment and over 70% tumor control rate for a limited number of BMs [4, 5]. The median OS for NSCLC-BM patients treated with first-line GKRS is expected to be approximately one year [6]. In addition, TKI drugs have shown promising results in treating NSCLC-BM patients with epidermal growth factor receptor (EGFR) mutations [7]. The EGFR is a protein involved in cell growth and division. The gene for the EGFR is located on the short arm of chromosome 7 and encodes a 170-kDa type I transmembrane growth factor receptor with tyrosine kinase activity. EGFR gene mutations cause abnormal proliferation, angiogenesis, metastasis, and decreased apoptosis, leading to cancer. The exon 19 deletions and exon 21 L858R mutations in EGFR gene are the most frequently detected oncogenic drivers in approximately 30–50% of NSCLC-BM patients. EGFR-TKIs such as afatinib, erlotinib, and gefitinib inhibit the overexpression of EGFR genes and are widely used in the treatment of NSCLC-BM patients with EGFR mutations. Knowledge of the prevalence of EGFR mutations in different patient subgroups can provide a reference for diagnosis and treatment strategies. The use of EGFR information and TKIs has been emphasized in the management of patients with BM [8, 9].

The combination of GKRS and other treatment modalities significantly improves the tumor control rate and OS in patients with BM [3]. Previous studies have proposed several prognostic factors for patients with NSCLC-BMs, such as the volume and number of BMs, EGFR mutation status, and applied treatments [10, 11]. The management of patients is facilitated based on the association of these characteristics with prognosis. Nevertheless, tumor heterogeneity reduces the reliability of prognostic factors, and therefore prediction of OS in patients remains challenging. The characteristics of magnetic resonance imaging (MRI) can be quantified using radiomics analysis to estimate tumor heterogeneity [12–14]. Several studies have further

suggested that combining radiomic features extracted from pre-GKRS MRI with clinical information and machine-learning algorithms could improve the prediction of local tumor control in BM patients after GKRS [15, 16].

However, several challenges remain in predicting the OS after GKRS in patients with BM. First, OS can be influenced by various factors, including functional impairment, extracranial metastases, concurrent/subsequent treatments, and control of the primary tumor [11, 17]. These complicated factors reduce the prediction performance of conventional survival estimations [18]. Second, conventional survival estimations, such as the Cox proportional-hazards model and Kaplan–Meier estimator, may only consider linear relationships between predictors or perform a group analysis rather than a personalized prediction [19]. Finally, patients with BMs frequently present with more than one lesion, and previous studies have been inconclusive regarding feature-extraction strategies for multiple lesions [15]. Accordingly, an advanced algorithm for personalized survival prediction based on comprehensive predictors and appropriate feature aggregation methods for multiple BM lesions is required to benefit management of BM patients with GKRS treatment.

The aim of this study was to develop a reliable and personalized approach to predict OS of patients with NSCLC-BM. This study contributed to the OS prediction of NSCLC-BM patients from three aspects. First, we applied a multi-layer perception neural network with imaging and clinical features as inputs to predict personalized survival curves. The deep learning architecture with nonlinear activation functions could better model the interaction of input covariates and provide a reliable estimate of personalized survival curve. Second, we compared four aggregation methods of radiomic features and identified the most suitable one for OS prediction in patients with multiple BM lesions. The image traits of multiple lesions might vary due to differences in origin and pathology type. Accordingly, adopting an appropriate aggregation strategy of radiomic features (either averaging across lesions or taking the representative lesion) could significantly influence the outcome prediction model. The third contribution of this study was the comprehensive investigation of prognostic factors in NSCLC-BM. Comprehensive clinical data, including the EGFR gene mutation, treatment strategies such as the dosage of GKRS, the use of TKI and chemotherapeutic drugs, and clinical staging of patients, were included and integrated with radiomic features to achieve a superior OS prediction. Finally, we suggested that the proposed deep learning model based on the comprehensive radiomic, clinical, and genetic features could effectively predict OS in NSCLC-BM patients.

Materials and methods

Patient cohort and clinical characteristics

This study retrospectively included 237 NSCLC-BM patients treated with GKRS at Taipei Veteran General Hospital between 2012 and 2017. The patient dataset was collected in accordance with the following inclusion criteria: (1) identification of NSCLC by lung biopsy or open surgery, (2) presence of at least one visible BM on MRI, (3) treatment of patients with GKRS, and (4) patients with complete MRI and clinical follow-up information. This study was approved by the Institutional Review Board of Taipei Veterans General Hospital, which waived the requirement for informed consent.

In this study, most patients received other therapies for NSCLC-BM in addition to GKRS, including chemotherapy, WBRT, and TKI therapy. Chemotherapy was performed for systemic metastases. These treatments have been shown to be beneficial for NSCLC-BM patients [20–22]. Furthermore, clinical characteristics, such as the patient's age, control of the primary lung cancer, presence of other metastases, number of BMs, Karnofsky performance status (KPS), EGFR mutation, and dose of GKRS, were recorded for OS prediction.

MRI data and image preprocessing

MRI data acquisition and subsequent treatment planning were performed prior to the implementation of GKRS. Pre-GKRS MRIs, including pre-contrast T1-weighted (T1w; TR/TE = 500/9 ms), contrast-enhanced T1-weighted (T1c; TR/TE = 500/9 ms), and T2-weighted (T2w; TR/TE = 4000/109 ms) images were acquired for each patient. Several preprocessing steps were applied to the acquired pre-GKRS MRIs before the subsequent radiomics analysis. First, the resolution of the MRI sequences was adjusted to the same dimension with a pixel size of $1 \times 1 \times 1 \text{ mm}^3$. Second, a rigid-body image registration of T1w and T2w images to T1c images was performed using the mutual information algorithm. Finally, the image intensities were transformed into standardized ranges (Z-score transformations) based on the mean and standard deviation of the entire image for each MRI set.

Radiomic feature extraction and multi-lesion aggregation

To delineate regions of interest (ROIs), a multidisciplinary team of experienced neurosurgeons and neuroradiologists reached a consensus on the BM contour for treatment planning. Wavelet decomposition was applied to obtain

additional information regarding the frequency and location of the images [23]. For each MRI sequence, low-pass (L) and high-pass (H) dimensional filtering (Morlet wavelet) were applied to the three image axes, resulting in eight wavelet sets: LLL, LLH, LHL, LHH, HLL, HLH, HHL, and HHH filtered images. Radiomic features, including histogram, geometry, and texture analysis (gray level co-occurrence matrix, GLCM; gray level run length matrix, GLRLM; local binary pattern, LBP) [24, 25], were then extracted from all image sets (eight wavelet-decomposed and original images of each MRI sequence). A total of 1763 radiomic features were generated for each lesion ROI. All wavelet filtering, image preprocessing procedures, and subsequent radiomics analysis were performed using the previously published Multimodal Radiomics Platform (available online: http://cfu.lab.nycu.edu.tw/MRP_MLIglioma.html, accessed on 18 August 2022) [15, 26] in compliance with the Image Biomarker Standardization Initiative (IBSI) [27, 28]. The formulae for the radiomics analysis are listed in supplementary Table S1.

For patients with more than one BM, the following feature aggregation methods were applied to integrate the radiomic features of multiple lesions for OS prediction [29–31]. *Method I - average*: the geometric features were summed, and other features were averaged across all BMs; *Method II - weighted average*: the geometric features were summed, and other features were averaged based on the volume of each BM; *Method III - weighted average of three largest BMs*: only the three largest BMs were considered in the weighted average calculation; *Method IV - the largest BM*: only the largest BM was considered. These aggregation methods were compared with regard to their prediction efficacy of patient OS.

Statistical analysis and prediction models

The hold-out method was performed to randomly split dataset into training dataset (70% of the patients) and test dataset (the remaining 30% of patients). To identify key radiomic and clinical features and to reduce redundancy for OS prediction, a two-step feature selection approach was applied to the training dataset. The initial statistical tests, including univariate Cox proportional regression for continuous variables and chi-squared test for categorical variables, were followed by a sequential forward selection (SFS) algorithm [32]. To maintain the complexity of the deep learning model (i.e., sufficient number of input features), we applied a selection criterion of $p < 0.1$ in the first step (the Cox and Chi-squared methods). The performance of the constructed OS prediction models was evaluated using the test dataset.

In this study, the DeepSurv survival model based on the Cox proportional hazards deep neural network was applied to evaluate the risk of death after GKRS [33]. In contrast to

the conventional Cox proportional hazard model, the DeepSurv approach uses a multi-layer perceptron to perform a non-linear simulation of the hazard function. Therefore, this model has potential to show superior performance on survival prediction by considering nonlinear effects of covariates. The loss function of the DeepSurv model was defined as the average negative logarithmic partial likelihood ($l(\theta)$, Eq. 1):

$$l(\theta) = -\frac{1}{N_{E=1}} \sum_{i: E_i=1} \left(\hat{h}_\theta(x_i) - \log \sum_{j \in R(T_i)} e^{\hat{h}_\theta(x_j)} \right) \quad (1)$$

where x is the input features, $N_{E=1}$ indicates the number of patients with an observable event, $\hat{h}_\theta(x)$ is estimated non-linear Cox hazard function, $R(T_i)$ represents patients with survival longer than T_i , and θ represents the model weights. Gradient descent optimization was applied to minimize $l(\theta)$ to determine the prediction performance of the hazard function for mortality events.

Hyper-parameters of deep learning models control the learning process and may significantly influence the performance. Some hyper-parameters are associated with the structure and complexity of the model, such as the number of hidden layers and nodes. Other hyper-parameters regulate the training process and the convergence speed of the model, such as the optimizer, learning rate, and dropout rate. The number of fully connected layers and nodes in each layer control the model capacity. The model with many nodes or layers has a great model capacity and thus can model complex relations/interactions between input features. However, too large a capacity may cause overfitting, resulting in poor model generalization or failure to optimize mapping function. The learning rate is related to the weight update during model training. Too large a learning rate may result in the model not converging to an optimal solution, while too small a value may result in significant time costs or model convergence to a local optimal solution. To stabilize the update of weights, the Adam optimizer was applied in our model. The Adam optimizer applies the momentum to adjust the value of learning rate during the training process so that the learning rate will be controlled in a defined range. Dropout and L2 regularization methods were also applied to avoid overfitting. The dropout is a regularization method that randomly drops hidden layer nodes with a certain probability in each iteration, and the dropped nodes are not updated with their weights. A low dropout rate may result in overfitting, while a high rate may result in small node size for model training. L2 regularization is the addition of a penalty term based on Lagrange multipliers to the loss function to reduce overfitting.

Searching for appropriate hyper-parameters contributes to the prediction performance and is therefore an essential step in training the DeepSurv model. In this study, the grid

search strategy was applied to determine hyper-parameters [34]. Each combination of hyper-parameters was assessed by a k-fold ($k=3$) cross validation in each training set, and the final setup of hyper-parameters was determined based on the average index of concordance (C-index) and time cost. The model performance associated with hyper-parameters is listed in supplementary Table S2. The final DeepSurv model applied consisted of an input layer (number of nodes equal to the number of selected features), three hidden layers (each containing eight nodes with rectified linear unit activation and batch normalization), and an output layer. An Adam optimizer with an initial learning rate of 0.01, a learning rate decay of 0.01, a dropout rate of 20%, and L2 regularization was applied for the training process. In this study, four DeepSurv models were generated based on clinical features and four different aggregation methods of radiomic features. Individual nonlinear logarithmic risk functions and corresponding personalized survival curves were generated using the DeepSurv models. The workflow of the radiomics analysis and deep learning is shown in Fig. 1. The feature selection and subsequent DeepSurv model training were performed on R DeepSurv package (available online: <https://rdrr.io/cran/survivalmodels/src/R/deepsurv.R>, accessed on 18 August 2022).

A log-rank test was applied to evaluate the statistical difference in the average of personalized survival curves between the good survival (OS > median survival) and poor survival (OS < median survival) groups based on median OS (12.2 months) of included patients. The statistical power of log-rank test was calculated based on the α of 0.05, estimated hazard ratio, and sample size. Time-dependent receiver operating characteristic (ROC) curves, area under the ROC curve (AUC), C-index, sensitivity, and specificity were estimated to assess the prediction performance of survival status at different time points (i.e., 3, 6, 12, and 24 months). A bootstrap random sampling method [35] and the paired t-test were applied to statistically compare the prediction efficacy of the four radiomic aggregation methods. The Shapiro-Wilk test was applied to each bootstrap dataset to check whether the dataset was normally distributed. The log-rank tests and paired t-tests were 2-sided, and statistical significance was set at 0.05 or less.

To construct a reference risk curve to intuitively represent individual death risk, we first determined the optimal thresholds based on time-dependent ROC curves for each of the four selected time points. Subsequently, the Weibull probability distribution function was applied for curve fitting using four time-dependent thresholds as it was indicated to accurately model the time-to-failure of real-world events [36]. The area between the reference risk curve and personalized survival curve was used to assess the patients' risk of death. A negative value of the area during the observed period indicated that the portion of the personalized survival

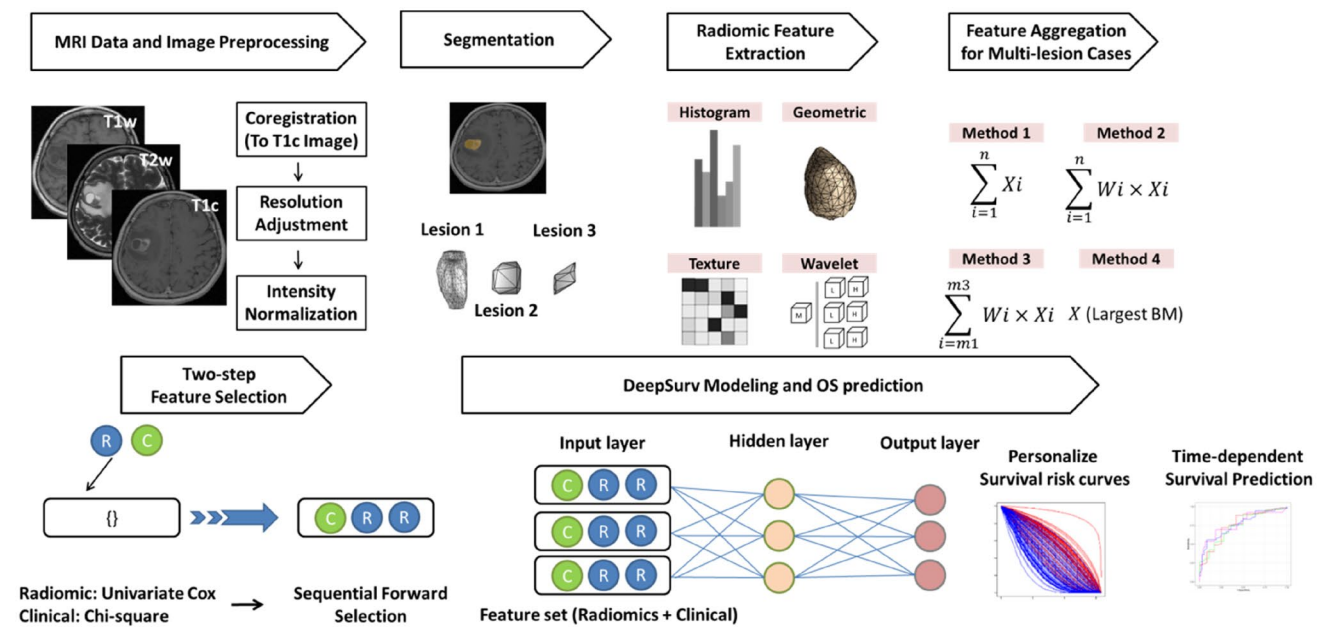


Fig. 1 Workflow of data analysis. Axial T1w and T2w images are co-registered to the T1c images followed by the resolution adjustment and intensity normalization. The radiomic features, including histogram, geometric, and texture features with the wavelet image decomposition, are extracted from the tumor ROIs. A two-step feature selec-

tion is applied to identify key features and to reduce the redundancy for modeling. Finally, the identified features (clinical and radiomic features) are input to the DeepSurv model to generate personalized survival curves for the OS prediction

curve was lower than the reference risk curve, resulting in a high risk of death. A schematic representation of the risk-of-death period is shown in supplementary Figure S1. All R language codes for training and validation of DeepSurv models were provided on Code Ocean (<https://codeocean.com/capsule/6451155/tree>).

The Graded Prognostic Assessment for Lung Cancer Using Molecular Markers (lung-molGPA) is a validated prognostic tool for patients with BMs from lung cancer [11]. The scoring criteria of lung-molGPA are listed in supplementary Table S3. The lung-molGPA was calculated for each patient and compared the prediction performance with DeepSurv models. However, we would like to emphasize that the lung-molGPA could not provide time-dependent survival prediction or the personalized survival curve.

Results

Clinical characteristics of recruited patients

All patients had a complete follow-up until death after GKRS, without any missing data or censoring during this period. Table 1 summarizes the clinical characteristics of the 237 patients. The age of the patients varied from 22.6 to 91.3 years (median = 60.8). The proportions of men (N = 115, 48.5%) and women (N = 122, 51.5%) were comparable.

Approximately 49.4% of patients had other metastases, 45.6% of patients had good control of primary NSCLC, 66.7% of patients presented with EGFR mutation, and 73.0% of patients had more than one BM. Most patients (98.4%) were histologically diagnosed with pure adenocarcinomas.

Selected features for OS prediction

The details of the selected features are listed in supplementary Table S4. In the clinical features, KPS, EGFR status and the use of TKIs were finally selected by the SFS algorithm. In the radiomic features, 3, 1, 26, and 40 features were selected by the SFS algorithm for *Methods I* to *IV*, respectively. These radiomic features included histogram describing variance of intensities and textural features describing low gray level regions in the T1w, T1c, and T2w images.

Performance of DeepSurv prediction models and lung-molGPA scores

Figure 2 shows the distribution of personalized survival curves generated by the DeepSurv models based on the four aggregation methods. Our results showed that all four DeepSurv models presented significant differences between the two survival groups ($p < 0.015$), indicating their prediction efficacy in differentiating patient outcomes. DeepSurv models based on *Methods I*, *III*, and *IV*

Table 1 Characteristics of 237 recruited NSCLC patients with BMs

Characteristics	Value
Age, median(IQR) ^a	62.7(15.9)
Gender	
Female, N(%)	122(51.5)
Male, N(%)	115(48.5)
Overall survival (months), median(IQR)	12.2(13.9)
Existence of other metastases	
Yes, N(%)	117(49.4)
No, N(%)	120(50.6)
KPS	
≥ 90, N(%)	164(69.2)
< 90, N(%)	73(30.8)
Primary NSCLC control	
Good, N(%)	108(45.6)
Poor, N(%)	129(54.4)
Number of BMs	
1, N(%)	64(27.0)
2, N(%)	50(21.1)
3, N(%)	29(12.2)
4, N(%)	25(10.6)
> 4, N(%)	69(29.1)
NSCLC histology	
Pure adenocarcinoma, N(%)	233(98.4)
Adenocarcinoma and Large cell carcinoma, N(%)	1(0.4)
Adenocarcinoma and Squamous cell carcinoma, N(%)	1(0.4)
Undifferentiated NSCLC, N(%)	2(0.8)
Additional treatment options	
Whole-brain radiotherapy, N(%)	30(12.7)
TKIs before GKRS, N(%)	155(65.4)
TKIs after GKRS, N(%)	163(68.8)
Chemotherapy, N(%)	137(57.8)
EGFR status	
Wild type, N(%)	74(31.2)
Mutation, N(%)	158(66.7)
Not available, N(%)	5(2.1)
GKRS dose (Gray)	
Tumor center, median(IQR)	28.6(6)
Tumor periphery, median(IQR)	19(2)
Lung-molGPA	
0.5, N(%)	5(2.1)
1, N(%)	8(3.4)
1.5, N(%)	18(7.6)
2, N(%)	27(11.4)
2.5, N(%)	52(21.9)
3, N(%)	67(28.3)
3.5, N(%)	28(11.8)
4, N(%)	32(13.5)

^a IQR interquartile range

achieved statistical powers larger than 0.84, and the model based on *Method II* achieved a statistical power of 0.75.

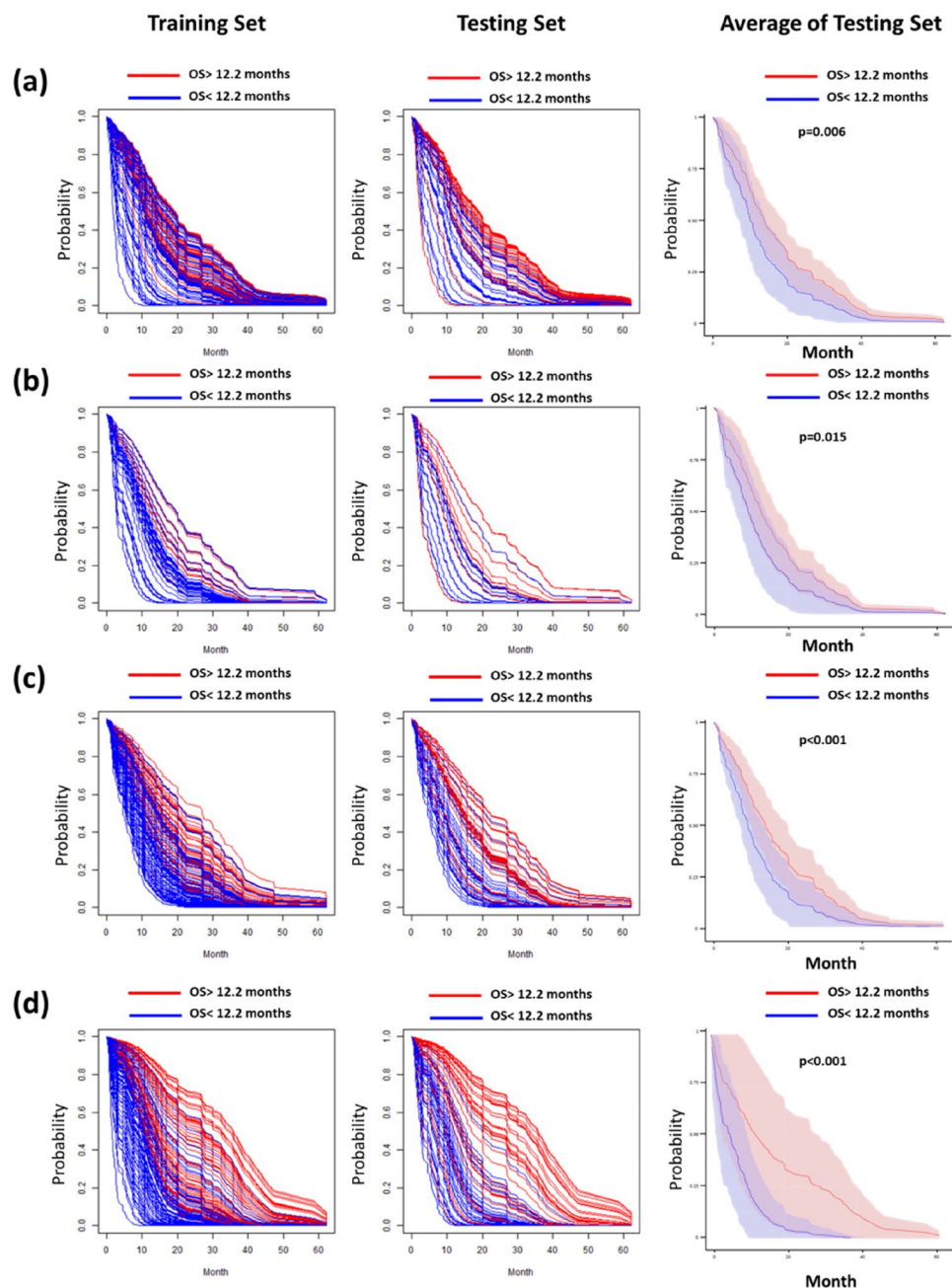
We estimated the prediction performance at four selected time points (3, 6, 12, and 24 months) using the test dataset. Table 2 lists the C-index, AUC, threshold, sensitivity, and specificity achieved using the four different aggregation methods. Among all the methods, the model based on *Method IV* achieved the best performance (C-index = 0.75) with the highest AUCs (0.82, 0.80, 0.84, and 0.92) sensitivities (75%, 73%, 75%, and 86%), and specificities (83%, 79%, 83%, and 90%) at 3, 6, 12, and 24 months, respectively. The time-dependent ROC curves for each model are shown in supplementary Figure S2. Bootstrap random sampling for 100 times was further applied for statistical comparisons between the different models. The results of the Shapiro-Wilk test showed that all the time-dependent AUCs estimated using the bootstrap random sampling were normally distributed ($p > 0.05$). The model based on aggregation *Method IV* showed significantly higher AUCs than the other models in predicting OS at 6, 12, and 24 months (Fig. 3).

Figure 4 illustrates the prediction of risk-of-death period for representative cases with short (0.9 months), moderate (9.9 months), and long OS (37.3 months), respectively. For the patient with a short OS (lung-molGPA = 2.5, three BMs and wild-type EGFR, Fig. 4a), the prediction models based on aggregation *Methods I, III, and IV* correctly predicted a risk-of-death period of less than three months. For the patient with a moderate OS (lung-molGPA = 4, two BMs and EGFR mutation, Fig. 4b), *Methods II and IV* correctly predicted OS with a risk-of-death period between 6 and 12 months. For the patient with a long OS (lung-molGPA = 3, one BM and EGFR mutation, Fig. 4c), only *Methods III and IV* correctly predicted a risk-of-death period beyond 24 months. These results suggested that the aggregation approach based on the largest BM (*Method IV*) provided the most stable and accurate estimate of the risk-of-death period in patients with BM after GKRS.

Finally, the lung-molGPA score achieved a C-index of 0.66 for the OS prediction in the test dataset. The general performance of OS prediction using DeepSurv model (C-index of 0.75) significantly outperformed ($p < 0.001$) that using the lung-molGPA score (C-index of 0.66). More importantly, DeepSurv model could provide time-dependent prediction and personalized survival curves.

The development and validation of DeepSurv models followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [37]. The items of the TRIPOD checklist are listed in supplementary Table S5.

Fig. 2 Distribution of personalized survival curves predicted by the DeepSurv models. The estimated personalized survival curves using training (left column) and testing set (middle column) based on aggregation Method I - average (a), Method II - weighted average (b), Method III - weighted average of three largest BMs (c), and Method IV – the largest BM (d). The red curves represent the patients with OS better than median OS (12.2 months), and blue curves indicate the patients with OS poorer than median OS. The personalized survival curves show significant difference ($p < 0.05$, log-rank test) between the good and poor OS groups in the testing set (right column)



Discussion

Several clinical and pretreatment imaging characteristics have been proposed for predicting OS in patients with BM from NSCLC [15, 38]. However, these studies have combined radiomics with traditional machine learning methods, such as support vector machines and random forests, to predict patient survival at a single time point. Accordingly, the implementation of survival prediction in personalized medicine remains challenging. This study proposed a deep learning approach based on radiomic features and EGFR status to estimate the personalized survival curve and improve the

OS prediction after GKRS. The DeepSurv model has been shown to be more appropriate than traditional statistical and machine-learning algorithms for handling nonlinear interactions between prognostic factors [33].

Currently, no standard rule defines the extraction of radiomic features from multiple lesions for survival prediction. In this study, the DeepSurv model based on *Method IV* (the largest BM) achieved the best performance for OS prediction (Fig. 2) and significantly outperformed the other models in predicting survival status at 6, 12, and 24 months (Fig. 3). These findings could be attributed to several factors. First, metastatic cancers of the brain are usually small and

Table 2 Performance of DeepSurv models based on the test dataset

Estimate	Prediction time points			
	3 months	6 months	12 months	24 months
Method I - average (C-index = 0.70)				
Threshold	87%	76%	50%	31%
AUC (95% CI ^a)	0.75 (0.60–0.90)	0.74 (0.64–0.84)	0.76 (0.68–0.86)	0.77 (0.67–0.86)
Sensitivity	72%	74%	64%	69%
Specificity	75%	67%	72%	81%
Method II - weighted average (C-index = 0.67)				
Threshold	83%	73%	45%	27%
AUC (95% CI)	0.83 (0.73–0.93)	0.73 (0.64–0.82)	0.75 (0.65–0.86)	0.77 (0.66–0.89)
Sensitivity	84%	73%	74%	78%
Specificity	73%	65%	70%	74%
Method III - weighted average of three largest BMs (C-index = 0.73)				
Threshold	83%	28%	50%	26%
AUC (95% CI)	0.82 (0.71–0.93)	0.78 (0.72–0.84)	0.80 (0.71–0.88)	0.87 (0.78–0.95)
Sensitivity	75%	74%	78%	85%
Specificity	91%	79%	78%	81%
Method IV – the largest BM (C-index = 0.75)				
Threshold	84%	74%	52%	46%
AUC (95% CI)	0.82 (0.71–0.90)	0.80 (0.72–0.88)	0.84 (0.74–0.90)	0.92 (0.86–0.98)
Sensitivity	75%	73%	75%	86%
Specificity	83%	79%	83%	90%

^a CI confidence interval estimated using bootstrap random sampling for 100 times

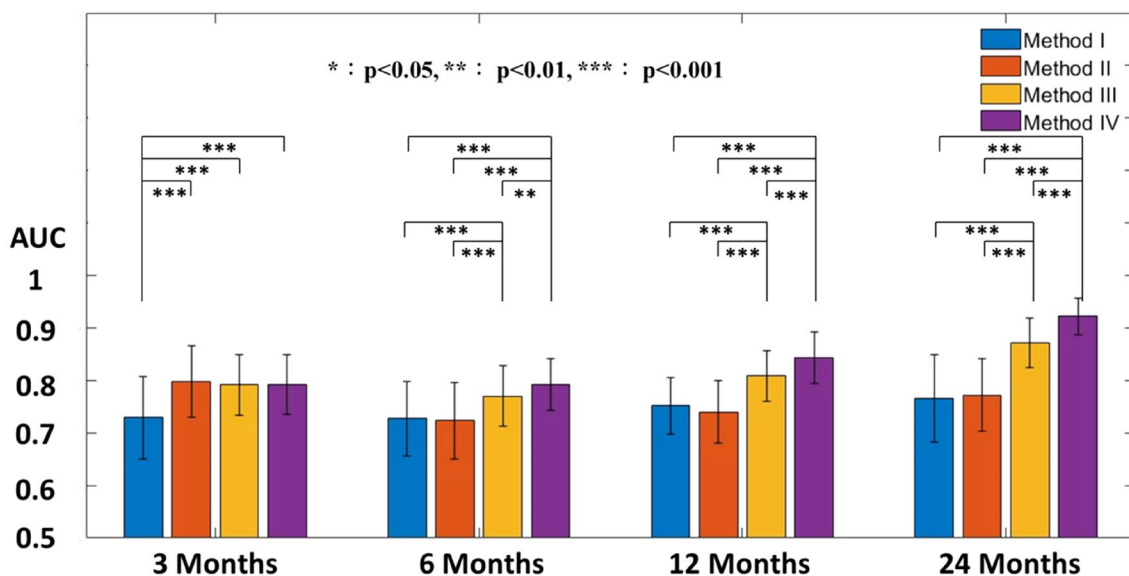


Fig. 3 Statistical comparisons of four radiomic aggregation methods. Statistical comparisons (paired t-test) of four aggregation methods are performed using bootstrap random resampling for 100 times in the testing set. Error bars: Standard deviations

present as multiple lesions. Previous studies have suggested that small BMs have a limited diagnostic value in clinical practice [39]. Second, previous studies have shown that the presence of large BMs predisposes patients to poor OS and local tumor control after GKRS [40]. Therefore, the largest

tumor may contain critical information for predicting OS. Finally, we found that the number of BMs was more important for OS prediction than the radiomic features extracted from limited voxels of small lesions. Accordingly, we suggested that the radiomic features extracted from the largest

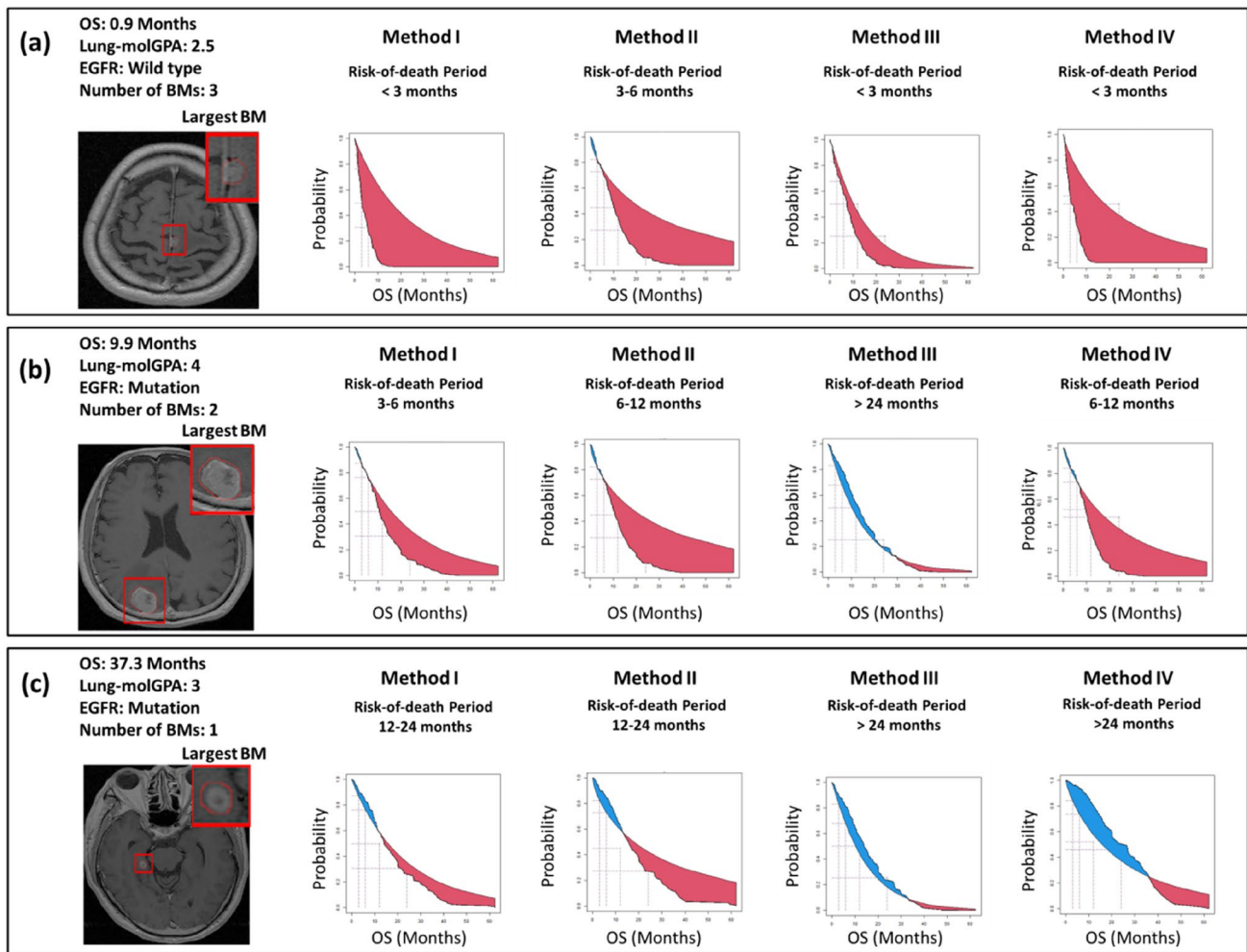


Fig. 4 Representative cases for OS prediction based on different aggregation methods. Figure shows MRIs and the predicted risk-of-death period based on DeepSurv models in **a** a patient with poor OS

(0.9 months), wild-type EGFR, and three BMs; **b** a patient with moderate OS (9.9 months), mutant EGFR, and two BMs; **c** a patient with good OS (37.3 months), mutant EGFR, and one BM.

lesion were sufficient to predict OS after GKRS in patients with multiple BMs.

The lung-molGPA is a clinically available tool for the prognosis assessment in NSCLC-BM. However, the prediction based on lung-molGPA only achieved a C-index of 0.66 in our patient cohort. Based on the results of DeepSurv model, we suggested that MRI radiomics could provide valuable information to enhance the OS prediction. A previous study combined radiomic and clinical features with support vector machines to predict 12 months OS after GKRS in 237 BM patients, achieving an AUC of 0.81 [15]. The DeepSurv model proposed in this study provided the time-dependent prediction at multiple time points, including 3, 6, 12, and 24 months with superior performance (AUCs of 0.82, 0.80, 0.84, and 0.92, respectively; C-index of 0.75). The prediction performance of the lung-molGPA, support vector machine, and DeepSurv methods for the OS of these 237

BM patients is compared in supplementary Table S6. This improvement in the prediction performance may be attributed to two possible reasons. First, the DeepSurv model can learn the effects of covariates and continuously update feature weights with multiple hidden layers and nonlinear activation functions. This model is appropriate for dealing with high-dimensional data because the weights of the key features gradually increase during the learning process. Second, in addition to KPS, existence of extracranial metastases, and number of BMs, we further included EGFR gene status and target therapy as key clinical features for OS prediction. EGFR status is a known prognostic factor for NSCLC-BM, and TKI-targeted therapy with GKRS has been shown to be more effective than TKI alone [41]. Accordingly, comprehensive clinical data, including well-known clinical factors, gene status, and target therapy, may benefit OS prediction in NSCLC-BM patients after GKRS.

Tuning of the hyper-parameters is an essential step in the DeepSurv model. The results of the grid search indicated that a small initial training rate (0.001) led to a reduction in prediction performance. This suggested that differences in the hazard function (i.e., survival status) between patients could not be effectively distinguished if a low initial learning rate and limited epochs were applied. The adjustment of other hyper-parameters, including number of hidden layers, number of nodes in each hidden layer, learning rate decay, and dropout rate, showed minor effects on the prediction performance. For example, a 263% change rate of node number in each hidden layer caused only a change in the C-index within 5.7% (see Table S2 for details). However, an appropriate selection of hyper-parameters largely reduced the model training time. For example, we selected three hidden layers as the final setup because it achieved the highest C-index (0.73) and a relatively low time cost (364 s). Therefore, the combination of hyper-parameters with the lowest time cost was selected when the prediction performance was similar.

MRI radiomic features have been reported to be associated with tumor control and OS after GKRS in patients with NSCLC–BM. For the DeepSurv model based on *Method IV*, high values of variance, standard deviation, and mean absolute deviation on T1w and T1c images indicated long patient OS. The high deviation in T1w and T1c intensities may indicate the presence of hypointensity components within the ROI, such as calcifications, cysts, and edema [42], in addition to contrast-enhanced (hyperintensity) tumor tissues. Furthermore, patients with long OS presented high values of two texture features, including the short run low gray-level emphasis and long run low gray-level emphasis, in T1w, T1c, and T2w images. Both features emphasize the spatial and intensity heterogeneity of the low-grayscale (hypointensity) components. This finding again supported that the presence and distribution of calcifications, cysts, and edema within the ROI may be useful imaging predictors of OS.

We proposed the application of the DeepSurv model to estimate the risk-of-death period. The estimated personalized survival curves provided information on the survival probability at different times after treatment. The thresholds implemented in the time-dependent ROC curves were used to predict the patient survival status at each time point. The risk-of-death period was estimated by comparing the reference risk curve with personalized survival curves (supplementary Figure S1). This approach provided a longitudinal description of patient survival. For patients with poor OS, most aggregation methods provided reliable estimates of the risk-of-death period. However, only aggregation *Method IV* provided an accurate estimate of the risk-of-death period in patients with moderate and good OS.

Several limitations of this study and further considerations are discussed below. First, the cases were collected

from a single institution in this study. An external validation dataset should be considered in future studies to validate the proposed model. Second, a multidisciplinary team of experienced neurosurgeons and neuroradiologists performed semi-automatic tumor segmentation on the MRIs. The development of automated segmentation approaches could reduce the time and cost of treatment planning and improve the reproducibility of radiomic features. Finally, EGFR status and TKI therapy implementation were identified as key predictors of OS in patients with NSCLC–BM. Our results showed that patients with wild-type EGFR status had poorer OS. Further investigation of survival prediction could focus on patients with wild-type EGFR. This may facilitate the management of patients with BM who have a potentially poor prognosis.

Conclusion

NSCLC–BM patients with EGFR mutations, who were treated with TKI, exhibited good OS. This study showed that the combination of a deep neural network, MRI quantitative features, and EGFR genetic information provided promising results for OS prediction in patients with BM after GKRS. The personalized survival curve and reference risk curve generated by the DeepSurv model deliver an intuitive prognostic assessment. The DeepSurv model could benefit patient management and treatment strategies for BM treated with GKRS.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s13246-023-01234-7>.

Author contributions Conception and design: C-YL, C-CL, C-FL. Acquisition of data: C-CL, H-CY, W-YC, H-MW, W-YG. Analysis and interpretation of data: C-YL, C-C, C-FL. Statistical analysis: C-YL, C-FL. Drafting the article: C-YL, C-CL, C-FL, C-JC. Critically revising the article: all authors. Reviewed and approved submitted version of manuscript: all authors. Study supervision: C-CL, R-SL, C-FL.

Funding This work was supported by the Ministry of Science and Technology, Taiwan (MOST109-2314-B-010-022-MY3) and Veterans General Hospitals and University System of Taiwan Joint Research Program (VGHUST112-G1-3-3). The funding sources had no role in the design and conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval This study was performed in accordance with the Declaration of Helsinki. The Institutional Review Board of Taipei Veterans General Hospital approved this retrospective study (2022-07-049BC) and waived the requirement of acquiring informed consent from patients.

References

- Schouten LJ, Rutten J, Huvencuers HA, Twijnstra A (2002) Incidence of brain metastases in a cohort of patients with carcinoma of the breast, colon, kidney, and lung and melanoma. *Cancer* 94:2698–2705
- Kelly K, Bunn PA Jr (1998) Is it time to reevaluate our approach to the treatment of brain metastases in patients with non-small cell lung cancer? *Lung Cancer* 20:85–91
- Suh JH, Kotecha R, Chao ST, Ahluwalia MS, Sahgal A, Chang EL (2020) Current approaches to the management of brain metastases. *Nat reviews Clin Oncol* 17:279–299
- Chao ST, De Salles A, Hayashi M, Levivier M, Ma L, Martinez R et al (2018) Stereotactic radiosurgery in the management of limited (1–4) brain metastases: systematic review and international stereotactic radiosurgery society practice guideline. *Neurosurgery* 83:345–353
- Hong AM, Fogarty GB, Dolven-Jacobsen K, Burmeister BH, Lo SN, Haydu LE et al (2019) Adjuvant whole-brain radiation therapy compared with observation after local treatment of melanoma brain metastases: a multicenter, randomized phase III trial. *J Clin oncology: official J Am Soc Clin Oncol* 37:3132–3141
- Pan H-C, Sheehan J, Stroila M, Steiner M, Steiner L (2005) Gamma knife surgery for brain metastases from lung cancer. *J Neurosurg* 102:128–133
- Magnuson WJ, Lester-Coll NH, Wu AJ, Yang TJ, Lockney NA, Gerber NK et al (2017) Management of brain metastases in tyrosine kinase inhibitor-naïve epidermal growth factor receptor-mutant non-small-cell lung cancer: a retrospective multi-institutional analysis. *J Clin Oncol* 35:1070–1077
- Hsu C-H, Tseng C-H, Chiang C-J, Hsu K-H, Tseng J-S, Chen K-C et al (2016) Characteristics of young lung cancer: analysis of Taiwan's nationwide lung cancer registry focusing on epidermal growth factor receptor mutation and smoking status. *Oncotarget* 7:46628
- Zhang Y-L, Yuan J-Q, Wang K-F, Fu X-H, Han X-R, Threapleton D et al (2016) The prevalence of EGFR mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis. *Oncotarget* 7:78985
- Andratschke N, Kraft J, Nieder C, Tay R, Califano R, Soffiatti R et al (2019) Optimal management of brain metastases in oncogenic-driven non-small cell lung cancer (NSCLC). *Lung Cancer* 129:63–71
- Sperduto PW, Yang TJ, Beal K, Pan H, Brown PD, Bangdiwala A et al (2017) Estimating survival in patients with lung cancer and brain metastases: an update of the graded prognostic assessment for lung cancer using molecular markers (Lung-molGPA). *JAMA Oncol* 3:827–831
- Kim JY, Park JE, Jo Y, Shim WH, Nam SJ, Kim JH et al (2019) Incorporating diffusion-and perfusion-weighted MRI into a radiomics model improves diagnostic performance for pseudoprogression in glioblastoma patients. *Neurooncology* 21:404–414
- Wang G, He L, Yuan C, Huang Y, Liu Z, Liang C (2018) Pretreatment MR imaging radiomics signatures for response prediction to induction chemotherapy in patients with nasopharyngeal carcinoma. *Eur J Radiol* 98:100–106
- Wang J, Wu C-J, Bao M-L, Zhang J, Wang X-N, Zhang Y-D (2017) Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. *Eur Radiol* 27:4082–4090
- Liao C-Y, Lee C-C, Yang H-C, Chen C-J, Chung W-Y, Wu H-M et al (2021) Enhancement of radiosurgical treatment outcome prediction using MRI radiomics in patients with non-small cell lung cancer brain metastases. *Cancers* 13:4030
- Mouraviev A, Detsky J, Sahgal A, Ruschin M, Lee YK, Karam I et al (2020) Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery. *Neurooncology* 22:797–805
- Sperduto PW, Yang TJ, Beal K, Pan H, Brown PD, Bangdiwala A et al (2016) The effect of gene alterations and tyrosine kinase inhibition on survival and cause of death in patients with adenocarcinoma of the lung and brain metastases. *Int J Radiation Oncology* Biology* Phys* 96:406–413
- Stankiewicz M, Tomasik B, Blamek S (2021) A new prognostic score for predicting survival in patients treated with robotic stereotactic radiotherapy for brain metastases. *Sci Rep* 11:1–10
- Bollschweiler E (2003) Benefits and limitations of Kaplan–Meier calculations of survival chance in cancer surgery. *Langenbeck's Arch Surg* 388:239–244
- Mehta MP, Paleologos NA, Mikkelsen T, Robinson PD, Ammirati M, Andrews DW et al (2010) The role of chemotherapy in the management of newly diagnosed brain metastases: a systematic review and evidence-based clinical practice guideline. *J Neurooncol* 96:71–83
- Bindal RK, Sawaya R, Leavens ME, Lee JJ (1993) Surgical treatment of multiple brain metastases. *J Neurosurg* 79:210–216
- Wang C, Lu X, Lyu Z, Bi N, Wang L (2018) Comparison of up-front radiotherapy and TKI with TKI alone for NSCLC with brain metastases and EGFR mutation: a meta-analysis. *Lung Cancer* 122:94–99
- Kumar R, Saini B (2012) Improved image denoising technique using neighboring wavelet coefficients of optimal wavelet with adaptive thresholding. *Int J Comput Theory Eng* 4:395
- Dhruv B, Mittal N, Modi M (2019) Study of Haralick's and GLCM texture analysis on 3D medical images. *Int J Neurosci* 129:350–362
- García-Olalla Ó, Fernández-Robles L, Alegre E, Castejón-Limas M, Fidalgo E (2019) Boosting texture-based classification by describing statistical information of gray-levels differences. *Sensors* 19:1048
- Lu C-F, Hsu F-T, Hsieh KL-C, Kao Y-CJ, Cheng S-J, Hsu JB-K et al (2018) Machine learning-based radiomics for molecular subtyping of gliomas. *Clin Cancer Res* 24:4429–4436
- Depeursinge A, Andrearczyk V, Whybra P, van Griethuysen J, Müller H, Schaer R et al (2020) Standardised convolutional filtering for radiomics. *arXiv preprint arXiv:200605470*. <https://doi.org/10.48550/arXiv.2006.05470>
- Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A et al (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 295:328–338
- Derclé L, Lu L, Schwartz LH, Qian M, Tejpar S, Eggleton P et al (2020) Radiomics response signature for identification of metastatic colorectal cancer sensitive to therapies targeting EGFR pathway. *JNCI: J Natl Cancer Inst* 112:902–912
- Sperduto PW, Kased N, Roberge D, Xu Z, Shanley R, Luo X et al (2012) Summary report on the graded prognostic assessment: an accurate and facile diagnosis-specific tool to estimate survival for patients with brain metastases. *J Clin Oncol* 30:419
- Chang E, Joel MZ, Chang HY, Du J, Khanna O, Omuro A et al (2021) Comparison of radiomic feature aggregation methods for patients with multiple tumors. *Sci Rep* 11:1–7
- Mao KZ (2004) Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Trans Syst Man Cybernetics Part B (Cybernetics)* 34:629–634
- Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y (2018) DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 18:1–12

34. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. *J Mach Learn Res* 13:281–305
35. Dixon PM (2006) Bootstrap resampling. *Encycl Environ*. <https://doi.org/10.1002/9780470057339.vab028>
36. Weibull W (1951) A statistical distribution function of wide applicability. *J Appl Mech*. <https://doi.org/10.1115/1.4010337>
37. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW et al (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 162:W1–W73
38. Zhou C, Shan C, Lai M, Zhou Z, Zhen J, Deng G et al (2021) Individualized nomogram for predicting survival in patients with brain metastases after stereotactic radiosurgery utilizing driver gene mutations and volumetric surrogates. *Front Oncol* 11:1525
39. Lin NU, Lee EQ, Aoyama H, Barani IJ, Barboriak DP, Baumert BG et al (2015) Response assessment criteria for brain metastases: proposal from the RANO group. *Lancet Oncol* 16:e270–e8
40. Routman DM, Bian SX, Diao K, Liu JL, Yu C, Ye J et al (2018) The growing importance of lesion volume as a prognostic factor in patients with multiple brain metastases treated with stereotactic radiosurgery. *Cancer Med* 7:757–764
41. Chiou G-Y, Chiang C-L, Yang H-C, Shen C-I, Wu H-M, Chen Y-W et al (2021) Combined stereotactic radiosurgery and tyrosine kinase inhibitor therapy versus tyrosine kinase inhibitor therapy alone for the treatment of non-small cell lung cancer patients with brain metastases. *J Neurosurg* 1:1–8
42. Wu Z, Mittal S, Kish K, Yu Y, Hu J, Haacke EM (2009) Identification of calcification with MRI using susceptibility-weighted imaging: a case study. *J Magn Reson Imaging: Official J Int Soc Magn Reson Med* 29:177–182

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.