



OPEN Development and validation of a machine learning-based survival prediction model for Asian glioblastoma patients using the SEER database and Chinese data

Denglin Li^{1,4}, Luxin Zhang^{2,4}, Lifei Xu^{1,4}, Renhe Zhai¹, Hanyu Gao¹, Junlan Gao³, Minghai Wei¹, Ningwei Che¹✉ & Yeting He¹✉

Glioblastoma is an aggressive, malignant primary brain tumour and the most prevalent histological type of glioma. Our study attempted to investigate the independent predictors of overall survival (OS) and cancer-specific survival (CSS) in Asian patients with glioblastoma and establish predictive models for the OS and CSS of Asian patients with glioblastoma based on the machine learning algorithms. Data from Asian patients with glioblastoma in the SEER database were retrieved and stochastically grouped into a training set ($n = 845$) and a validation set ($n = 362$), and patients in our centre were assigned to the test set ($n = 172$). Univariate and multivariate Cox regression analyses were performed to evaluate the prognostic factors. Predictive models for OS and CSS were established based on eight machine learning algorithms, including Lasso Cox, random survival forest, CoxBoost, generalized boosted regression modelling (GBM), stepwise Cox and survival support vector machine, eXtreme Gradient Boosting, supervised principal component and partial least squares regression for Cox, and the selected predictive models were evaluated by the area under the ROC curves (AUC) and 95% confidence interval (CI), calibration curves and decision curve analyses in the training set, validation set and test set. In our retrospective study, age, tumour history, histologic type, surgery and chemotherapy were confirmed to be predictors of OS ($p < 0.05$); age, tumour history, histologic type, surgery and chemotherapy were identified as independent factors for CSS ($p < 0.05$). The predictive model for OS based on the GBM algorithm exhibited excellent predictive performance at 6 months (AUC = 0.837, 95% CI: 0.803–0.870), 12 months (AUC = 0.809, 95% CI: 0.780–0.839) and 24 months (AUC = 0.750, 95% CI: 0.717–0.783) in the training set, and the powerful predictive performance of the GBM model was confirmed in the validation and test sets, with good concordance between the predicted and observed OS rates demonstrated by calibration curves and clinical decision making performance suggested by the decision curve analyses curves. The predictive model based on the GBM algorithm for CSS also performed best in the training set at 6 months (AUC = 0.808, 95% CI: 0.770–0.847), 12 months (AUC = 0.755, 95% CI: 0.721–0.789) and 24 months (AUC = 0.692, 95% CI: 0.657–0.728) in the training set, and convincing predictive effectiveness was also confirmed in the validation and test sets with good calibration and clinical utility. Age, tumour history, histologic type, surgery and chemotherapy were confirmed to be independent factors for OS; and age, tumour history, histologic type, surgery and chemotherapy were identified as prognostic factors for CSS in our retrospective study. The predictive model constructed for OS and CSS based on the GBM algorithm in Asian patients with glioblastoma can be used to accurately predict OS and CSS in clinical practice, which may help tailor personalized treatment regimens and provide significant benefits for these patients.

Keywords Glioblastoma, Survival, Risk factor, Machine learning, Predictive model

¹Department of Neurosurgery, Second Affiliated Hospital of Dalian Medical University, No.467, Zhongshan Road, Dalian 116011, Liaoning Province, China. ²Department of Urology, Second Affiliated Hospital of Dalian Medical University, Dalian 116011, Liaoning Province, China. ³Department of Emergency, Second Affiliated Hospital of

Dalian Medical University, Dalian 116011, Liaoning Province, China. ⁴These authors contributed equally to this work: Denglin Li, Luxin Zhang and Lifei Xu. ✉email: screamingeagles@163.com; heyeting1985@live.com

Glioblastoma is the most prevalent histological type of glioma, accounting for more than half of the total number of gliomas, with a frequency of approximately four incidences among 100,000 people per year^{1–3}. Glioblastoma is more likely to occur in people aged 55–85 years, with a median age of 64 years. According to the WHO grading system, glioblastoma represents a highly invasive and fast-growing neoplasm originating in brain tissue and is categorized under the most severe classification level of central nervous system tumours⁴. Patients diagnosed with glioblastoma face challenging clinical outcomes, typically exhibiting limited life expectancy with a median survival of only 15 months and a median progression-free survival of 6.2–7.5 months despite standard therapies, including maximal safe surgical resection followed by concurrent radiotherapy and adjuvant chemotherapy^{5–7}. The identification of prognostic indicators and the development of reliable prediction tools for patients with glioblastoma have significant clinical value, as they enable timely therapeutic strategies and personalized care approaches, ultimately leading to improved treatment outcomes. Such advancements in risk stratification and predictive analytics may contribute to improved mortality rates through optimized clinical decision-making processes.

Machine learning (ML), which demonstrates superior capabilities in pattern recognition and predictive analytics and enables more robust data modelling and precise parameter estimation through advanced computational algorithms, has been widely used in various areas of clinical research^{8,9} and practice. Compared with traditional survival prediction models built employing multivariate Cox regression, ML enables the extraction of meaningful patterns from complex datasets, facilitates the identification of underlying correlations among variables, and supports the development of sophisticated forecasting frameworks through diverse algorithmic approaches¹⁰. Thus, our study aimed to investigate the independent predictors of overall survival (OS) and cancer-specific survival (CSS) and establish ML model to OS and CSS of Asian patients with glioblastoma. The model was trained and validated using data from the SEER database and tested through the data in our centre and can be used to provide guidance for clinical treatment decisions.

Materials and methods

Patient information acquisition

We searched the publicly accessible SEER database (<https://seer.cancer.gov/>), which consists of 17 registries covering more than 26.5% of the United States population between 2000 and 2021, to obtain the majority of information on patients with glioblastoma available in February 2025 through SEER*Stat (version 8.4.3). The inclusion criteria were as follows: (1) patients were diagnosed with glioblastoma between 2010 and 2021; (2) the race of the patient was Asian; (3) the histologic type ICD-O-3 site code was 9440/3, 9441/3, 9442/3 or 9445/3; (4) the pathology was confirmed; and (5) the surgery codes were “00”, “20”, “21”, “30”, or “55”. The exclusion criteria consisted of the following conditions: (1) patient age less than 18 years; (2) unknown tumour size; (3) unknown laterality; (4) unknown survival time or status; and (5) survival time of 0 months. The main outcome measures in our study were OS and CSS. OS was defined as the time span from the initial diagnosis of PUC to the time of death or the terminal follow-up. CSS was measured as the time from the diagnosis of glioblastoma to glioblastoma-specific death. Moreover, we collected the above patient information as a test set from the Second Affiliated Hospital of Dalian Medical University. This study obtained the approval of the Institutional Research Ethics Committees of the Second Affiliated Hospital of Dalian Medical University and the informed consent of all the enrolled patients and/or their legal guardians. All studies were performed in accordance with the Declaration of Helsinki and relevant guidelines.

Statistical analysis

Counting data were converted into categorical variables. Univariate and multivariate Cox analyses of the independent predictors of OS and CSS, presented as hazard ratios (HRs) and 95% confidence intervals (CIs), were performed, and a *p* value < 0.05 was considered to indicate statistical significance. Inverse probability of treatment weighting (IPTW) uses standardized methods to control and minimize potential biases and confounding effects on survival analyses by assigning appropriate weights to each observation through propensity score values. Cox regression analysis was subsequently performed to confirm the independent factors after IPTW. We randomly divided the SEER database data into training (*n* = 845) and validation (*n* = 362) sets at a proportion of 7:3, and the patients in our centre were assigned to the test set (*n* = 172). Univariate Cox analyses in the training set were performed to screen variables to construct the machine learning predictive model. Eight machine learning algorithms were implemented, including random survival forest (RSF), generalized boosted regression modelling (GBM), Lasso Cox, CoxBoost, stepwise Cox and survival support vector machine (Survival-SVM), eXtreme Gradient Boosting (XGBoost), supervised principal component (SuperPC) and partial least squares regression for Cox (plsRcox) to establish predictive models for OS and CSS. OS and CSS were evaluated at discrete time points (6, 12, and 24 months) to assess short-, medium-, and long-term prognostic performance. Discrimination was assessed using Harrell's concordance index (C-index) to present global measure and time-dependent ROC curves using the timeROC and riskRegression packages with AUCs calculated separately for each time point. The 95% confidence intervals for AUC values were estimated using bootstrap resampling (500 repetitions) to account for variability in model performance. To further validate the models, calibration curves were utilized to examine the alignment between predicted probabilities and actual outcomes. Decision curve analyses (DCAs) were used to evaluate the practical clinical value of these methods. Using the best-performing model, patients were stratified into high-risk and low-risk groups based on the median predicted risk score. This cutoff ensured balanced group sizes and maximized discriminative power for survival outcomes. Kaplan–Meier (KM) survival

curves and log-rank tests were performed to confirm the predictive model discrimination capacity between high-risk and low-risk groups. All the above analyses were performed in R version 4.4.2.

Results

Baseline characteristics

The baseline characteristics of glioblastoma patients in SEER set and test set were presented in the Table 1, key differences emerged in age, tumor history prevalence, and histologic type ($p < 0.05$) which may implying the data heterogeneity, while other demographic and clinical features remained comparable between cohorts.

Identification of the risk factors for patients with glioblastoma

Identification of the risk factors for OS

Univariate and multivariate Cox regression analyses were performed to screen for significant predictors of OS in the SEER set ($n = 1207$) and test set ($n = 172$). As shown in Table 2, age, histologic type, combined summary stage, surgery, radiotherapy and chemotherapy were confirmed as independent prognostic factors for OS in the

		SEER set ($n = 1207$)	Test set ($n = 172$)	<i>P</i> value
Year of diagnosis	≤2016	621 (51.4%)	86 (50.0%)	0.784
	>2016	586 (48.6%)	86 (50.0%)	
Age	≤60	569 (47.1%)	96 (55.8%)	0.041
	>60	638 (52.9%)	76 (44.2%)	
Gender	Female	542 (44.9%)	82 (47.7%)	0.548
	Male	665 (55.1%)	90 (52.3%)	
Tumor history	No	1075 (89.1%)	166 (96.5%)	0.004
	Yes	132 (10.9%)	6 (3.5%)	
Histologic type	Glioblastoma_NOS	1141 (94.5%)	128 (74.4%)	<0.001
	Giant cell glioblastoma	15 (1.2%)	18 (10.5%)	
	Gliosarcoma	32 (2.7%)	10 (5.8%)	
	Glioblastoma_IDH-mutant	19 (1.6%)	16 (9.3%)	
Primary site	Cerebrum	65 (5.4%)	10 (5.8%)	0.898
	Frontal lobe	449 (37.2%)	59 (34.3%)	
	Temporal lobe	301 (24.9%)	47 (27.3%)	
	Parietal lobe	197 (16.3%)	27 (15.7%)	
	Occipital lobe	26 (2.2%)	5 (2.9%)	
	Ventricle	3 (0.2%)	1 (0.6%)	
	Overlapping lesion of brain	119 (9.9%)	14 (8.1%)	
	Brain_NOS	47 (3.9%)	9 (5.2%)	
Laterality	Left	582 (48.2%)	81 (47.1%)	0.113
	Right	603 (50.0%)	83 (48.3%)	
	Midline	4 (0.3%)	2 (1.2%)	
	Bilateral	18 (1.5%)	6 (3.5%)	
Tumor size	≤20 mm	74 (6.1%)	12 (7.0%)	0.937
	≤40 mm	383 (31.7%)	55 (32.0%)	
	≤60 mm	502 (41.6%)	73 (42.4%)	
	≤80 mm	234 (19.4%)	31 (18.0%)	
	>80 mm	14 (1.2%)	1 (0.6%)	
Combined summary stage	Regional	218 (18.1%)	27 (15.7%)	0.749
	Localized	968 (80.2%)	142 (82.6%)	
	Distant	21 (1.7%)	3 (1.7%)	
Surgery	No	202 (16.7%)	26 (15.1%)	0.171
	Local	196 (16.2%)	38 (22.1%)	
	Subtotal	398 (33.0%)	52 (30.2%)	
	Radical	374 (31.0%)	47 (27.3%)	
	Gross_total	37 (3.1%)	9 (5.2%)	
Radiotherapy	No/Unknown	220 (18.2%)	34 (19.8%)	0.702
	Yes	987 (81.8%)	138 (80.2%)	
Chemotherapy	No/Unknown	275 (22.8%)	43 (25.0%)	0.583
	Yes	932 (77.2%)	129 (75.0%)	

Table 1. Baseline characteristics of glioblastoma patients in SEER set and test set.

Variable		SEER set (n = 1207)			Test set (n = 172)		
			HR (univariate)	HR (multivariate)		HR (univariate)	HR (multivariate)
Year of diagnosis	<=2016	621 (51.4%)			86 (50.0%)		
	>2016	586 (48.6%)	0.98 (0.86–1.12, p=0.739)		86 (50.0%)	1.08 (0.78–1.50, p=0.637)	
Age	<=60	569 (47.1%)			96 (55.8%)		
	>60	638 (52.9%)	1.96 (1.72–2.23, p<0.001)	1.77 (1.55–2.02, p<0.001)	76 (44.2%)	1.62 (1.17–2.25, p=0.004)	1.86 (1.30–2.67, p=0.001)
Gender	Female	542 (44.9%)			82 (47.7%)		
	Male	665 (55.1%)	0.98 (0.86–1.11, p=0.752)		90 (52.3%)	0.90 (0.65–1.25, p=0.541)	
Tumor history	No	1075 (89.1%)			166 (96.5%)		
	Yes	132 (10.9%)	1.28 (1.05–1.57, p=0.015)	1.21 (0.99–1.49, p=0.067)	6 (3.5%)	5.32 (2.30–12.30, p<0.001)	3.70 (1.55–8.78, p=0.003)
Histologic type	Glioblastoma_NOS	1141 (94.5%)			128 (74.4%)		
	Giant cell glioblastoma	15 (1.2%)	0.57 (0.31–1.03, p=0.063)	0.77 (0.42–1.40, p=0.392)	18 (10.5%)	0.59 (0.34–1.01, p=0.056)	0.55 (0.31–0.97, p=0.038)
	Gliosarcoma	32 (2.7%)	1.68 (1.15–2.47, p=0.008)	1.78 (1.21–2.62, p=0.004)	10 (5.8%)	0.69 (0.34–1.43, p=0.321)	0.58 (0.27–1.22, p=0.150)
	Glioblastoma_IDH-mutant	19 (1.6%)	0.33 (0.16–0.69, p=0.003)	0.51 (0.24–1.08, p=0.078)	16 (9.3%)	0.47 (0.26–0.83, p=0.010)	0.41 (0.22–0.75, p=0.004)
Primary site	Cerebrum	65 (5.4%)			10 (5.8%)		
	Frontal lobe	449 (37.2%)	1.09 (0.80–1.48, p=0.599)		59 (34.3%)	1.28 (0.61–2.69, p=0.519)	
	Temporal lobe	301 (24.9%)	1.02 (0.74–1.40, p=0.911)		47 (27.3%)	1.05 (0.49–2.26, p=0.892)	
	Parietal lobe	197 (16.3%)	1.23 (0.89–1.72, p=0.216)		27 (15.7%)	0.97 (0.43–2.20, p=0.950)	
	Occipital lobe	26 (2.2%)	1.19 (0.70–2.02, p=0.515)		5 (2.9%)	0.55 (0.17–1.83, p=0.331)	
	Ventricle	3 (0.2%)	0.95 (0.30–3.07, p=0.937)		1 (0.6%)	0.76 (0.09–6.09, p=0.796)	
	Overlapping lesion of brain	119 (9.9%)	1.26 (0.88–1.78, p=0.203)		14 (8.1%)	1.18 (0.48–2.90, p=0.718)	
	Brain_NOS	47 (3.9%)	1.47 (0.96–2.26, p=0.075)		9 (5.2%)	1.33 (0.49–3.57, p=0.573)	
Laterality	Left	582 (48.2%)			81 (47.1%)		
	Right	603 (50.0%)	0.93 (0.82–1.05, p=0.247)	1.01 (0.89–1.15, p=0.839)	83 (48.3%)	0.80 (0.57–1.12, p=0.187)	
	Midline	4 (0.3%)	0.72 (0.23–2.25, p=0.574)	0.76 (0.24–2.38, p=0.637)	2 (1.2%)	0.44 (0.10–1.81, p=0.253)	
	Bilateral	18 (1.5%)	2.44 (1.46–4.10, p<0.001)	1.57 (0.92–2.68, p=0.095)	6 (3.5%)	1.24 (0.50–3.08, p=0.647)	
Tumor size	<=20 mm	74 (6.1%)			12 (7.0%)		
	<=40 mm	383 (31.7%)	0.77 (0.58–1.01, p=0.056)		55 (32.0%)	1.03 (0.50–2.12, p=0.929)	
	<=60 mm	502 (41.6%)	0.82 (0.63–1.07, p=0.139)		73 (42.4%)	1.29 (0.64–2.59, p=0.479)	
	<=80 mm	234 (19.4%)	0.82 (0.62–1.10, p=0.183)		31 (18.0%)	1.29 (0.61–2.76, p=0.507)	
	>80 mm	14 (1.2%)	1.48 (0.80–2.75, p=0.213)		1 (0.6%)	1.56 (0.20–12.38, p=0.674)	
Combined summary stage	Regional	218 (18.1%)			27 (15.7%)		
	Localized	968 (80.2%)	0.63 (0.53–0.74, p<0.001)	0.63 (0.53–0.75, p<0.001)	142 (82.6%)	0.65 (0.41–1.02, p=0.062)	
	Distant	21 (1.7%)	3.06 (1.90–4.93, p<0.001)	2.45 (1.51–3.97, p<0.001)	3 (1.7%)	2.06 (0.48–8.84, p=0.329)	
Surgery	No	202 (16.7%)			26 (15.1%)		
	Local	196 (16.2%)	0.66 (0.53–0.81, p<0.001)	0.70 (0.56–0.87, p=0.001)	38 (22.1%)	0.79 (0.46–1.36, p=0.404)	0.76 (0.42–1.35, p=0.343)
	Subtotal	398 (33.0%)	0.55 (0.45–0.66, p<0.001)	0.65 (0.53–0.78, p<0.001)	52 (30.2%)	0.91 (0.55–1.52, p=0.732)	1.09 (0.64–1.89, p=0.745)
	Radical	374 (31.0%)	0.45 (0.38–0.55, p<0.001)	0.56 (0.46–0.68, p<0.001)	47 (27.3%)	0.63 (0.37–1.05, p=0.078)	0.68 (0.39–1.19, p=0.176)
	Gross_total	37 (3.1%)	0.43 (0.29–0.62, p<0.001)	0.52 (0.35–0.76, p=0.001)	9 (5.2%)	0.40 (0.16–0.99, p=0.049)	0.31 (0.12–0.80, p=0.015)

Continued

Variable		SEER set (n = 1207)			Test set (n = 172)		
			HR (univariate)	HR (multivariate)		HR (univariate)	HR (multivariate)
Radiotherapy	No/Unknown	220 (18.2%)			34 (19.8%)		
	Yes	987 (81.8%)	0.36 (0.31–0.43, $p < 0.001$)	0.63 (0.51–0.79, $p < 0.001$)	138 (80.2%)	0.63 (0.43–0.94, $p = 0.025$)	0.92 (0.55–1.52, $p = 0.734$)
Chemotherapy	No/Unknown	275 (22.8%)			43 (25.0%)		
	Yes	932 (77.2%)	0.36 (0.31–0.42, $p < 0.001$)	0.53 (0.43–0.64, $p < 0.001$)	129 (75.0%)	0.38 (0.26–0.55, $p < 0.001$)	0.40 (0.25–0.64, $p < 0.001$)

Table 2. Univariate and multivariate Cox regression analysis for the overall survival in the SEER and test set.

Outcomes	Before IPTW		After IPTW	
	HR (95%CI)	P value	HR (95%CI)	P value
SEER set				
No (ref)	1		1	
Yes	1.28 (1.05–1.57)	0.015	1.11 (0.87–1.42)	0.411
Test set				
No (ref)	1		1	
Yes	5.32 (2.30–12.30)	<0.001	2.06 (1.30–3.25)	0.002

Table 3. Cox regression analysis of tumor history on the overall survival before and after IPTW. Significant values are in bold. Adjusted for year of diagnosis, age, gender, histologic type, primary site, laterality, tumor size, combined summary stage, surgery, radiotherapy, chemotherapy.

SEER set ($p < 0.05$). In the test set, age, tumour history, histologic type, surgery and chemotherapy were screened as statistically significant prognostic factors for patients with glioblastoma ($p < 0.05$). To address potential selection bias of the tumor history on OS, we employed IPTW strategy to generate weighted cohorts using year of diagnosis, age, gender, histologic type, primary site, laterality, tumor size, combined summary stage, surgery, radiotherapy, chemotherapy. As shown in Table 3, tumor history continued to be an independent predictor of worse OS after employing IPTW (HR: 2.06, 95% CI: 1.30–3.25, $p = 0.002$), reinforcing the robustness of the association, while no significant difference was explored after IPTW in SEER set.

Identification of the risk factors for CSS

The detailed univariate and multivariate Cox regression analysis results for CSS in the SEER set and test set are presented in Table 4. Age, histologic type, primary site, combined summary stage, surgery, radiotherapy and chemotherapy were identified as statistically significant factors after univariate and multivariate Cox regression analyses in the SEER set ($p < 0.05$). In the test set, age, tumour history, histologic type, surgery and chemotherapy were confirmed to be independently significant factors for CSS ($p < 0.05$). We used the IPTW technique to create weighted cohorts based on year of diagnosis, age, gender, histologic type, primary site, laterality, tumor size, combined summary stage, surgery, radiotherapy, and chemotherapy to address the potential selection bias of tumor history on CSS. As demonstrated in Table 5, tumor history remained an independent predictor of worse OS after IPTW (HR: 1.93, 95% CI: 1.30–2.86, $p = 0.001$) in the test set, confirming the association strength.

Machine learning based predictive model for OS

We used univariate Cox regression to select features in the training set. As presented in Table 6, nine features, including age, tumour history, histologic type, laterality, tumour size, combined summary stage, surgery, radiotherapy and chemotherapy, were statistically significant and were selected for the construction of predictive models. Multiple machine learning algorithms, including RSF, GBM, Lasso Cox, CoxBoost, Survival-SVM, XGBoost, SuperPC and plsRcox, were employed to develop prognostic frameworks capable of estimating OS probabilities at 6-month, 12-month, and 24-month intervals. As the ROC curves in Fig. 1; Table 7 show, the GBM model exhibited excellent predictive performance at 6 months (AUC = 0.837, 95% CI: 0.803–0.870), 12 months (AUC = 0.809, 95% CI: 0.780–0.839) and 24 months (AUC = 0.750, 95% CI: 0.717–0.783) in the training set, and the powerful predictive performance of the GBM model was confirmed in the validation and test sets. The DCA curve also suggested that the GBM model holds considerable utility in making clinical decisions (Fig. 2). The GBM model showed good agreement between the predicted and observed OS rates at 6 months, 12 months and 24 months in the training, validation, and test sets, as shown by the calibration curves (Fig. 3). The survival curves in Fig. 4 demonstrate the differentiation capability between low-risk and high-risk OS patients in the training, validation and test sets ($p < 0.05$).

Variable		SEER set			Test set		
			HR (univariate)	HR (multivariate)		HR (univariate)	HR (multivariate)
Year of diagnosis	<=2016	621 (51.4%)			86 (50.0%)		
	>2016	586 (48.6%)	0.95 (0.83–1.09, <i>p</i> =0.474)		86 (50.0%)	1.10 (0.79–1.52, <i>p</i> =0.576)	
Age	<=60	569 (47.1%)			96 (55.8%)		
	>60	638 (52.9%)	1.89 (1.65–2.17, <i>p</i> <0.001)	1.71 (1.48–1.96, <i>p</i> <0.001)	76 (44.2%)	1.60 (1.15–2.22, <i>p</i> =0.005)	1.84 (1.28–2.64, <i>p</i> =0.001)
Gender	Female	542 (44.9%)			82 (47.7%)		
	Male	665 (55.1%)	1.01 (0.89–1.16, <i>p</i> =0.860)		90 (52.3%)	0.91 (0.66–1.26, <i>p</i> =0.570)	
Tumor history	No	1075 (89.1%)			166 (96.5%)		
	Yes	132 (10.9%)	1.27 (1.03–1.57, <i>p</i> =0.024)	1.22 (0.98–1.52, <i>p</i> =0.073)	6 (3.5%)	4.38 (1.76–10.90, <i>p</i> =0.001)	3.05 (1.19–7.79, <i>p</i> =0.020)
Histologic type	Glioblastoma_NOS	1141 (94.5%)			128 (74.4%)		
	Giant cell glioblastoma	15 (1.2%)	0.62 (0.34–1.13, <i>p</i> =0.119)	0.86 (0.47–1.57, <i>p</i> =0.618)	18 (10.5%)	0.59 (0.34–1.02, <i>p</i> =0.059)	0.55 (0.32–0.97, <i>p</i> =0.040)
	Gliosarcoma	32 (2.7%)	1.64 (1.09–2.46, <i>p</i> =0.017)	1.75 (1.15–2.64, <i>p</i> =0.008)	10 (5.8%)	0.70 (0.34–1.44, <i>p</i> =0.330)	0.58 (0.27–1.22, <i>p</i> =0.149)
	Glioblastoma_IDH-mutant	19 (1.6%)	0.25 (0.11–0.61, <i>p</i> =0.002)	0.38 (0.16–0.92, <i>p</i> =0.032)	16 (9.3%)	0.47 (0.26–0.84, <i>p</i> =0.011)	0.41 (0.22–0.75, <i>p</i> =0.004)
Primary site	Cerebrum	65 (5.4%)			10 (5.8%)		
	Frontal lobe	449 (37.2%)	1.19 (0.85–1.67, <i>p</i> =0.303)	1.29 (0.91–1.83, <i>p</i> =0.147)	59 (34.3%)	1.25 (0.60–2.64, <i>p</i> =0.551)	
	Temporal lobe	301 (24.9%)	1.07 (0.76–1.51, <i>p</i> =0.711)	1.14 (0.80–1.64, <i>p</i> =0.464)	47 (27.3%)	1.06 (0.50–2.27, <i>p</i> =0.878)	
	Parietal lobe	197 (16.3%)	1.32 (0.93–1.89, <i>p</i> =0.123)	1.31 (0.91–1.89, <i>p</i> =0.152)	27 (15.7%)	0.94 (0.42–2.13, <i>p</i> =0.882)	
	Occipital lobe	26 (2.2%)	1.19 (0.67–2.11, <i>p</i> =0.553)	1.34 (0.75–2.41, <i>p</i> =0.327)	5 (2.9%)	0.55 (0.17–1.83, <i>p</i> =0.330)	
	Ventricle	3 (0.2%)	0.74 (0.18–3.07, <i>p</i> =0.679)	1.23 (0.29–5.12, <i>p</i> =0.777)	1 (0.6%)	0.76 (0.09–6.08, <i>p</i> =0.795)	
	Overlapping lesion of brain	119 (9.9%)	1.38 (0.95–2.01, <i>p</i> =0.096)	1.48 (1.01–2.18, <i>p</i> =0.046)	14 (8.1%)	1.21 (0.49–2.97, <i>p</i> =0.681)	
	Brain_NOS	47 (3.9%)	1.62 (1.03–2.55, <i>p</i> =0.036)	1.16 (0.73–1.84, <i>p</i> =0.536)	9 (5.2%)	1.38 (0.52–3.71, <i>p</i> =0.519)	
Laterality	Left	582 (48.2%)			81 (47.1%)		
	Right	603 (50.0%)	0.94 (0.82–1.08, <i>p</i> =0.379)	1.01 (0.88–1.16, <i>p</i> =0.852)	83 (48.3%)	0.82 (0.59–1.14, <i>p</i> =0.239)	
	Midline	4 (0.3%)	0.80 (0.26–2.49, <i>p</i> =0.697)	0.79 (0.25–2.50, <i>p</i> =0.693)	2 (1.2%)	0.24 (0.03–1.72, <i>p</i> =0.155)	
	Bilateral	18 (1.5%)	2.15 (1.21–3.83, <i>p</i> =0.009)	1.38 (0.76–2.51, <i>p</i> =0.292)	6 (3.5%)	1.26 (0.51–3.14, <i>p</i> =0.620)	
Tumor size	<=20 mm	74 (6.1%)			12 (7.0%)		
	<=40 mm	383 (31.7%)	0.81 (0.60–1.08, <i>p</i> =0.147)		55 (32.0%)	1.14 (0.54–2.41, <i>p</i> =0.738)	
	<=60 mm	502 (41.6%)	0.85 (0.64–1.13, <i>p</i> =0.260)		73 (42.4%)	1.40 (0.67–2.92, <i>p</i> =0.369)	
	<=80 mm	234 (19.4%)	0.87 (0.64–1.18, <i>p</i> =0.355)		31 (18.0%)	1.45 (0.66–3.21, <i>p</i> =0.355)	
	>80 mm	14 (1.2%)	1.70 (0.91–3.18, <i>p</i> =0.096)		1 (0.6%)	1.72 (0.21–13.83, <i>p</i> =0.610)	
Combined summary stage	Regional	218 (18.1%)			27 (15.7%)		
	Localized	968 (80.2%)	0.61 (0.52–0.72, <i>p</i> <0.001)	0.63 (0.52–0.75, <i>p</i> <0.001)	142 (82.6%)	0.64 (0.40–1.01, <i>p</i> =0.055)	
	Distant	21 (1.7%)	2.37 (1.37–4.10, <i>p</i> =0.002)	1.94 (1.11–3.39, <i>p</i> =0.020)	3 (1.7%)	2.09 (0.49–8.94, <i>p</i> =0.322)	
Continued							

Variable		SEER set			Test set		
			HR (univariate)	HR (multivariate)		HR (univariate)	HR (multivariate)
Surgery	No	202 (16.7%)			26 (15.1%)		
	Local	196 (16.2%)	0.66 (0.53–0.82, $p < 0.001$)	0.69 (0.55–0.88, $p = 0.002$)	38 (22.1%)	0.78 (0.46–1.34, $p = 0.372$)	0.74 (0.42–1.33, $p = 0.316$)
	Subtotal	398 (33.0%)	0.54 (0.45–0.66, $p < 0.001$)	0.64 (0.52–0.78, $p < 0.001$)	52 (30.2%)	0.88 (0.53–1.47, $p = 0.625$)	1.06 (0.62–1.83, $p = 0.832$)
	Radical	374 (31.0%)	0.45 (0.37–0.55, $p < 0.001$)	0.54 (0.44–0.67, $p < 0.001$)	47 (27.3%)	0.61 (0.36–1.03, $p = 0.064$)	0.67 (0.38–1.18, $p = 0.166$)
	Gross_total	37 (3.1%)	0.42 (0.28–0.62, $p < 0.001$)	0.50 (0.33–0.75, $p < 0.001$)	9 (5.2%)	0.40 (0.16–0.98, $p = 0.044$)	0.30 (0.12–0.78, $p = 0.014$)
Radiotherapy	No/Unknown	220 (18.2%)			34 (19.8%)		
	Yes	987 (81.8%)	0.36 (0.31–0.43, $p < 0.001$)	0.62 (0.49–0.78, $p < 0.001$)	138 (80.2%)	0.63 (0.42–0.94, $p = 0.023$)	0.90 (0.54–1.51, $p = 0.700$)
Chemotherapy	No/Unknown	275 (22.8%)			43 (25.0%)		
	Yes	932 (77.2%)	0.37 (0.32–0.43, $p < 0.001$)	0.54 (0.43–0.66, $p < 0.001$)	129 (75.0%)	0.37 (0.26–0.54, $p < 0.001$)	0.39 (0.24–0.63, $p < 0.001$)

Table 4. Univariate and multivariate Cox regression analysis for the cancer-specific survival in the SEER and test set.

Outcomes	Before IPTW		After IPTW	
	HR (95%CI)	P value	HR (95%CI)	P value
SEER set				
No (ref)	1		1	
Yes	1.27 (1.03–1.57)	0.024	1.09 (0.84–1.41)	0.503
Test set				
No (ref)	1		1	
Yes	4.38 (1.76–10.90)	0.001	1.93 (1.30–2.86)	0.001

Table 5. Cox regression analysis of tumor history on the cancer-specific survival before and after IPTW. Significant values are in bold. Adjusted for year of diagnosis, age, gender, histologic type, primary site, laterality, tumor size, combined summary stage, surgery, radiotherapy, chemotherapy.

Machine learning based predictive model for CSS

As shown in Table 8, seven features, including age, primary site, laterality, combined summary stage, surgery, radiotherapy and chemotherapy, were statistically significant, as confirmed by univariate Cox regression, and were incorporated to establish the predictive models. Among the eight machine learning algorithm-based models, RSF, GBM, Lasso Cox, CoxBoost, Survival-SVM, XGBoost, SuperPC and plsRcox, the GBM model exhibited the best predictive performance in the training set at 6 months (AUC = 0.808, 95% CI: 0.770–0.847), 12 months (AUC = 0.755, 95% CI: 0.721–0.789) and 24 months (AUC = 0.692, 95% CI: 0.657–0.728) in the training set, as shown in Table 9; Fig. 5, and convincing predictive effectiveness was also confirmed in the validation and test sets. The utility of the GBM model for clinical decision making was suggested by the DCA curve (Fig. 6). The calibration curves of the GBM model in the training, validation, and test sets showed good concordance between the predicted and observed overall survival rates at 6 months, 12 months and 24 months (Fig. 7). As illustrated in Fig. 8, Kaplan-Meier analysis revealed a statistically significant stratification ($p < 0.05$) of CSS outcomes between the low-risk and high-risk subgroups in the training, validation, and test sets ($p < 0.05$).

Discussion

This study clarified the survival-associated factors of Asian patients with glioblastoma in the SEER database and our retrospective cohort and developed a novel machine learning predictive model based on the data from the SEER database, which was validated with retrospective cohort data. Age, histologic type, surgery and chemotherapy were identified as independent factors for OS and CSS in the SEER database and retrospective cohort, which was also validated by other studies^{11,12}. Maximal surgical resection followed by simultaneous radiotherapy and temozolomide chemotherapy is the recommended standard for the treatment of patients with glioblastoma^{12,13} but not all patients receive standard treatment for several reasons. Moreover, we found that tumour history was an important risk factor for the OS and CSS of patients with glioblastoma in our

Variables			HR (univariate)	HR (multivariate)
Year of diagnosis	<=2016	444 (52.5%)		
	>2016	401 (47.5%)	1.01 (0.86–1.18, <i>p</i> =0.900)	
Age	<=60	410 (48.5%)		
	>60	435 (51.5%)	2.04 (1.75–2.38, <i>p</i> <0.001)	1.76 (1.50–2.06, <i>p</i> <0.001)
Gender	Female	381 (45.1%)		
	Male	464 (54.9%)	1.02 (0.87–1.18, <i>p</i> =0.834)	
Tumor history	No	746 (88.3%)		
	Yes	99 (11.7%)	1.38 (1.09–1.73, <i>p</i> =0.006)	1.28 (1.01–1.63, <i>p</i> =0.039)
Histologic type	Glioblastoma_NOS	796 (94.2%)		
	Giant cell glioblastoma	10 (1.2%)	0.58 (0.29–1.16, <i>p</i> =0.125)	0.72 (0.35–1.45, <i>p</i> =0.357)
	Gliosarcoma	21 (2.5%)	1.88 (1.20–2.93, <i>p</i> =0.006)	1.89 (1.20–2.97, <i>p</i> =0.006)
	Glioblastoma_IDH-mutant	18 (2.1%)	0.32 (0.15–0.68, <i>p</i> =0.003)	0.50 (0.23–1.06, <i>p</i> =0.071)
Primary site	Cerebrum	43 (5.1%)		
	Frontal lobe	318 (37.6%)	0.85 (0.58–1.25, <i>p</i> =0.410)	
	Temporal lobe	198 (23.4%)	0.86 (0.58–1.27, <i>p</i> =0.444)	
	Parietal lobe	141 (16.7%)	1.05 (0.70–1.56, <i>p</i> =0.818)	
	Occipital lobe	18 (2.1%)	0.90 (0.48–1.69, <i>p</i> =0.738)	
	Ventricle	2 (0.2%)	0.53 (0.13–2.21, <i>p</i> =0.383)	
	Overlapping lesion of brain	91 (10.8%)	0.90 (0.59–1.38, <i>p</i> =0.637)	
	Brain_NOS	34 (4.0%)	1.44 (0.87–2.38, <i>p</i> =0.155)	
Laterality	Left	401 (47.5%)		
	Right	426 (50.4%)	0.92 (0.79–1.07, <i>p</i> =0.288)	0.99 (0.85–1.16, <i>p</i> =0.898)
	Midline	4 (0.5%)	0.70 (0.22–2.20, <i>p</i> =0.542)	0.75 (0.24–2.37, <i>p</i> =0.625)
	Bilateral	14 (1.7%)	2.30 (1.29–4.10, <i>p</i> =0.005)	1.60 (0.88–2.92, <i>p</i> =0.125)
Tumor size	<=20 mm	55 (6.5%)		
	<=40 mm	257 (30.4%)	0.78 (0.57–1.07, <i>p</i> =0.123)	0.89 (0.65–1.23, <i>p</i> =0.484)
	<=60 mm	344 (40.7%)	0.86 (0.63–1.17, <i>p</i> =0.335)	1.05 (0.76–1.44, <i>p</i> =0.784)
	<=80 mm	178 (21.1%)	0.84 (0.60–1.17, <i>p</i> =0.301)	1.07 (0.76–1.50, <i>p</i> =0.704)
	>80 mm	11 (1.3%)	2.02 (1.02–4.00, <i>p</i> =0.045)	2.15 (1.07–4.35, <i>p</i> =0.032)
Combined summary stage	Regional	155 (18.3%)		
	Localized	676 (80.0%)	0.63 (0.52–0.77, <i>p</i> <0.001)	0.69 (0.56–0.85, <i>p</i> <0.001)
	Distant	14 (1.7%)	2.76 (1.55–4.91, <i>p</i> <0.001)	2.29 (1.27–4.12, <i>p</i> =0.006)
Surgery	No	145 (17.2%)		
	Local	142 (16.8%)	0.66 (0.52–0.85, <i>p</i> =0.001)	0.65 (0.50–0.84, <i>p</i> =0.001)
	Subtotal	263 (31.1%)	0.53 (0.42–0.66, <i>p</i> <0.001)	0.59 (0.47–0.75, <i>p</i> <0.001)
	Radical	266 (31.5%)	0.43 (0.35–0.55, <i>p</i> <0.001)	0.52 (0.41–0.65, <i>p</i> <0.001)
	Gross_total	29 (3.4%)	0.44 (0.29–0.68, <i>p</i> <0.001)	0.48 (0.31–0.75, <i>p</i> =0.001)
Radiotherapy	No/Unknown	168 (19.9%)		
	Yes	677 (80.1%)	0.39 (0.32–0.46, <i>p</i> <0.001)	0.76 (0.59–0.98, <i>p</i> =0.035)
Chemotherapy	No/Unknown	198 (23.4%)		
	Yes	647 (76.6%)	0.33 (0.28–0.40, <i>p</i> <0.001)	0.47 (0.37–0.60, <i>p</i> <0.001)

Table 6. Univariate and multivariate Cox regression analysis for the overall survival in the training set.

retrospective cohort before and after IPTW, which has not yet been widely recognized. In the research of Ebad Ur Rehman et al., tumour history was an independent OS-associated factor, and the most frequent prior cancers reported were genitourinary, breast, haematologic and lymphatic, and gastrointestinal malignancies; however, the exact mechanism underlying the connection was unclear¹⁴. A common hypothesis is a shared genetic pathway, and Turcotte et al. reported that primary cancers can give rise to syndromes that show an autosomal dominant pattern of inheritance, such as gliomas¹⁵. Another hypothesis is that an important risk factor for the development of CNS tumours in cancer survivors is prior cranial radiotherapy and chemotherapy, which has been validated by various studies^{16–20}. In the retrospective study for the second primary gliomas conducted by Maluf et al., most had received chemotherapy and/or radiation before their glioma diagnosis²¹. What's more, both intravenous and intrathecal methotrexate administration have been found to be associated to an increased risk of meningioma development^{22,23}. Diagnostic delays in some Chinese patient populations often result in initial cancer detection at advanced stages, consequently strengthening the observed correlation between prior tumor history and adverse prognosis. The relevant connections and underlying mechanisms still need to be explored in depth.

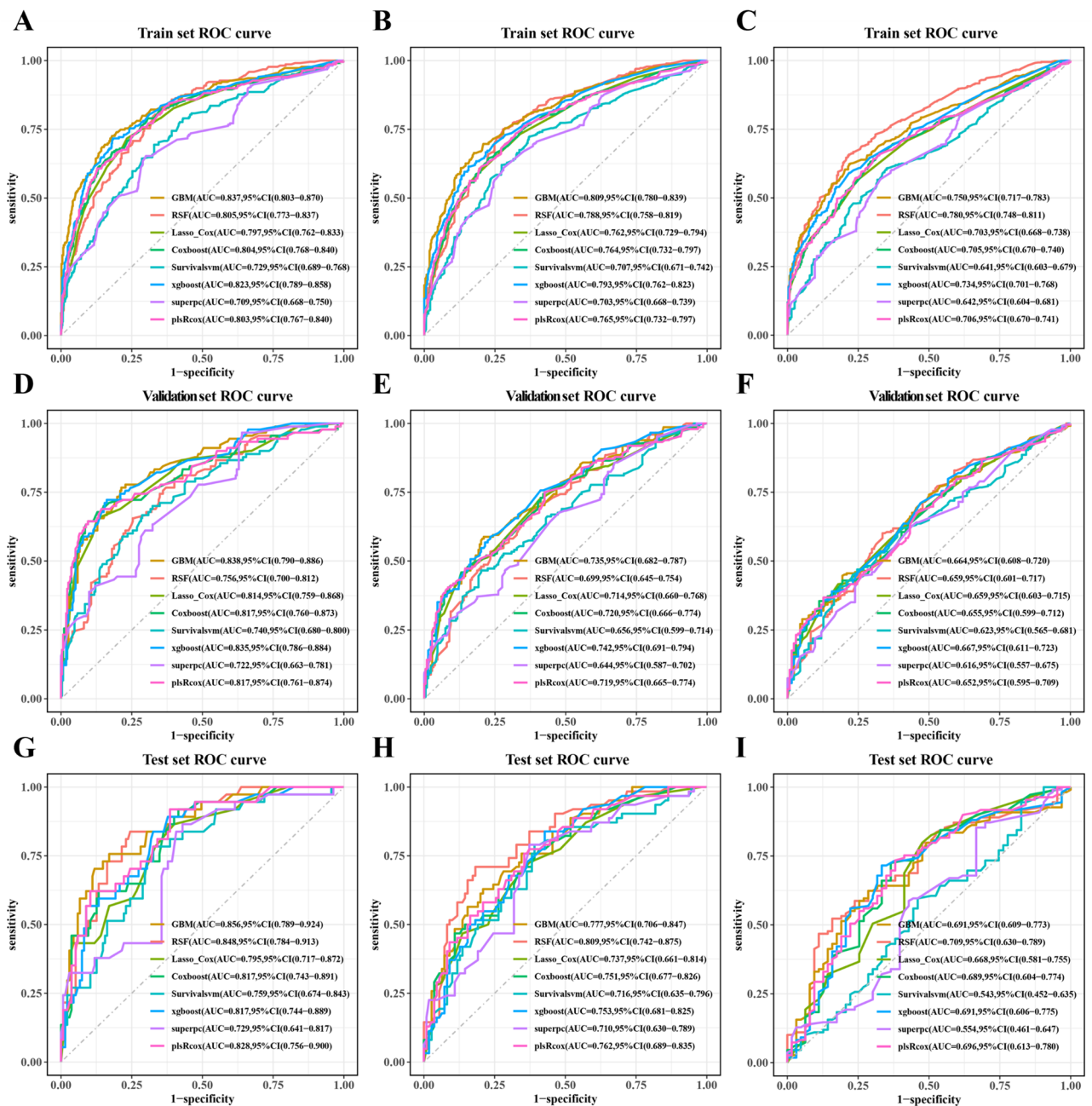


Fig. 1. Receiver operating characteristic curves of the 6-month, 12-month, and 24-month OS prediction models for glioblastoma patients in the training (A–C), validation (D–F), and test (G–I) set.

We were the first to apply machine learning models in the construction of predictive models for patients with glioblastoma based on clinical data. The predictive performance of the generalized boosted regression modelling algorithm was more stable and better than those of the other algorithms, and the overall survival of patients with glioblastoma reached AUROCs of 0.856, 0.777 and 0.691 for 6-month, 12-month and 24-month survival, respectively, in the test set. The predictive model for cancer-specific survival attained AUROCs of 0.805, 0.736 and 0.703 for 6-month, 12-month and 24-month survival, respectively, in the test set. The performance of our machine learning-based predictive models was slightly preferable to that of the prognostic nomogram models, which were developed on the clinical data of Niu et al. using traditional Cox regression analysis²⁴. A variety of predictive models have been created as artificial intelligence continues to evolve. The radiomics-based machine learning prediction model using pretreatment multiparametric magnetic resonance imaging achieved AUROCs of 0.791 and 0.708 in internal and external validation, respectively²⁵. The SVM and LR pathomics models based on glioblastoma tissue and normal brain tissue were constructed with AUC values of 0.779 and 0.785, respectively²⁶. The effectiveness of multimodal prediction tends to be better than that of a single type of data. The comprehensive clinical-radiomics-TME model for predicting OS constructed by Zhou et al. exhibited

	6-month			12-month			24-month		
	Training set	Validation set	Test set	Training set	Validation set	Test set	Training set	Validation set	Test set
GBM	0.837	0.838	0.856	0.809	0.735	0.777	0.750	0.664	0.691
RSF	0.805	0.756	0.848	0.788	0.699	0.809	0.780	0.659	0.709
Lasso Cox	0.797	0.814	0.795	0.762	0.714	0.737	0.703	0.659	0.668
Coxboost	0.804	0.817	0.817	0.764	0.720	0.751	0.705	0.655	0.689
Survivalsvm	0.729	0.740	0.759	0.707	0.656	0.716	0.641	0.623	0.543
Xgboost	0.823	0.835	0.817	0.793	0.742	0.753	0.734	0.667	0.691
Superpc	0.709	0.722	0.729	0.703	0.644	0.710	0.642	0.616	0.554
PlsRcox	0.803	0.817	0.828	0.765	0.719	0.762	0.706	0.652	0.696

Table 7. Area under curve of the predictive models based on eight machine learning algorithms for overall survival.

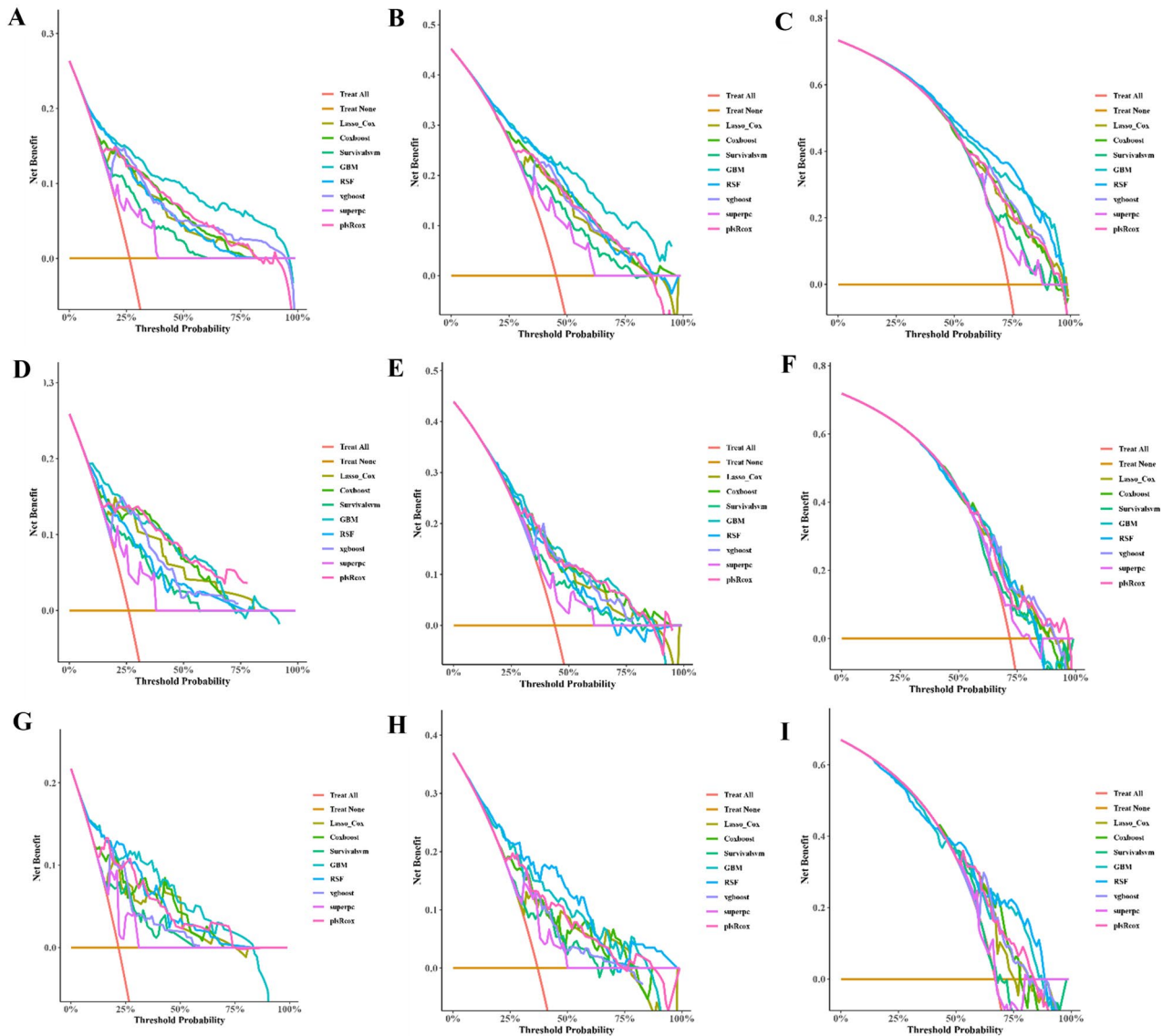


Fig. 2. Decision curve analyses (DCAs) of the 6-month, 12-month, and 24-month OS prediction models for glioblastoma patients in the training (A-C), validation (D-F), and test (G-I) set.

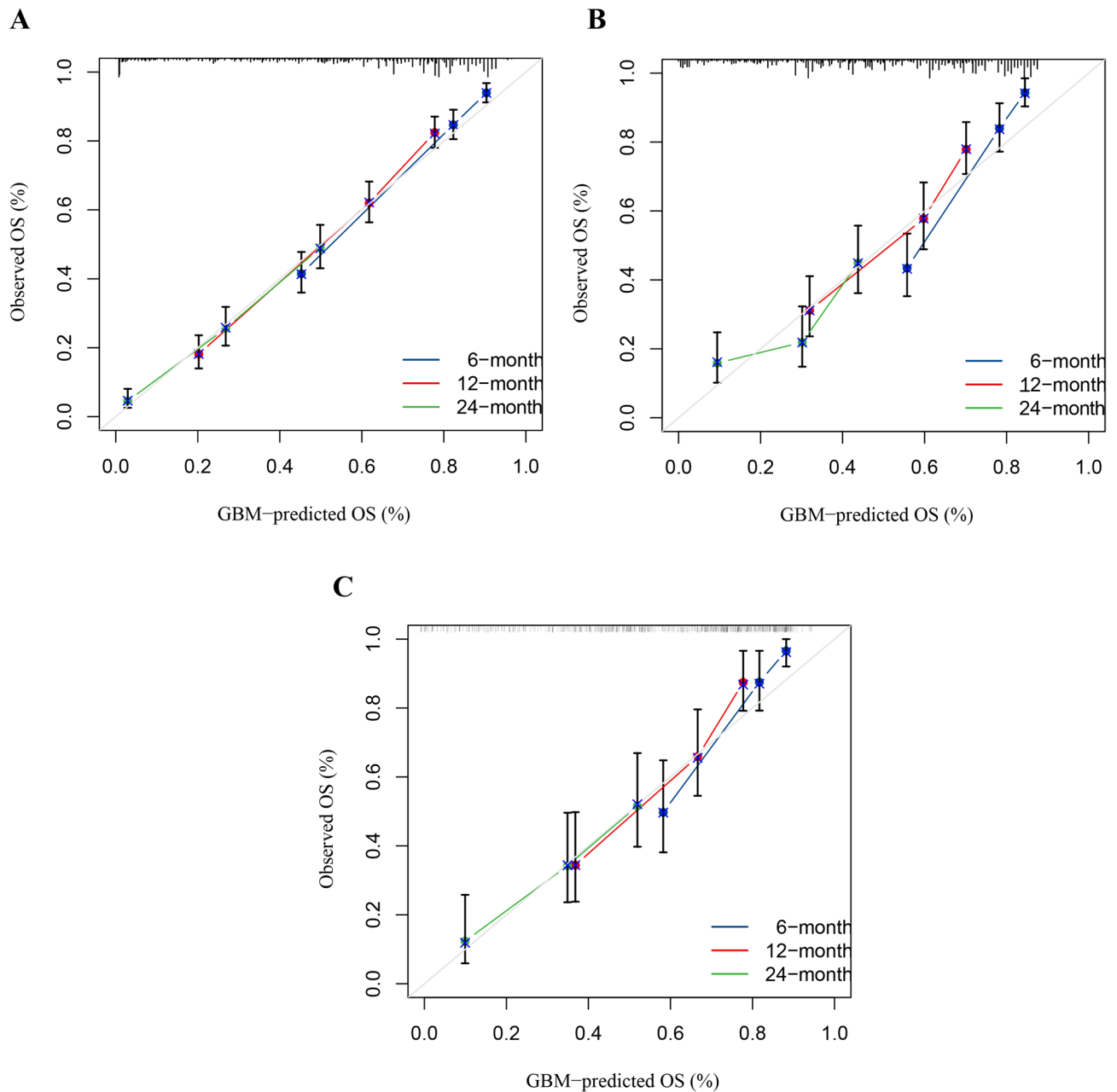


Fig. 3. Calibration curves of 6-month, 12-month, and 24-month OS prediction in the training (A), validation (B), and test (C) set.

satisfactory survival prediction ability, with AUROCs of 1-, 2-, and 3-year OS prediction of 0.842, 0.844, and 0.795, respectively, in the test set²⁷. The combination of clinical data and genetic information through artificial neural networks accurately predicts 15-month survival, with an accuracy rate of 83.3% and an AUROC of 0.81²⁸. Combining multiple types of data, including clinical, radiomics, pathomics and genetic data, may help further improve the predictive efficacy of prognostic models.

There are several limitations to our study. Since the patient data in the training and validation sets were collected from the SEER database, the retrospective nature and missing data were unavoidable, and the single-centre data of patients with glioblastoma also limited the sample size of the test set, which may have led to bias or heterogeneity. Thus, large-scale prospective studies and detailed characteristic analyses, such as analyses of chemotherapeutic regimens, are needed to explore and verify the prognostic factors in our centre and construct a predictive model with greater predictive effectiveness and the capability for further survival prediction. Moreover, the combination of additional data, including radiomics, pathomics and genetic data, may contribute to further enhancing the predictive power of the model.

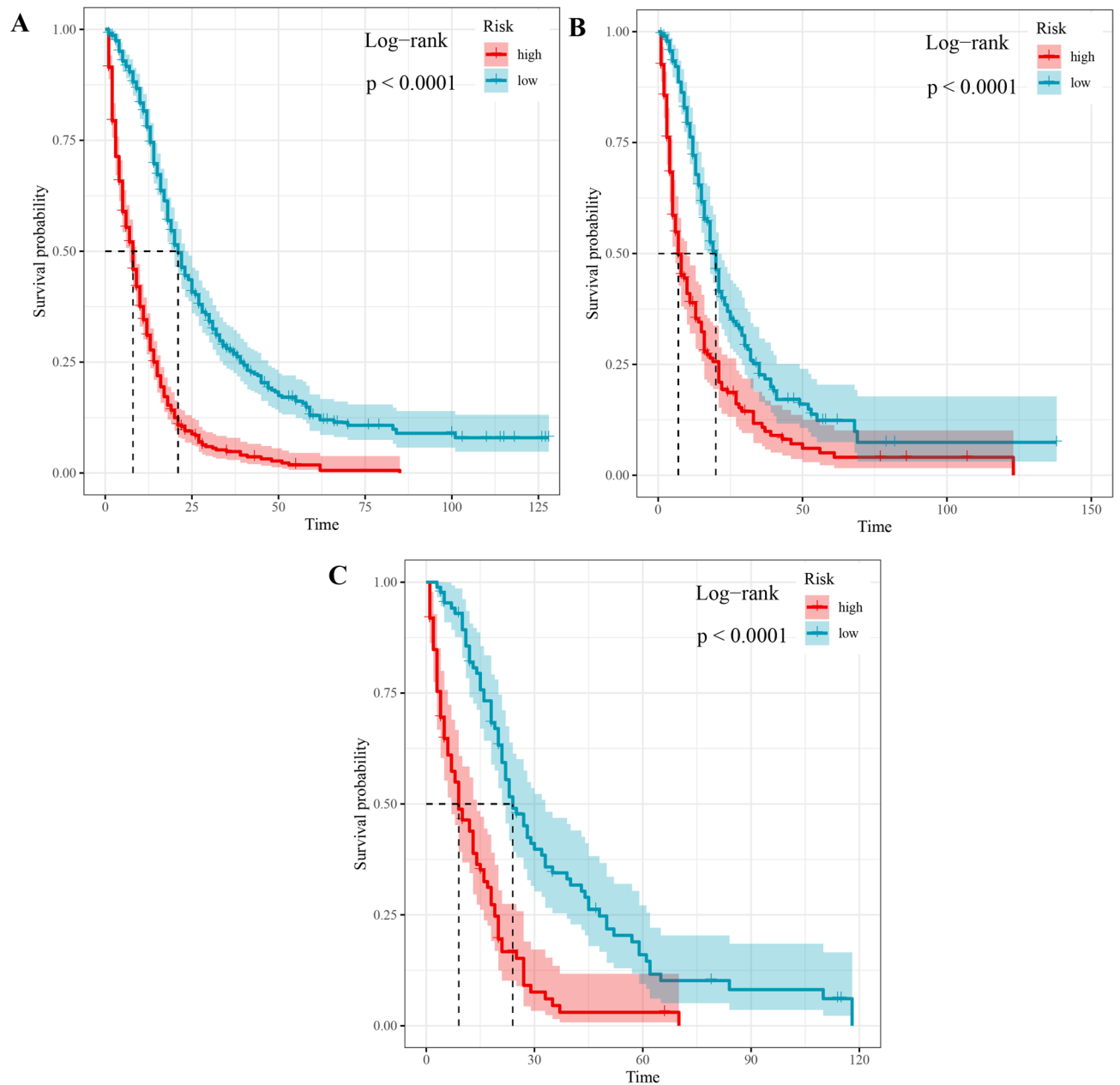


Fig. 4. The survival curves between low-risk and high-risk glioblastoma patients for OS.

Conclusion

Age, tumour history, histologic type, surgery and chemotherapy were confirmed to be independent factors for OS, and age, tumour history, histologic type, surgery and chemotherapy were identified as prognostic factors for CSS in our retrospective study. The predictive model constructed for OS and CSS based on the generalized boosted regression modelling algorithm in Asian patients with glioblastoma can be used to accurately predict OS and CSS in clinical practice, which may help tailor personalized treatment regimens and provide significant benefits for Asian patients with glioblastoma.

Variables			HR (univariate)	HR (multivariate)
Year of diagnosis	<=2016	439 (52.0%)		
	>2016	406 (48.0%)	0.91 (0.77–1.08, <i>p</i> =0.291)	
Age	<=60	407 (48.2%)		
	>60	438 (51.8%)	1.89 (1.60–2.22, <i>p</i> <0.001)	1.66 (1.40–1.97, <i>p</i> <0.001)
Gender	Female	382 (45.2%)		
	Male	463 (54.8%)	1.02 (0.87–1.20, <i>p</i> =0.769)	
Tumor history	No	759 (89.8%)		
	Yes	86 (10.2%)	1.27 (0.98–1.66, <i>p</i> =0.075)	
Histologic type	Glioblastoma_NOS	802 (94.9%)		
	Giant cell glioblastoma	8 (0.9%)	0.46 (0.17–1.23, <i>p</i> =0.122)	
	Gliosarcoma	22 (2.6%)	1.44 (0.86–2.40, <i>p</i> =0.164)	
	Glioblastoma_IDH-mutant	13 (1.5%)	0.43 (0.18–1.04, <i>p</i> =0.062)	
Primary site	Cerebrum	40 (4.7%)		
	Frontal lobe	312 (36.9%)	1.49 (0.95–2.34, <i>p</i> =0.081)	1.60 (1.01–2.56, <i>p</i> =0.047)
	Temporal lobe	212 (25.1%)	1.42 (0.90–2.24, <i>p</i> =0.133)	1.63 (1.01–2.64, <i>p</i> =0.046)
	Parietal lobe	143 (16.9%)	1.71 (1.07–2.74, <i>p</i> =0.025)	1.71 (1.05–2.78, <i>p</i> =0.031)
	Occipital lobe	18 (2.1%)	1.73 (0.85–3.53, <i>p</i> =0.130)	1.89 (0.91–3.92, <i>p</i> =0.088)
	Ventricle	2 (0.2%)	0.93 (0.12–6.92, <i>p</i> =0.943)	1.50 (0.20–11.22, <i>p</i> =0.694)
	Overlapping lesion of brain	85 (10.1%)	1.63 (0.99–2.67, <i>p</i> =0.053)	1.84 (1.10–3.05, <i>p</i> =0.019)
	Brain_NOS	400 (47.3%)	2.09 (1.17–3.74, <i>p</i> =0.013)	1.57 (0.86–2.85, <i>p</i> =0.141)
Laterality	Left	430 (50.9%)		
	Right	3 (0.4%)	0.91 (0.77–1.07, <i>p</i> =0.251)	0.93 (0.79–1.10, <i>p</i> =0.418)
	Midline	12 (1.4%)	0.61 (0.15–2.48, <i>p</i> =0.495)	1.02 (0.25–4.17, <i>p</i> =0.980)
	Bilateral	12 (1.4%)	2.07 (1.02–4.18, <i>p</i> =0.044)	1.27 (0.60–2.68, <i>p</i> =0.528)
Tumor size	<=20 mm	48 (5.7%)		
	<=40 mm	263 (31.1%)	0.80 (0.55–1.15, <i>p</i> =0.220)	
	<=60 mm	359 (42.5%)	0.86 (0.61–1.23, <i>p</i> =0.418)	
	<=80 mm	166 (19.6%)	0.95 (0.65–1.38, <i>p</i> =0.771)	
	>80 mm	9 (1.1%)	1.75 (0.81–3.77, <i>p</i> =0.154)	
Combined summary stage	Regional	145 (17.2%)	0.80 (0.55–1.15, <i>p</i> =0.220)	
	Localized	685 (81.1%)	0.68 (0.55–0.84, <i>p</i> <0.001)	0.70 (0.56–0.87, <i>p</i> =0.002)
	Distant	15 (1.8%)	2.63 (1.32–5.23, <i>p</i> =0.006)	2.82 (1.40–5.67, <i>p</i> =0.004)
Surgery	No	131 (15.5%)		
	Local	126 (14.9%)	0.69 (0.53–0.91, <i>p</i> =0.009)	0.64 (0.48–0.86, <i>p</i> =0.003)
	Subtotal	286 (33.8%)	0.56 (0.44–0.71, <i>p</i> <0.001)	0.56 (0.44–0.72, <i>p</i> <0.001)
	Radical	269 (31.8%)	0.42 (0.33–0.54, <i>p</i> <0.001)	0.45 (0.35–0.58, <i>p</i> <0.001)
	Gross_total	33 (3.9%)	0.41 (0.26–0.63, <i>p</i> <0.001)	0.43 (0.28–0.68, <i>p</i> <0.001)
Radiotherapy	No/Unknown	150 (17.8%)		
	Yes	695 (82.2%)	0.40 (0.32–0.48, <i>p</i> <0.001)	0.67 (0.51–0.87, <i>p</i> =0.002)
Chemotherapy	No/Unknown	188 (22.2%)		
	Yes	657 (77.8%)	0.39 (0.33–0.47, <i>p</i> <0.001)	0.53 (0.42–0.67, <i>p</i> <0.001)

Table 8. Univariate and multivariate Cox regression analysis for the cancer-specific survival in the training set.

	6-month			12-month			24-month		
	Training set	Validation set	Test set	Training set	Validation set	Test set	Training set	Validation set	Test set
GBM	0.808	0.830	0.805	0.755	0.774	0.736	0.692	0.687	0.703
RSF	0.772	0.722	0.748	0.760	0.669	0.723	0.753	0.611	0.690
Lasso Cox	0.784	0.839	0.812	0.719	0.770	0.729	0.651	0.670	0.652
Coxboost	0.784	0.843	0.813	0.723	0.781	0.739	0.654	0.686	0.664
Survivalsvm	0.709	0.731	0.740	0.634	0.668	0.679	0.575	0.546	0.525
Xgboost	0.798	0.843	0.804	0.738	0.779	0.724	0.672	0.681	0.679
Superpc	0.554	0.493	0.494	0.574	0.463	0.482	0.578	0.563	0.480
PlsRcox	0.782	0.841	0.810	0.723	0.784	0.738	0.655	0.691	0.666

Table 9. Area under curve of the predictive models based on eight machine learning algorithms for cancer-specific survival.

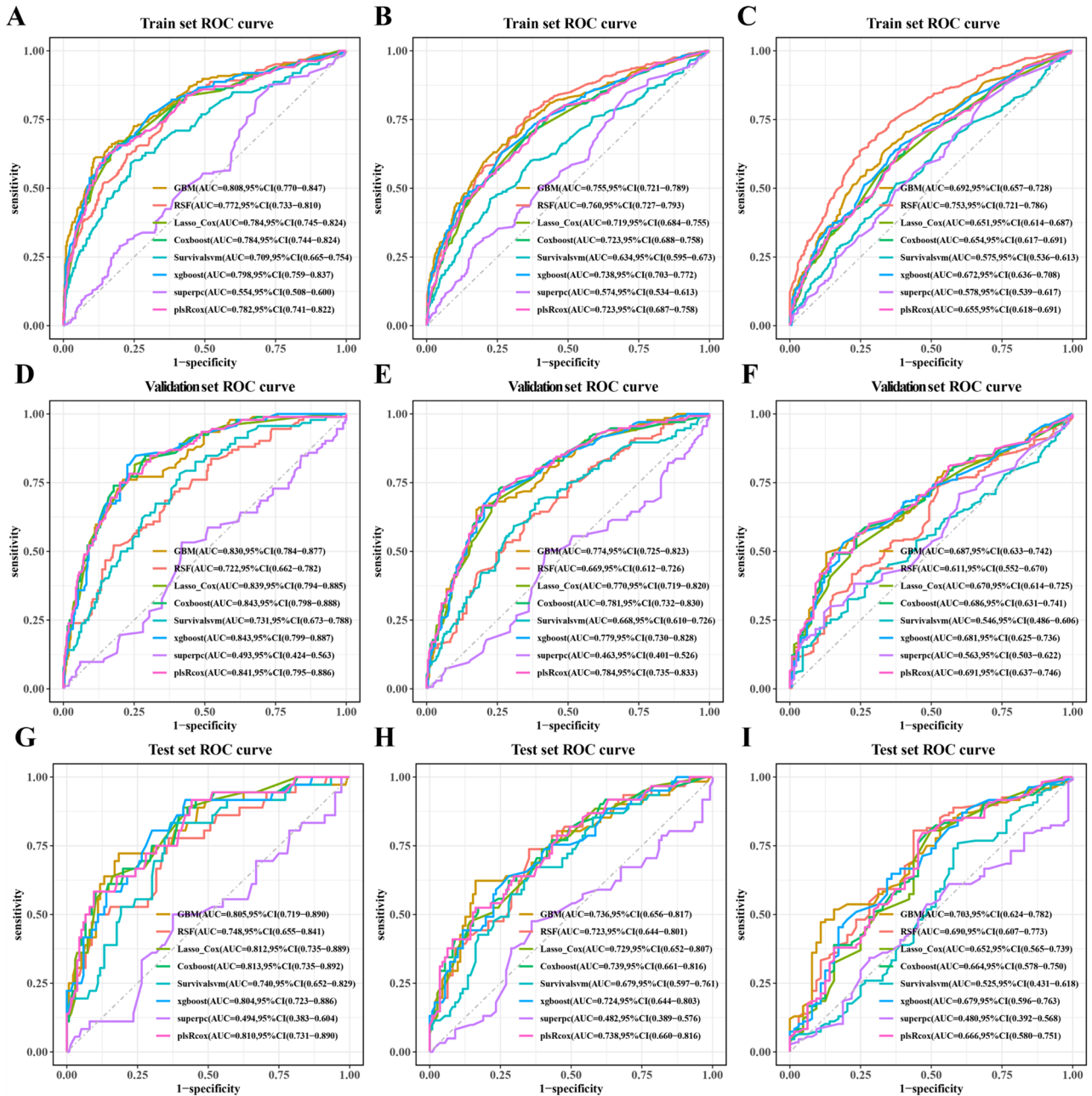


Fig. 5. Receiver operating characteristic curves of the 6-month, 12-month, and 24-month CSS prediction models for glioblastoma patients in the training (A-C), validation (D-F), and test (G-I) set.

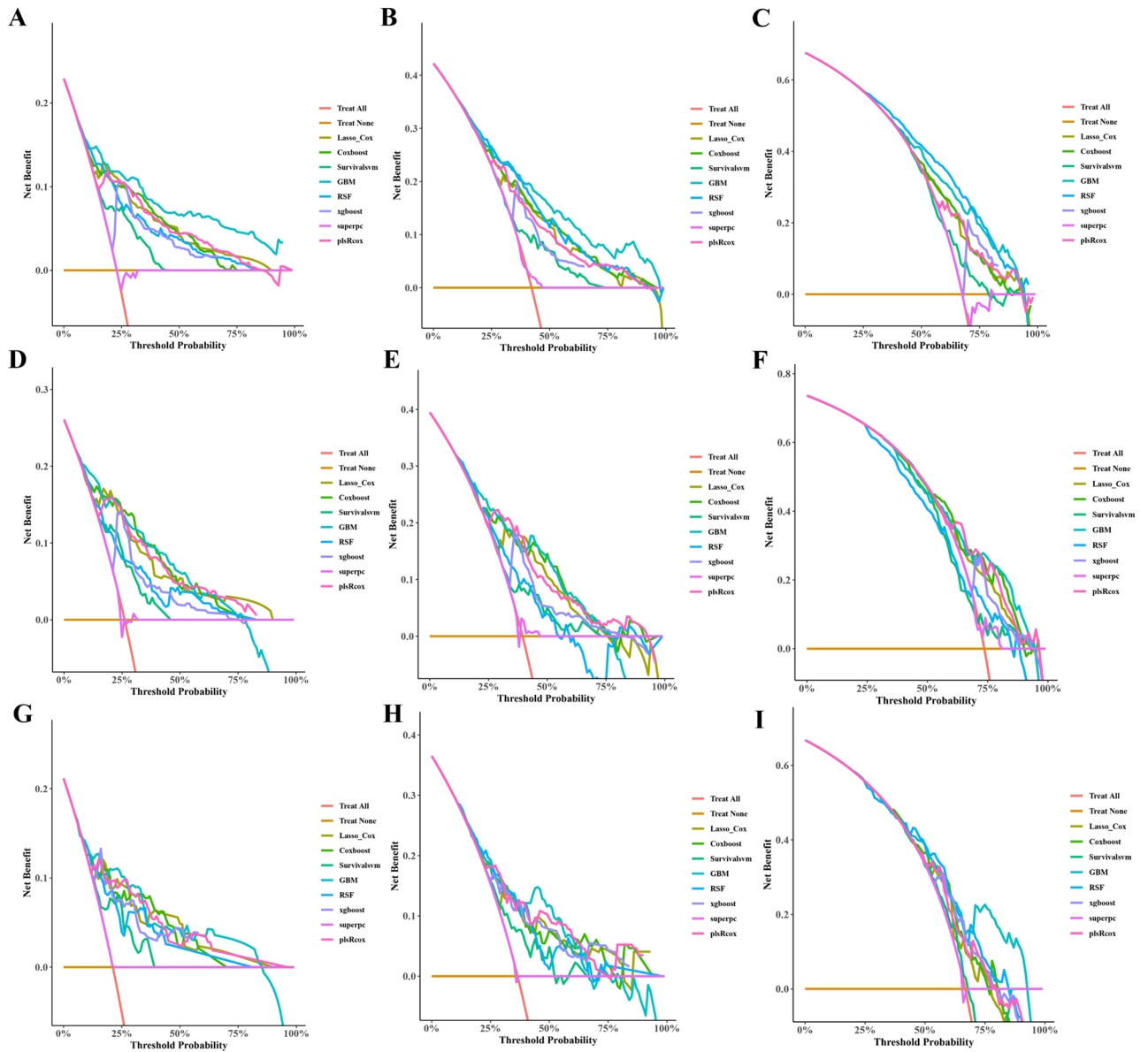


Fig. 6. Decision curve analyses (DCAs) of the 6-month, 12-month, and 24-month CSS prediction models for glioblastoma patients in the training (A-C), validation (D-F), and test (G-I) set.

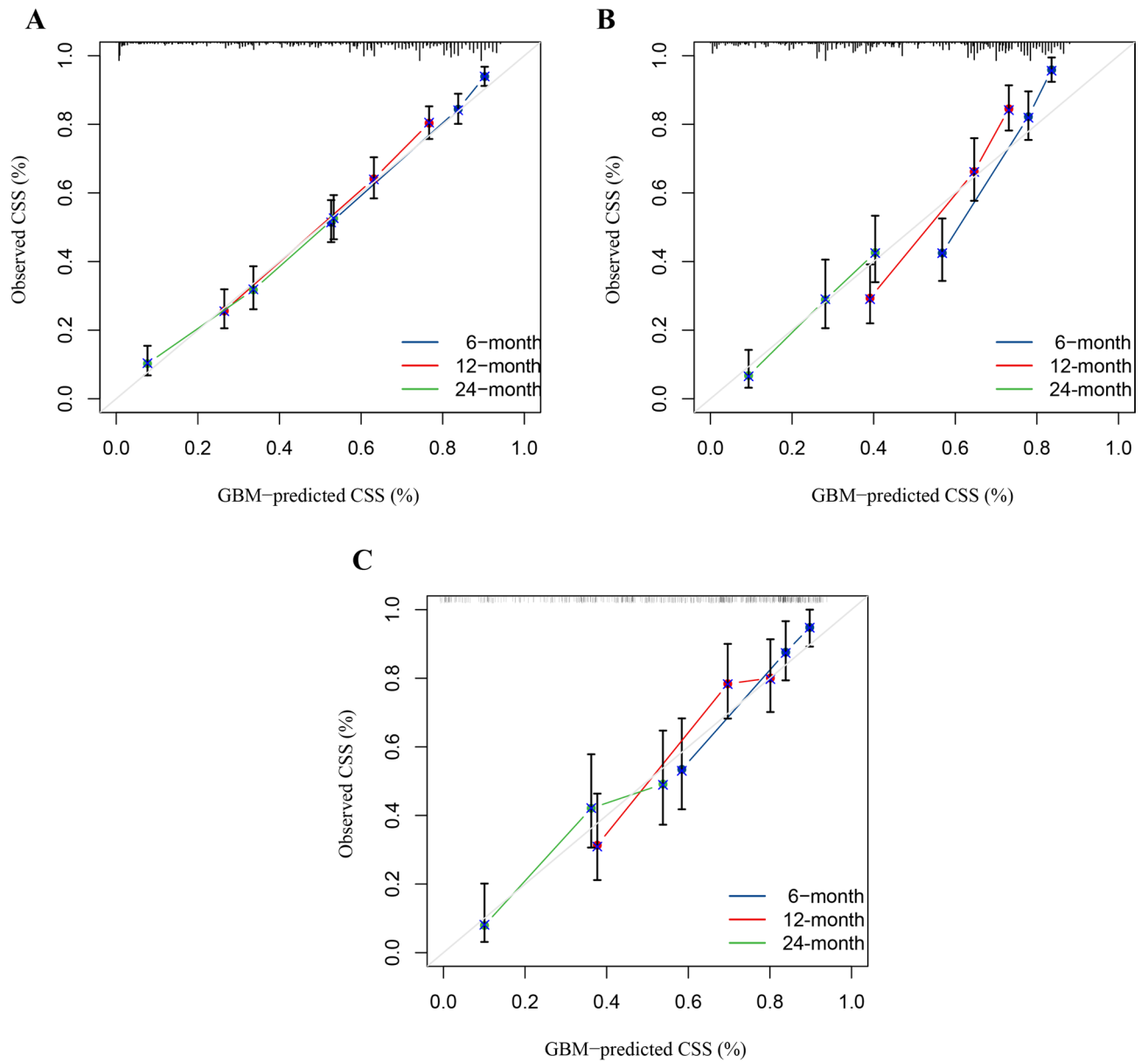


Fig. 7. Calibration curves of 6-month, 12-month, and 24-month CSS prediction in the training (A), validation (B), and test (C) set.

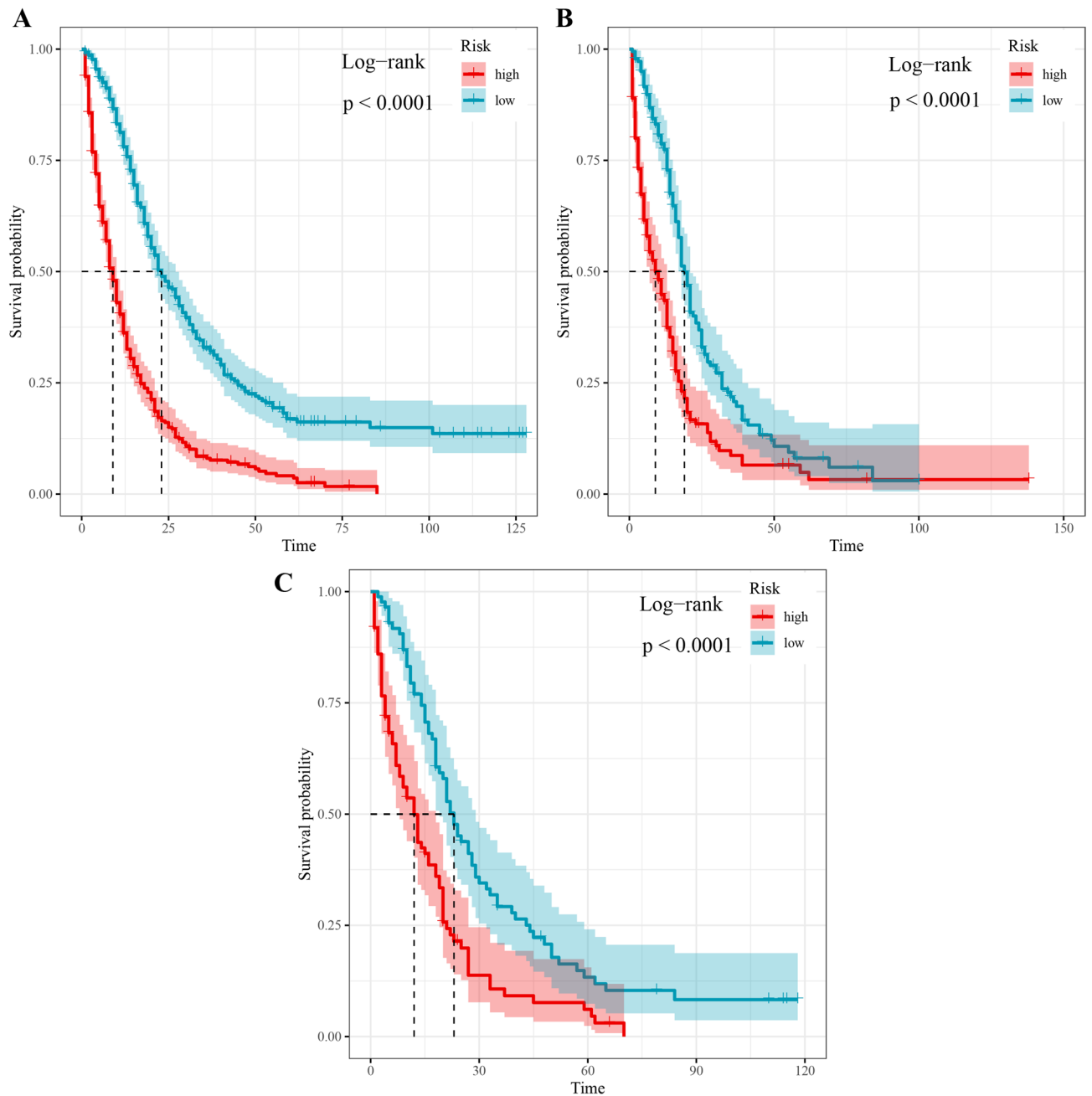


Fig. 8. The survival curves between low-risk and high-risk glioblastoma patients for CSS.

Data availability

For patient privacy reasons, datasets generated and/or analyzed in this study are not publicly available, but are available from the corresponding authors upon request.

Received: 21 March 2025; Accepted: 8 August 2025

Published online: 24 August 2025

References

- Ostrom, Q. T., Gittleman, H., Kruchko, C. & Barnholtz-Sloan, J. S. Primary brain and other central nervous system tumors in appalachia: regional differences in incidence, mortality, and survival. *J. Neurooncol.* **142**, 27–38 (2019).
- Lapointe, S., Perry, A. & Butowski, N. A. Primary brain tumours in adults. *Lancet* **392**, 432–446 (2018).
- Grochans, S. et al. Baranowska-Bosiacka, I. Epidemiology of glioblastoma Multiforme-Literature review. *Cancers (Basel)*. **14**, 2412 (2022).
- Louis, D. N. et al. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).

5. Stupp, R. et al. Cilengitide combined with standard treatment for patients with newly diagnosed glioblastoma with methylated MGMT promoter (CENTRIC EORTC 26071–22072 study): a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol.* **15**, 1100–1108 (2014).
6. Weller, M. et al. EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *Lancet Oncol.* **15**, e395–403 (2014).
7. Stupp, R. et al. Maintenance therapy with Tumor-Treating fields plus Temozolomide vs Temozolomide alone for glioblastoma: A randomized clinical trial. *JAMA* **314**, 2535–2543 (2015).
8. Skrede, O. J. et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* **395**, 350–360 (2020).
9. Le Berre, C. et al. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology* **158**, 76–94e2 (2020).
10. Achilonu, O. J. et al. Predicting colorectal cancer recurrence and patient survival using supervised machine learning approach: A South African Population-Based study. *Front. Public Health.* **9**, 694306 (2021).
11. Kim, M. et al. Glioblastoma as an age-related neurological disorder in adults. *Neurooncol Adv.* **3**, vdab125 (2021).
12. Gilbert, M. R. & Lohin, M. The treatment of malignant gliomas. *Curr. Treat. Options Neurol.* **7**, 293–303 (2005).
13. Zhao, M. J., Lu, T., Ma, C., Wang, Z. F. & Li, Z. Q. A narrative review on the management of glioblastoma in China. *Chin. Clin. Oncol.* **11**, 29 (2022).
14. Ebad Ur Rehman, M. et al. Impact of prior cancer history on survival in brain malignancy: A propensity score-adjusted, population-based study. *Cancer Rep (Hoboken)* **7**, e (2024). (1984).
15. Turcotte, L. M. et al. Risk factors, and surveillance of subsequent malignant neoplasms in survivors of childhood cancer: A review. *J. Clin. Oncol.* **36**, 2145–2152 (2018). Risk.
16. Bowers, D. C. et al. Subsequent neoplasms of the CNS among survivors of childhood cancer: a systematic review. *Lancet Oncol.* **14**, e321–328 (2013).
17. Yamanaka, R. & Hayano, A. Secondary glioma following acute lymphocytic leukemia: therapeutic implications. *Neurosurg. Rev.* **40**, 549–557 (2017).
18. Pettorini, B. L. et al. Radiation-induced brain tumours after central nervous system irradiation in childhood: a review. *Childs Nerv. Syst.* **24**, 793–805 (2008).
19. Relling, M. V. et al. High incidence of secondary brain tumours after radiotherapy and antimetabolites. *Lancet* **354**, 34–39 (1999).
20. Little, M. P. et al. Risks of brain tumour following treatment for cancer in childhood: modification by genetic factors, radiotherapy and chemotherapy. *Int. J. Cancer.* **78**, 269–275 (1998).
21. Maluf, F. C., DeAngelis, L. M., Raizer, J. J. & Abrey, L. E. High-grade gliomas in patients with prior systemic malignancies. *Cancer* **94**, 3219–3224 (2002).
22. Taylor, A. J. et al. Population-based risks of CNS tumors in survivors of childhood cancer: the British childhood cancer survivor study. *J. Clin. Oncol.* **28**, 5287–5293 (2010).
23. Chojnacka, M. et al. Second brain tumors following central nervous system radiotherapy in childhood. *Br. J. Radiol.* **87**, 20140211 (2014).
24. Niu, X., Chang, T., Yang, Y. & Mao, Q. Prognostic nomogram models for predicting survival probability in elderly glioblastoma patients. *J. Cancer Res. Clin. Oncol.* **149**, 14145–14157 (2023).
25. Karabacak, M., Patil, S., Gersey, Z. C., Komotar, R. J. & Margetis, K. Radiomics-Based machine learning with natural gradient boosting for continuous survival prediction in glioblastoma. *Cancers (Basel).* **16**, 3614 (2024).
26. Li, W. et al. Pathomics models for CD40LG expression and prognosis prediction in glioblastoma. *Sci. Rep.* **14**, 24350 (2024).
27. Zhou, Q. et al. Integrated MRI radiomics, tumor microenvironment, and clinical risk factors for improving survival prediction in patients with glioblastomas. *Strahlenther Onkol.* **201**, 398–410 (2024).
28. Azimi, P., Yazdani, T., Zohrevand, A. & Ahmadiani, A. Predicting survival in glioblastoma using gene expression databases: A neural network analysis. *Int. J. Mol. Cell. Med.* **13**, 79–90 (2024).

Author contributions

YTH and NWC designed the study, revised the manuscript and was responsible to undertake project leadership and guarantee this study. DLL, LXZ and LFX contributed to write this manuscript and analyzed the data of patient, RHZ, HYG, JLG and MHW collected the data of patients. All authors contributed to the article and approved the final version of manuscript for submission.

Funding

The present study was supported by Dalian Institute of Chemical Physics-Dalian University of Medical Science and Technology Joint Innovation Fund (DMU-2&DICP UN202309) and Liaoning Provincial Science and Technology Department (2022-MS-15, 2022-MS-16).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

The Ethics Committee of the Second Affiliated Hospital of Dalian Medical University approved the research.

Patient consent for publication

The patient provided informed consent.

Additional information

Correspondence and requests for materials should be addressed to N.C. or Y.H.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025