Neuro-Oncology Advances

6(1), vdae096, 2024 | https://doi.org/10.1093/noajnl/vdae096 | Advance Access date 11 June 2024

Advancing precision prognostication in neuro-oncology: Machine learning models for data-driven personalized survival predictions in IDH-wildtype glioblastoma

Mert Karabacak, Pemla Jagtiani, Long Di, Ashish H. Shah, Ricardo J. Komotar, and Konstantinos Margetis

All author affiliations are listed at the end of the article

Corresponding Author: Konstantinos Margetis, MD, PhD, Department of Neurosurgery, Mount Sinai Health System, 1468 Madison Ave, New York, NY 10029, USA (konstantinos.margetis@mountsinai.org).

Abstract

Background. Glioblastoma (GBM) remains associated with a dismal prognoses despite standard therapies. While population-level survival statistics are established, generating individualized prognosis remains challenging. We aim to develop machine learning (ML) models that generate personalized survival predictions for GBM patients to enhance prognostication.

Methods. Adult patients with histologically confirmed IDH-wildtype GBM from the National Cancer Database (NCDB) were analyzed. ML models were developed with TabPFN, TabNet, XGBoost, LightGBM, and Random Forest algorithms to predict mortality at 6, 12, 18, and 24 months postdiagnosis. SHapley Additive exPlanations (SHAP) were employed to enhance the interpretability of the models. Models were primarily evaluated using the area under the receiver operating characteristic (AUROC) values, and the top-performing models indicated by the highest AUROCs for each outcome were deployed in a web application that was created for individualized predictions.

Results. A total of 7537 patients were retrieved from the NCDB. Performance evaluation revealed the top-performing models for each outcome were built using the TabPFN algorithm. The TabPFN models yielded mean AUROCs of 0.836, 0.78, 0.732, and 0.724 in predicting 6, 12, 18, and 24 month mortality, respectively.

Conclusions. This study establishes ML models tailored to individual patients to enhance GBM prognostication. Future work should focus on external validation and dynamic updating as new data emerge.

Key Points

- This study developed machine learning models predicting survival outcomes postdiagnosis for IDH-wildtype glioblastoma patients.
- Machine learning has the potential to transform cancer prognostication from population statistics to patient-tailored predictions.

Glioblastoma (GBM) represents the most prevalent and aggressive form of primary brain tumors in adults.[1,2] Despite incremental advancements in standard-of-care involving maximal safe resection followed by concurrent chemoradiotherapy, GBM continues to be associated with a dismal prognosis, with a median survival of only 12–18 months postdiagnosis.[2–5] As new treatments emerge, the overall survival for patients may increase. For example, the addition of tumor-treating fields (TTF) to maintenance temozolomide chemotherapy has been shown to extend median overall survival to 20.9 months.[6] The complex progression patterns coupled with inadequate response to current treatment modalities pose challenges for prognostication in GBMs. Therefore, accurately forecasting survival and anticipating disease trajectory constitutes a key area of interest within GBM literature.[7] While survival statistics at the population level are well-established,[1,2] and factors impacting

Importance of the Study

This study demonstrates the potential of machine learning models to transform glioblastoma prognostication from population statistics to personalized predictions tailored to individual patients. Generating precise, patient-specific survival estimates is challenging due to glioblastoma's complexity and heterogeneity. This work introduces machine learning models leveraging a large dataset to forecast mortality at multiple time points postdiagnosis. The top-performing models are presented in an accessible web application that provides individualized prognostic calculations. By elucidating data-driven, customized prognoses, this study establishes a framework for enhancing personalized care in glioblastoma. Precision prognostication can potentially inform patient counseling based on an individual's risk profile rather than subjective judgments or generic averages. Overall, this work highlights the potential for machine learning to advance personalized medicine in neuro-oncology.

survival like age, molecular markers, the extent of resection, and tumor location have been identified,[8–13] generating precise survival predictions at the individual patient level remains an unmet clinical need.[7]

Existing efforts to generate time-dependent prognostic estimates for GBM patients' have utilized various approaches, including Cox proportional hazards regression-based nomograms and, more recently, machine learning (ML) approaches.[7] Compared to conventional statistical techniques like Cox regression, ML offers several advantages for survival prediction in GBM. First, ML methods can process extensive, multifaceted, and heterogeneous datasets, uncovering subtle patterns and relationships potentially overlooked by traditional statistical methods.[14,15] Second, ML facilitates integrating diverse variables spanning clinical, genomic, and imaging parameters, enabling a more comprehensive, personalized prognosis forecasting.[16,17] Finally, ML eliminates the need for rigid assumptions mandated by classical models, enabling the identification of nonlinear interactions and associations within high-dimensional data.[14,15] Collectively, these capabilities make ML a potentially more robust methodology compared to conventional statistics for generating individualized survival predictions in GBM patients.

In this study, we aim to develop accurate prediction models for GBM using ML approaches. While a few prior studies have also utilized ML for this purpose, their implications have been limited to demonstrating feasibility due to the lack of practical frameworks for clinicians to derive patient-specific predictions, with a few exceptions.[7,18,19] In our study, we intend to overcome some of the limitations of the majority of existing studies that have prevented the adoption of developed ML in real-world settings by exploiting the capabilities of ML to create a web application designed to provide interpretable survival predictions for patients diagnosed with IDH-wildtype GBM.

Materials and Methods

Ethical Approval

This study was deemed exempt by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai. The National Cancer Database (NCDB) Participant User Files contain national deidentified data for which obtaining consent was not applicable.

Data Source

The data used in this study were obtained from the 2020 iteration of the NCDB. The NCDB constitutes an extensive, prospectively maintained repository jointly developed by the Commission on Cancer (CoC) of the American College of Surgeons (ACS) and the American Cancer Society.[20]

Guidelines

Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis (TRIPOD)[21] and Journal of Medical Internet Research (JMIR) Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research[22] were adhered to.

Study Population

Adult patients aged 18 years and older, with histologically confirmed IDH-wildtype GBMs [ICD-O-3 (International Classification of Diseases for Oncology, third edition) morphology code 9440/3], were identified in the NCDB-Brain Participant User File (PUF) using Brain Molecular Markers data item code 05 [Glioblastoma, IDH-wildtype (9440/3)].[2] We restricted our study population to only patients with IDH-wildtype GBM and did not include grade 4 IDH-mutant astrocytoma as our exploratory analysis yielded only 100 patients with grade 4 IDH-mutant astrocytoma. The study population was limited to diagnoses made between January 1, 2018 and December 31, 2019. This limitation was due to the NCDB reporting new Brain Molecular Marker and MGMT promoter methylation data items for diagnoses made after January 1, 2018.[2] Additionally, the 2020 iteration of the PUF, which was the latest iteration obtainable from the ACS at the time the study was conducted, contained data for diagnoses made up to December 31, 2019. Patients with missing data in the "Vital Status" and "Last Contact or Death" data items were excluded.

Neuro-Oncology Advances

Predictor Variables and Outcomes of Interest

The predictor variables utilized by the ML models comprised: (1) demographics: age, sex, race, Hispanic ethnicity, insurance status, and Charlson–Deyo score (modified version of the Charlson Comorbidity Index that considers both the existence of comorbidities a patient has as well as the severity of those comorbidities)[23]; (2) facility characteristics: facility type (grouped into academic/research program, community cancer program, and integrated network program) and facility geographic location (grouped into 5 geographic regions in the United States: Central, Atlantic, Pacific, New England, and Mountain; Supplementary Figure 1); (3) diagnostic information: maximum tumor dimension and MGMT promoter methylation status; 4) treatment-related characteristics: the extent of resection, radiotherapy, chemotherapy, and immunotherapy.

The ML models were trained to predict survival outcomes at 6-month intervals postdiagnosis at the following time points: 6, 12, 18, and 24 months. The data items "Vital Status" and "Last Contact or Death" were combined to derive these binary outcomes of interest. For example, to identify patients who died within 6 months of their diagnosis, we searched for cases where the "Vital Status" data item was reported as "Dead" and the "Last Contact or Death" data item was reported as less than 6 months. These patients were assigned a 6-month mortality status of "Yes." Patients with a "Vital Status" listed as "Alive" and a "Last Contact or Death" beyond 6 months, or "Vital Status" listed as "Dead" but "Last Contact or Death" was more than 6 months, were given a 6-month mortality status of "No." In cases where a patient was marked as alive, but their latest follow-up data was recorded before the survival time point in question, they were excluded from the relevant survival analyses. This approach was replicated for survival outcomes at 12, 18, and 24 months.

Data Preprocessing

A custom dataset was created by filtering the NCDB-Brain PUF for the study population and extracting the data items used to generate the predictor variables and outcomes of interest, as outlined above. The code for the process regarding how these variables were derived from the NCDB data elements is shared on GitHub (https://github.com/mertkarabacak/NCDB-GBM/blob/main/Label_Renaming.ipynb). After constructing this custom dataset, categorical variables were label encoded. To handle missing data and prevent potential bias from excluding patients with incomplete records, we utilized the k-nearest neighbor (kNN) imputation algorithm, which predicts and imputes missing values based on the most similar complete data points.[24] Since kNN imputation operates on continuous data, imputed values for the label-encoded categorical variables were rounded to the nearest integer. The code for these later processes, label encoding and missing data imputation is shared on GitHub (https://github.com/mertkarabacak/NCDB-GBM/blob/main/Preprocessing.ipynb) as well.

Model Development and Evaluation

The datasets curated for each outcome of interest were divided into 3 subsets using a 60:20:20 distribution for training, validation, and test sets, respectively. The training sets were used to train the ML models, the validation sets were utilized for hyperparameter tuning and model calibration, and the test sets were used to evaluate the models' performance.

Before the initiation of model training, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training sets.[25] SMOTE addresses imbalances in class distribution within a dataset by generating synthetic samples from the minority class, thereby augmenting the number of instances in the underrepresented class instead of merely replicating existing samples. This technique effectively enlarges the sample size of the underrepresented class, which in turn potentially enhances the performance of the ML models.

Five supervised ML algorithms were used to build models: TabPFN, TabNet, XGBoost, LightGBM, and Random Forest. Supervised ML models are trained on labeled datasets with known outcomes. These models can then assimilate the patterns in the training data, allowing them to make accurate predictions on new, unseen data. Each of these algorithms was selected for its unique capabilities that have demonstrated high performance in differentiating or categorizing data, handling many variables concurrently, and flexibility for tuning. TabPFN represents a class of transformer-based models adept at discerning intricate patterns in point cloud data, which comprises points organized in a spatial layout.[26] TabNet is a deep learning architecture providing an interpretable framework suited for diverse structured data.[27] XGBoost and LightGBM are gradient-boosting frameworks renowned for their remarkable performance on classification tasks.[28,29] The Random Forest algorithm operates by constructing an ensemble of decision trees, with the final prediction synthesized from the aggregate consensus across the trees.[30]

Models built with these ML algorithms for each of the survival outcomes underwent hyperparameter optimization using the Optuna library, with the goal of maximizing the area under the receiver operating characteristics curve (AUROC) as the optimization metric.[31] Optuna is a versatile Python library designed to automate and streamline the optimization of ML model hyperparameters through a robust and flexible framework. To establish a baseline for the optimization procedure, the Tree-Structured Parzen Estimator Sampler algorithm was employed to generate estimates of AUROC. The models fitted with training data and optimized hyperparameters were then calibrated using a nonparametric approach called isotonic calibration, which adjusts the predicted probabilities to better match the distribution of observed labels, with the *CalibratedClassifierCV* function from the *scikit-learn* library.[32] Isotonic calibration aims to find the optimal monotonic transformation of the predicted probabilities to yield well-calibrated probability estimates.[33,34] The code for the model development process is shared on GitHub (https://github.com/mertkarabacak/NCDB-GBM/blob/main/Survival_Modeling.ipynb).

The performance of the calibrated models was evaluated both visually and quantitatively. The visual appraisal was conducted using 2 graphs: the receiver operating characteristics (ROC) and precision-recall curves (PRC). ROC curves visually showcase the diagnostic ability of binary classifiers across various discrimination thresholds by graphically depicting the true positive rate plotted against the false positive rate. PRCs, on the other hand, graphically depict precision and recall.

An important step in our evaluation was determining the optimal binary classification threshold based on predicted probabilities. For this purpose, we identified the optimal threshold as the point on the ROC curve corresponding to the maximum value of the Youden Index ($J = sensitivity + specificity - 1$), which is a common metric in diagnostic or prognostic test evaluation.[35,36] This method facilitated a balanced trade-off between sensitivity and specificity, potentially optimizing the performance of our models.

Upon establishing the optimal threshold using the Youden Index, we calculated the binary predictions based on predicted probabilities and proceeded with the quantitative appraisal of the models' performance, which encompassed metrics such as sensitivity, specificity, accuracy, area under the PRC (AUPRC), and AUROC. Additionally, we assessed the calibration of our models employing the Brier score, representing the average squared difference between predicted and actual probabilities.[34,37] A well-calibrated model will exhibit a Brier score close to zero, indicating no difference between the predicted and actual probabilities. Confusion matrices were also generated to interpret the performance of the models by providing a clear snapshot of correct and incorrect predictions.

Radar charts were plotted to facilitate a comparative assessment of models' performance across diverse metrics for each 1 of the 4 survival outcomes. These charts act as a tool for visualizing multidimensional data, where each of the 5 axes represents a distinct performance metric. The position on each corresponding axis denotes the model's performance in association with a single metric. To enhance the interpretability of our models, we ascertained the relative importance of predictor variables utilizing SHapley Additive exPlanations (SHAP).[38] SHAP is a method that assigns each feature an importance value for a particular prediction. The SHAP values indicate how much each feature contributed, either positively or negatively, to the model's output. SHAP bar plots were generated, delineating the cumulative impact that discrete features exert on the prognostications associated with each survival outcome. The length of each bar is a visual representation of the mean SHAP value for that feature, thereby denoting the impact of that feature's influence on the predicted outcome. Features are displayed in a hierarchical manner, with the most important positioned at the top. Additionally, partial dependence plots (PDPs) were employed to delineate the influence of discrete variables on the predictions.[39] PDPs illustrate the isolated effect of a single feature on the model's predicted output, elucidating the degree to which individual features sway the predictions.

Web Application

For each of the 4 survival outcomes, we selected 1 "top-performing" model for deployment to an open-access web application. These top models were selected based on their AUROC values, a recognized performance metric for ML models that are particularly relevant for binary classification tasks.[40] The AUROC encapsulates a model's ability to distinguish between positive and negative examples across the range of classification thresholds. Using AUROC as the primary criteria for selecting the top models is justified for 3 main reasons. First, it is unaffected by class imbalance, making it suitable for datasets with skewed class distributions. Second, it incorporates all possible classification thresholds, thereby providing a comprehensive evaluation of performance at different thresholds. Finally, it facilitates the comparative analysis of different models or algorithms by condensing model performance into a single value.

Upon selecting the top models for deployment, we developed an open-access web application through Hugging Face, a platform that enables sharing ML models via web interfaces and provides access to their source code. Our web application allows users to obtain personalized probabilistic survival predictions for IDH-wildtype GBM patients at 6-month intervals postdiagnosis up to 24 months. Users can input demographic, clinical, and treatment information and receive prognostic estimates at 6, 12, 18, and 24 months postdiagnosis. This easy-to-use tool provides clinicians and researchers a means to leverage our ML models to get survival predictions for individual patients based on their characteristics. The user interface and instructions for use are demonstrated in Supplementary Video 1. This web application and its source code are available at https://huggingface.co/spaces/MSHS-Neurosurgery-Research/NCDB-GBM.

Descriptive Statistics

For continuous variables with normal distribution, means (±standard deviations); for continuous variables without normal distribution, medians (interquartile ranges); and for categorical variables, proportions (%) were reported.

Results

A total of 7537 IDH-wildtype GBM patients were retrieved from the NCDB. The mean age was 65 (±15), and 3050 (40.5 %) of these patients were females. For the 6-month survival analysis, 7293 patients (25.5% with 6-month mortality); for the 12-month survival analysis, 7110 patients (46.6% with 12-month mortality); for the 18-month, 6868 patients (65% with 18-month mortality); and for the 24-month survival analysis, 6422 patients (78.6% with 24-month mortality) were included. The characteristics of the patient population before the application of the time point-specific exclusion criteria are presented in Table 1.

Figure 1 presents radar charts, each corresponding to 1 of the 4 survival outcomes under investigation, and Table 2 presents the performance metrics of the models built with the 5 ML algorithms for each survival

Neuro-Oncology Advances

Table 1. Patient Characteristics

Variables		Mean (±SD), Median (IQR), or *n* (%)
Age		65.0 (15.0)
Sex	Male	4487 (59.5%)
	Female	3050 (40.5%)
Race	White	6781 (90.0%)
	Black	446 (5.9%)
	Other	310 (4.1%)
Hispanic ethnicity	No	7080 (93.9%)
	Yes	457 (6.1%)
Insurance status	Medicare	3612 (47.9%)
	Private insurance	3028 (40.2%)
	Medicaid	557 (7.4%)
	Other government	161 (2.1%)
	Not insured	179 (2.4%)
Facility type	Academic/Research Program	3644 (48.4%)
	Community Cancer Program	2422 (32.1%)
	Integrated Network Cancer Program	1471 (19.5%)
Facility location	Central	2902 (38.5%)
	Atlantic	2748 (36.5%)
	Pacific	1011 (13.4%)
	New England	509 (6.8%)
	Mountain	367 (4.9%)
Percent no high school education quartiles	<6.3%	2108 (28.0%)
	6.3–10.8%	2648 (35.1%)
	10.9–17.6%	1764 (23.4%)
	>17.6%	1017 (13.5%)
Census median income quartiles	<$40 227	884 (11.7%)
	$40 227–$50 353	1440 (19.1%)
	$50 354–$63 333	2275 (30.2%)
	>$63 333	2938 (39.0%)
Rurality	Metro	6471 (85.9%)
	Urban	953 (12.6%)
	Rural	113 (1.5%)
Charlson-deyo score	0	6581 (87.3%)
	1	570 (7.6%)
	>2	386 (5.1%)
Tumor size		44.0 (21.0)
MGMT methylation	Unmethylated	4550 (60.4%)
	Methylated	2987 (39.6%)
Extent of resection	No resective surgery was performed	2122 (28.2%)
	Subtotal resection	2327 (30.9%)
	Gross total resection	3088 (41.0%)
Radiotherapy	No	1656 (22.0%)
	Yes	5881 (78.0%)
Chemotherapy	No	1850 (24.6%)
	Yes	5687 (75.4%)
Immunotherapy	No	7025 (93.2%)
	Yes	512 (6.8%)

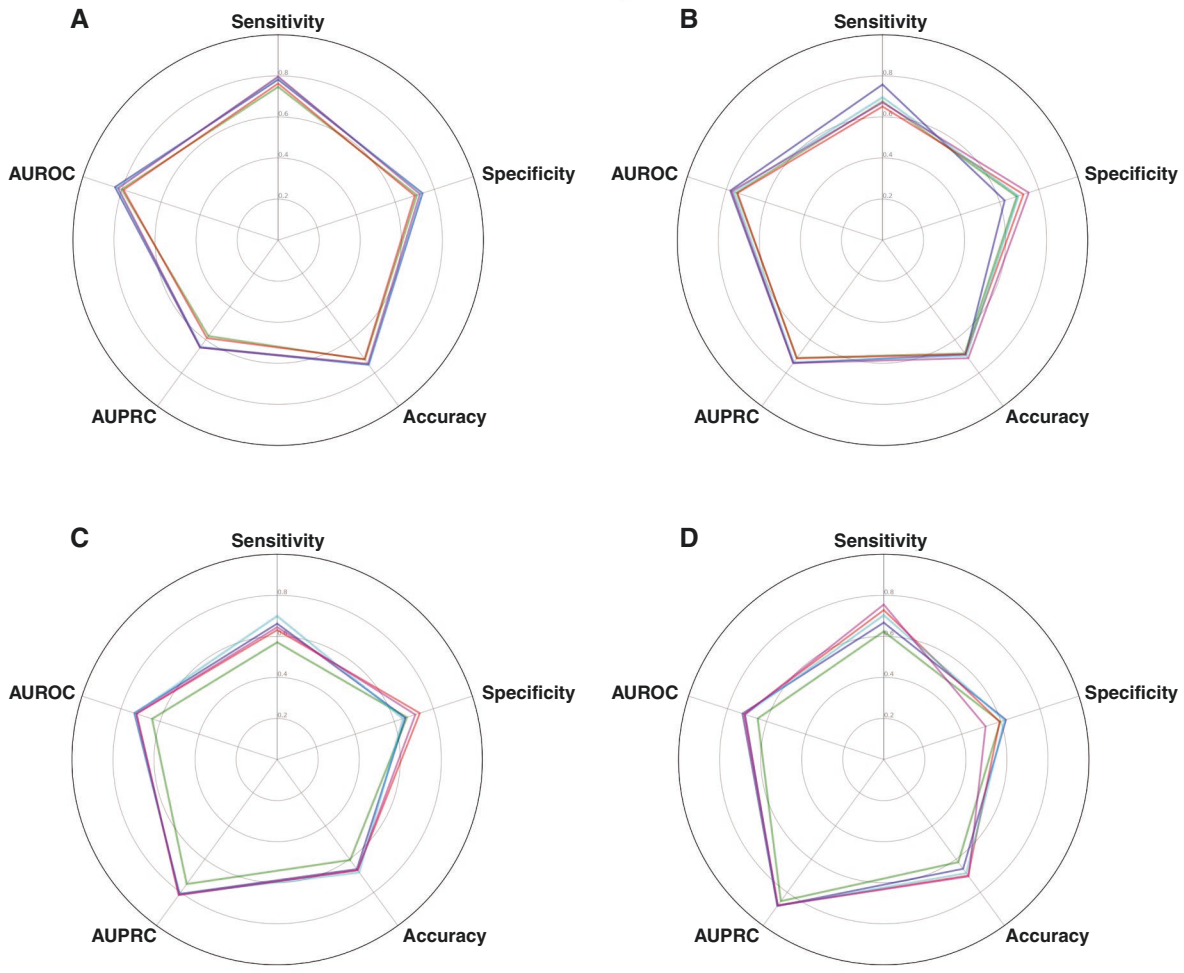

Figure 1. Algorithms' radar plots for the outcomes: (A) 6-month, (B) 12-month, (C) 18-month, and (D) 24-month mortality (AUROC, area under the receiver operating characteristics curve; AUPRC, area under the precision-recall curve).

outcome. Supplementary Figure 2 shows confusion matrices for the top-performing models per outcome, while Supplementary Figures 3–6 display confusion matrices for the remaining models. Performance evaluation revealed the top-performing models for each outcome were built using the TabPFN algorithm. The TabPFN models yielded mean AUROCs of 0.836, 0.78, 0.732, and 0.724 in predicting 6, 12, 18, and 24 month mortality, respectively. These results demonstrate good discriminatory ability in distinguishing patients who had 6-month mortality and fair discriminatory ability in distinguishing patients who had 12, 18, and 24 month mortality from those who did not.[41]

The ROC curves (Figure 2) illustrate the trade-off between sensitivity and specificity across probability cutoffs. As described in the methods, we determined the optimal classification threshold for each model using the Youden Index to find the optimal balance of sensitivity and specificity. The optimum thresholds were 15.07% for 6-, 45.62% for 12-, 63.18% for 18-, and 76.83% for 24-month mortality for top-performing models. Using these cutoffs to binarize the predicted probabilities, these models showed good discrimination for 6-month mortality and fair discrimination for longer-term mortality.

The PRCs (Figure 3) show the trade-off between precision and recall for different probability cutoffs. Precision refers to the proportion of positive predictions that are correct, while recall refers to the proportion of actual positives that are correctly predicted. The AUPRC values reflect the ability of the models to minimize false positives. The mean AUPRC values for the top-performing models were 0.647 for 6, 0.74 for 12, 0.809 for 18, and 0.882 for 24 month mortality. These results show the models' capacity to distinguish true positives improves over longer time horizons.

Figure 4 displays SHAP bar plots for the top models per outcome, while Supplementary Figures 7–10 show SHAP plots for the other models. For the 6-month top model, the variables with the highest absolute SHAP values were chemotherapy, radiotherapy, extent of resection, age, and MGMT methylation status—indicating they had the biggest influence on prediction. Similar interpretations can be

Neuro-Oncology Advances

Table 2. Performance Metrics of the Models

	Algorithm	Optimum classification threshold	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	Brier score (95% CI)
6-Month mortality	**TabPFN**	**15.07%**	**0.782 (0.761–0.803)**	**0.74 (0.717–0.763)**	**0.751 (0.729–0.773)**	**0.647 (0.622–0.672)**	**0.836 (0.805–0.853)**	**0.135 (0.117–0.153)**
	TabNet	22.22%	0.747 (0.725–0.769)	0.711 (0.688–0.734)	0.72 (0.697–0.743)	0.578 (0.553–0.603)	0.802 (0.773–0.824)	0.147 (0.129–0.165)
	XGBoost	30.43%	0.763 (0.741–0.785)	0.7 (0.676–0.724)	0.716 (0.693–0.739)	0.59 (0.565–0.615)	0.792 (0.773–0.825)	0.147 (0.129–0.165)
	LightGBM	23.08%	0.792 (0.771–0.813)	0.736 (0.713–0.759)	0.751 (0.729–0.773)	0.642 (0.617–0.667)	0.827 (0.802–0.851)	0.136 (0.118–0.154)
	Random Forest	20.00%	0.797 (0.776–0.818)	0.726 (0.703–0.749)	0.744 (0.722–0.766)	0.646 (0.621–0.671)	0.819 (0.801–0.842)	0.135 (0.117–0.153)
12-month mortality	**TabPFN**	**45.62%**	**0.758 (0.736–0.78)**	**0.626 (0.601–0.651)**	**0.689 (0.665–0.713)**	**0.74 (0.717–0.763)**	**0.78 (0.748–0.795)**	**0.196 (0.175–0.217)**
	TabNet	51.88%	0.676 (0.652–0.7)	0.688 (0.664–0.712)	0.682 (0.658–0.706)	0.709 (0.685–0.733)	0.748 (0.721–0.771)	0.206 (0.185–0.227)
	XGBoost	46.15%	0.651 (0.626–0.676)	0.722 (0.699–0.745)	0.688 (0.664–0.712)	0.712 (0.688–0.736)	0.743 (0.715–0.765)	0.209 (0.188–0.23)
	LightGBM	44.77%	0.696 (0.672–0.72)	0.696 (0.672–0.72)	0.696 (0.672–0.72)	0.735 (0.712–0.758)	0.762 (0.749–0.797)	0.195 (0.174–0.216)
	Random Forest	50.99%	0.67 (0.646–0.694)	0.749 (0.726–0.772)	0.711 (0.687–0.735)	0.74 (0.717–0.763)	0.772 (0.741–0.796)	0.194 (0.173–0.215)
18-Month mortality	**TabPFN**	**63.18%**	**0.662 (0.637–0.687)**	**0.657 (0.632–0.682)**	**0.66 (0.635–0.685)**	**0.809 (0.788–0.83)**	**0.732 (0.695–0.75)**	**0.2 (0.179–0.221)**
	TabNet	65.31%	0.571 (0.545–0.597)	0.665 (0.64–0.69)	0.605 (0.579–0.631)	0.75 (0.727–0.773)	0.641 (0.624–0.684)	0.215 (0.193–0.237)
	XGBoost	73.73%	0.631 (0.605–0.657)	0.731 (0.708–0.754)	0.667 (0.642–0.692)	0.817 (0.797–0.837)	0.721 (0.691–0.742)	0.197 (0.176–0.218)
	LightGBM	70.00%	0.699 (0.675–0.723)	0.645 (0.62–0.67)	0.68 (0.655–0.705)	0.81 (0.789–0.831)	0.729 (0.697–0.752)	0.198 (0.177–0.219)
	Random Forest	68.63%	0.645 (0.62–0.67)	0.708 (0.684–0.732)	0.667 (0.642–0.692)	0.814 (0.793–0.835)	0.721 (0.704–0.756)	0.198 (0.177–0.219)
24-Month mortality	**TabPFN**	**76.83%**	**0.667 (0.641–0.693)**	**0.626 (0.6–0.652)**	**0.658 (0.632–0.684)**	**0.882 (0.864–0.9)**	**0.724 (0.683–0.749)**	**0.154 (0.134–0.174)**
	TabNet	81.48%	0.624 (0.598–0.65)	0.597 (0.57–0.624)	0.618 (0.591–0.645)	0.853 (0.834–0.872)	0.646 (0.605–0.677)	0.165 (0.145–0.185)
	XGBoost	78.57%	0.728 (0.704–0.752)	0.594 (0.567–0.621)	0.699 (0.674–0.724)	0.877 (0.859–0.895)	0.715 (0.667–0.733)	0.156 (0.136–0.176)
	LightGBM	73.33%	0.703 (0.678–0.728)	0.619 (0.592–0.646)	0.685 (0.66–0.71)	0.882 (0.864–0.9)	0.71 (0.672–0.736)	0.15 (0.13–0.17)
	Random Forest	80.34%	0.755 (0.731–0.779)	0.522 (0.495–0.549)	0.704 (0.679–0.729)	0.88 (0.862–0.898)	0.711 (0.671–0.737)	0.152 (0.132–0.172)

made for the other outcome models based on their SHAP plots. Across the top models, treatment variables, age, and MGMT status consistently emerged as the most important predictors. Supplementary Figures 11–14 present partial dependence plots showing the individual effects of the 9 variables with the highest SHAP values on the predictions from each top model.

Discussion

This study demonstrates the potential of ML models to improve prognostication for GBM patients by developing models that can predict survival outcomes at multiple time points postdiagnosis. A key novelty is the incorporation of these models into an accessible web application that provides healthcare professionals with a practical tool to obtain individualized survival predictions. This study enables patient-specific, data-driven risk assessments tailored to the individual, unlike conventional practice, which depends on qualitative judgments based on a physician's limited experience or generalized population estimates. Such subjective assessments often have restricted applicability beyond a provider's experience, while population averages may not fit an individual patient's specific prognosis. By leveraging robust ML models trained on thousands of patients, this study overcomes these limitations to generate

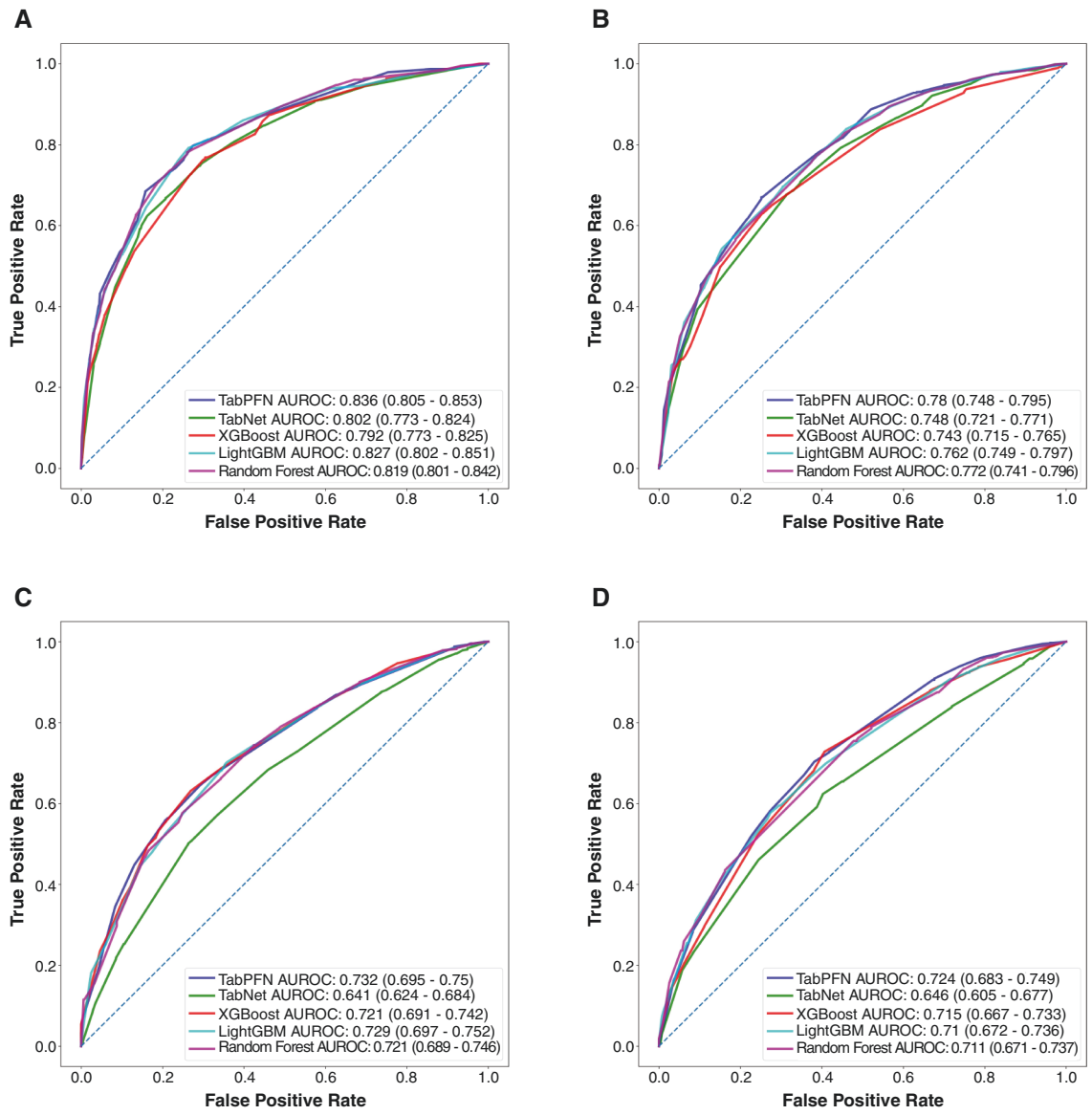

Figure 2. Algorithms' receiver operating characteristics curves for the outcomes: (A) 6-month, (B) 12-month, (C) 18-month, and (D) 24-month mortality (AUROC, area under the receiver operating characteristics curve).

tailored, quantitative predictions for each patient. Access to personalized prognosis estimates can empower clinicians to have informed discussions with patients about likely outcomes. This approach also has quality assurance applications, where worse-than-predicted outcomes could prompt a re-evaluation of protocols. Additionally, individualized prognosis assessment allows proper stratification of patients for research purposes and clinical trial design. Overall, this study exemplifies the value of ML in translating big data into precision, personalized prognostication.

In recent years, there has been a burgeoning interest in developing prognostic models to forecast survival outcomes for individual GBM patients.[7] These models employ a range of statistical and ML approaches to analyze multifaceted data and generate patient-specific survival estimates. However, according to a systematic review encompassing prognostic models constructed between 2010 and 2019, only 3 studies have operationalized their models into practical frameworks, such as online prediction tools, which are imperative for enhancing clinical utility and accessibility.[7] One such translational effort by Senders et al. utilized data from 20 821 patients who underwent resection for histopathologically confirmed GBM extracted from the Surveillance Epidemiology and End Results (SEER) database between 2005 and 2015.[19] Fifteen statistical and ML models were developed based on 13 demographic, socioeconomic, clinical, and radiographic

Neuro-Oncology Advances

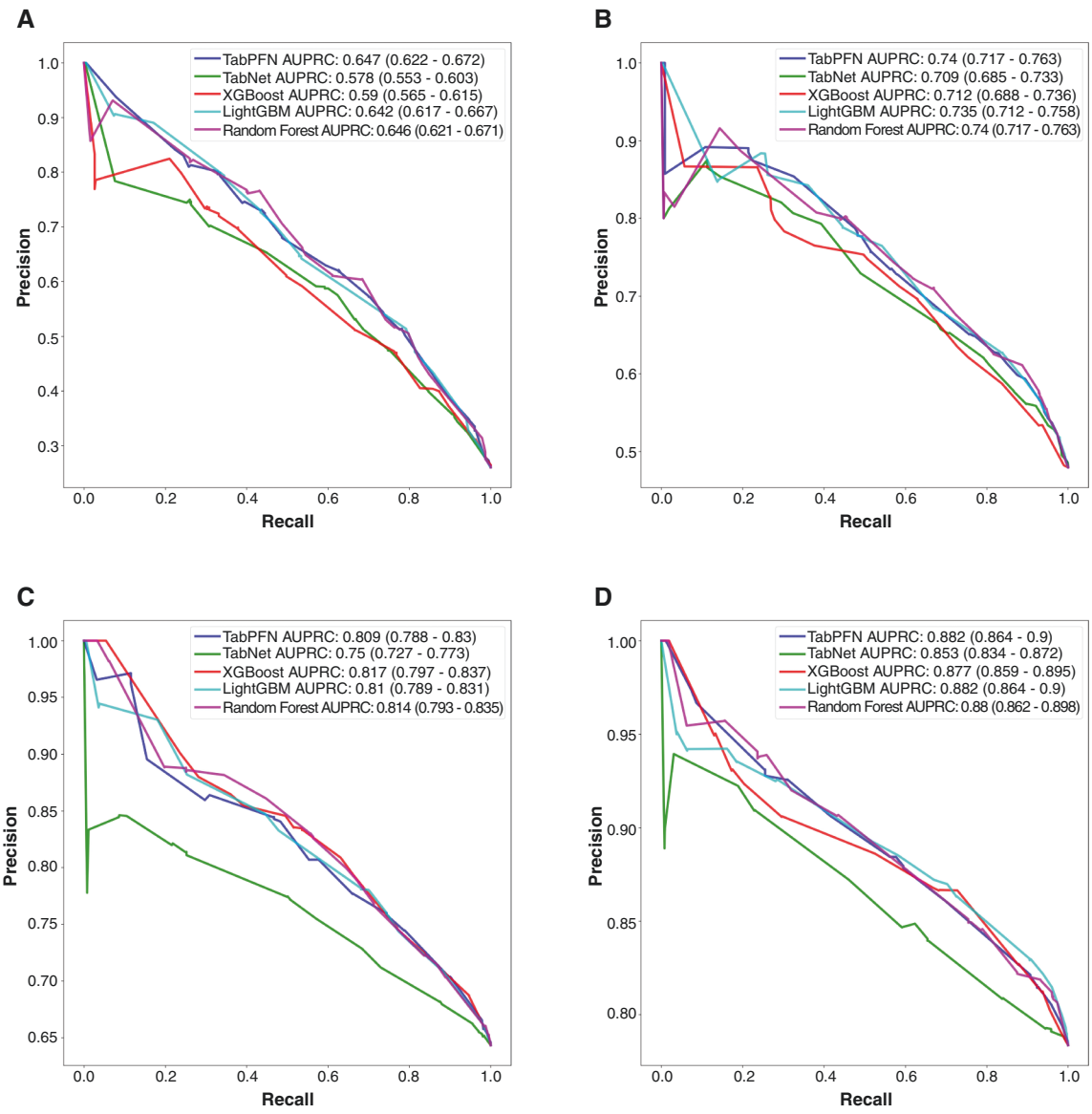

Figure 3. Algorithms' precision–recall curves for the outcomes: (A) 6-month, (B) 12-month, (C) 18-month, and (D) 24-month mortality (AUPRC, area under the precision-recall curve).

variables to predict overall survival, 1-year survival status, and generate individualized survival curves. Among these models, the accelerated failure time model demonstrated optimal discrimination with a concordance index of 0.70, outperforming Cox proportional hazards regression and other ML algorithms. Notably, the models developed in our current study, which utilized the TabPFN algorithm, achieved high discriminatory performance across all survival outcomes based on mean AUROC values ranging from 0.724 to 0.836. While the concordance index and AUROC are not directly equivalent metrics, they both provide an assessment of a model's discriminatory ability. The AUROC values achieved by our TabPFN models suggest comparable or potentially improved prognostic ability relative to Senders et al.'s top-performing model, though a direct numerical comparison is not straightforward given the different model evaluation metrics. Senders et al. deployed their top-performing model via a freely accessible web interface to facilitate clinical use. While laudable as a pioneering effort in operationalizing an ML-based prognostic model, several limitations temper the clinical applicability of this study. Notably, the online calculator generates survival predictions dependent on adjuvant radiotherapy, chemotherapy, or both, which may engender spurious assumptions of causality among users, as causal mechanisms were not explicitly analyzed by the authors. Additionally, while offering global model interpretations, their application lacks local explanations of individual-level

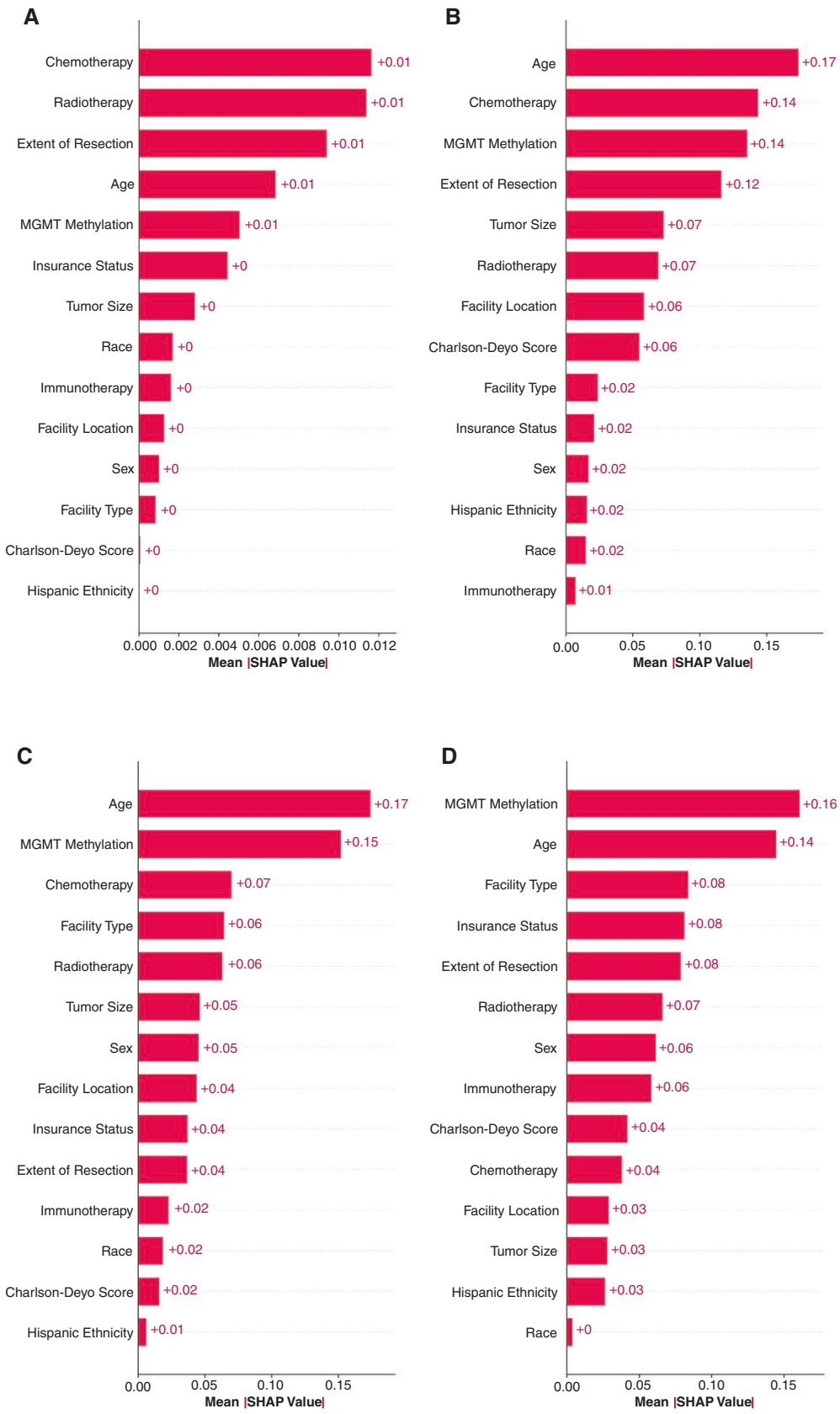

Figure 4. The 15 most important features and their mean SHAP values for the TabPFN models predicting: (A) 6-month, (B) 12-month, (C) 18-month, and (D) 24-month mortality (SHAP, SHapley Additive exPlanations).

predictions that contextualize real-world utility. The absence of model-agnostic methods like SHAP plots that provide granular, instance-wise interpretability poses a barrier to adoption in clinical practice, as it may promote the perception of ML models as inscrutable "black boxes."

Our study overcomes this latter limitation by integrating local explanatory techniques to build trust and transparency. The SHAP plots in our methodology furnish both global and local explanations. While global explanations elucidate overall model patterns, local explanations unpack the reasoning behind each patient-specific prediction. By revealing the rationale applied to each individual, the SHAP plots promote trust in model outputs rather than depicting them as black boxes. Moreover, they enable clinicians to critically evaluate alignment with their expertise. Through confirming or challenging predictions contextually, providers can integrate SHAP-enhanced ML prognostication into real-world settings. It should be mentioned here that global SHAP analysis revealed socioeconomic determinants of health, including insurance status, facility location, and facility type, emerged as significant components of our model. These factors are crucial in understanding the disparities in healthcare access and outcomes among glioblastoma patients. Acknowledging these socioeconomic predictors within our model not only enhances the accuracy of our survival predictions but also underscores the importance of addressing healthcare inequities in neuro-oncology. Future research should continue to explore these socioeconomic factors to develop more equitable healthcare strategies and improve prognostic tools for diverse patient populations.

In a more recent study, Kim et al. aimed to develop a clinically applicable prediction model predicting overall survival and progression-free survival in GBM patients treated with concurrent chemoradiotherapy.[18] Their analysis was limited by a small sample size of 467 patients from a single institution, compared to our more extensive dataset of over 7000 patients from the NCDB. They developed Cox proportional hazards, random survival forest, and survival support vector machine models based on 16 clinical variables. For both endpoints, the random survival forest model outperformed the other 2, yielding mean concordance indices of 0.72 and 0.70 for overall survival and progression-free survival, respectively. While not directly comparable metrics, as discussed previously when comparing our models to those by Senders et al., the AUROC values attained by our TabPFN models suggest comparable or potentially superior prognostic ability. Moreover, Kim et al. relied on 100-fold cross-validation for model assessment, rather than separate test sets like our study. Cross-validation risks overestimating performance when model parameters are optimized, as is often done. Additionally, their study lacked local interpretability methods to explain individual predictions, another limitation shared with prior work. Overall, our analysis addressed key restrictions of Kim et al.'s study regarding sample size, model transparency, and potentially better discrimination.

Similar to the comparable or potentially superior prognostic performance exhibited by our top models relative to prior studies, the performance evaluation within our study revealed a subtle yet consistent superiority of models built using the TabPFN algorithm over other ML algorithms across the distinct survival outcomes, as quantified by the AUROC metric. The advantage of TabPFN lies in its unique meta-learning framework, which facilitates learning from a variety of data, thereby enabling the algorithm to quickly adapt to new, unseen data.[26,42] This characteristic significantly bolsters its performance with the structured tabular data integral to this study. In contrast, a notable facet of TabPFN is its identity as a Prior-Data Fitted Network (PFN).[26] Unlike meta-learning, which is centered around enhancing the learning process, PFNs are pretrained on synthetic data to approximate Bayesian inference on new data.[43] Bayesian inference is a statistical method that allows for the quantification and management of uncertainty, a crucial feature in clinical prognosis scenarios.[44] The pretraining on synthetic data enables TabPFN to adeptly navigate complex patterns within real-world data, showcasing a nuanced approach to tabular data handling. This pretraining aspect also promotes a seamless transition to new datasets, enhancing its adaptability across diverse data scenarios. The algorithm's design minimizes the risk of overfitting and negates the need for extensive gradient-based training or hyperparameter tuning, aligning well with the primary objective of this study—to provide precise, individualized survival predictions for GBM patients. The subtle yet consistent superior performance of TabPFN accentuates the potential of employing meta-learned algorithms and prior-data fitted networks in enhancing prognostic performance, thus advancing the frontier of ML-driven, personalized prognostication in clinical oncology.

Beyond utilizing clinical, demographic, pathologic, or imaging data in a tabular format, several approaches leveraging modalities such as raw imaging and genomic data have been proposed to construct survival models for GBM patients. Pease et al. developed an MRI-based radiomic approach to discern patients with survival exceeding 12 months, leveraging preoperative MRI data from 235 individuals with pathologically confirmed GBMs in The Cancer Genome Atlas and an institutional cohort.[45] Ensuing manual segmentation of tumor volumes, radiomic features were extracted, with the 100 most relevant selected via the maximum relevance minimum redundancy technique. Prognostic models were constructed with a support vector machine classifier and validated through leave-one-out cross-validation and on external datasets. Both internal and external validations achieved AUROCs surpassing 0.91 and 0.71, respectively, for predicting 12-month survival. However, this study, like many studies following a similar methodology, did not sufficiently demonstrate seamless integration into clinical practice to enlighten prognosis. In contrast, Jia et al. aimed to construct and validate a radiomics nomogram for preoperative survival stratification in GBM patients, harnessing radiomic features from multiparametric MRI.[46] Following feature extraction and selection, classifier models were constructed, and a radiomics-based nomogram was created using logistic regression. While accessible for clinical utilization, this nomogram necessitates a "Radscore" input. Despite assertions it may facilitate preoperative planning and counseling, the means by which clinicians could procure this score remain ambiguous.

The existing body of literature offering predictive models based on genomic data is even more substantial,

with numerous studies proposing various gene signatures as prognostic biomarkers for survival outcomes in GBM patients.[47–52] While these gene signatures may provide valuable insights into the molecular basis of GBM, their clinical applicability is often limited.[53] This stems from the infrequent incorporation of the genomic profiling techniques used to develop these signatures into the standard of care for GBM management. It is more common to rely predominantly on traditional histopathological diagnosis rather than comprehensive genomic or transcriptomic analysis for clinical decision-making. Consequently, many proposed gene signature methodologies remain restricted to feasibility studies and are rarely adopted in real-world clinical practice. Though promising for elucidating the intricate genomic landscape of GBM, translation of these techniques into routine use has been modest, with barriers including cost, inadequate validation in heterogeneous populations, and lack of demonstrated superiority over conventional factors. Thus, despite extensive research on leveraging genomic data for enhanced prognostication, a sizable gap persists between theoretical modeling and clinical adoption. Further efforts are needed to demonstrate the feasibility and practical utility of modern "omics" technologies, in order to promote greater assimilation into neuro-oncology care.

Survival prediction plays a crucial role in clinical decision-making and enhancing patient counseling for those with GBM. Although the current prognosticator provides a user-friendly interface with potential clinical utility for approximating survival, it is designed as a research application and, at its current stage, should not be utilized in clinical settings to provide recommendations.[54] Validation of diverse cohorts, including single-institution and multicenter data, is vital to confirm the predictive capacity of this calculator for GBM patients. It is our aspiration that efforts to better delineate survival and prognostication of outcomes for GBM patients will not cease here—rather, we intend our model to serve as the first step in constructing more comprehensive calculators that incorporate other clinically relevant factors, such as biomarkers and imaging findings specific to this patient population.

This study has certain limitations, primarily stemming from the inherent biases of retrospective database analyses. As a registry database, the accuracy of coding and completeness of data capture are potential issues. Important clinical details like symptom presentation, surgical considerations, extent of resection determination, and adjuvant treatment specifics may be inconsistently reported or omitted entirely. As new treatments emerge to hopefully change the prognosis for this grim diagnosis, their incorporation into large databases may lag behind clinical practice. For example, tumor treating fields (TTF) therapy, a relatively new treatment modality, is not captured in the NCDB. Notably absent in the present analysis are data items for patients diagnosed in 2018–2019 that were available for prior years in NCDB, including Karnofsky Performance Scale scores, Ki-67 indexing, and tumor focality. Other unavailable clinical data like imaging findings, resection extent determination methods, and additional performance status factors potentially influencing the models were also not recorded by NCDB. Information is limited to initial management, with subsequent treatments unaccounted for, which may influence observed survival durations. The survival outcome was confined to overall survival, precluding analysis of progression-free survival or malignant transformation rates. Restricting the study to only IDH-wildtype GBM grade 4 means results may not extend to IDH-mutant astrocytoma grade 4, which would be considered a distinct entity in the WHO 2021 classification.[55] While the NCDB captures approximately 70% of US cancer cases, it may have selection bias towards larger, urban, Commission on Cancer accredited centers, and thus findings may not fully generalize. External validation is needed to evaluate model transportability and generalizability across different populations. While we performed internal validation by splitting the NCDB data into training and test sets, external validation in an entirely separate dataset is still needed. This would evaluate the model's transportability and generalizability in new populations. Access to an ideal external dataset that is significantly different from the NCDB was unfortunately unavailable for the current study. However, external validation in new datasets remains an important future direction. As a static analysis, dynamic updating of models over time as new data accrues was not undertaken, which may affect continued accuracy. Prospective validation in disparate datasets and modeling updates, as evidence emerges, are warranted to strengthen conclusions. Finally, the prognostic associations identified should not be interpreted as implying causal relationships between variables and survival outcomes. Proper experimental studies or causal inference methods are required to make causal claims, which is not possible within the scope of this retrospective database analysis. The models provide prognostic predictions only, and no causal relationships should be inferred.

Conclusions

This study demonstrates the potential of ML models to improve prognostication in IDH-wildtype GBM by developing an instrument for generating individualized survival predictions. Leveraging a large dataset and novel algorithms, we created an accessible web application that provides individualized prognostic estimates. Our top-performing models achieved satisfactory discriminatory performance, with mean AUROCs up to 0.836. This approach enables data-driven risk assessments tailored to each individual, overcoming the limitations of previous studies relying on subjective judgments or population averages. By elucidating prognosis at a granular level, this work exemplifies how ML can translate big data into precision medicine. Although retrospective database studies have intrinsic limitations, this study provides a framework for developing robust survival prediction models with modern ML techniques. Future efforts may focus on incorporating emerging data modalities, validating predictions prospectively, and updating models dynamically as new evidence accrues. Overall, our study highlights the potential for ML to transform cancer prognostication from the general population level to the individual patient.

Neuro-Oncology Advances

Supplementary material

Supplementary material is available at *Neuro-Oncology Advances* (https://academic.oup.com/noa).

Keywords

glioblastoma | machine learning | personalized medicine | predictive modeling | prognosis

Lay Summary

Glioblastoma (GBM) is a common form of brain cancer. Most patients with GBM eventually die from the disease, and it is challenging to predict how long each patient will live after being diagnosed. In this study, researchers used data from the National Cancer Database, including information about the patients, their tumors, and their treatments, to create a computer model that predicts individual survival times. Their results show that these models can estimate the chances of being alive at 6, 12, 18, and 24 months after surgery with reasonable, but not perfect, accuracy.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Acknowledgments

None.

Conflict of interest statement

The authors have no relevant financial or nonfinancial interests to disclose.

Authorship statement

Conceptualization: M.K., P.J., and K.M.; Methodology: M.K. and K.M.; Software: M.K.; Formal Analysis: M.K.; Data Curation: M.K.; Writing—Original Draft Preparation: M.K. and P.J.; Writing—Review & Editing: L.D., R.J.K., A.H.S., and K.M.; Visualization: M.K.; Supervision: K.M.; Project Administration: M.K. and K.M.

Ethics approval

No Institutional Review Board (IRB) approval or informed consent was required due to the use of de-identified patient data. The study was deemed exempt by the Icahn School of Medicine at Mount Sinai's IRB.

Source code

The source code for preprocessing and analyzing the data is available on GitHub (https://github.com/mertkarabacak/NCDB-GBM).

Statement for NCDB

The Commission on Cancer (CoC) of the American College of Surgeons and the American Cancer Society is the source of the data used herein; none of these institutions have verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

Data availability

Restrictions apply to the availability of these data. Data were obtained from the NCDB, a prospectively maintained repository collaboratively developed by the Commission on Cancer (CoC) of the American College of Surgeons and the American Cancer Society.

Affiliations

Department of Neurosurgery, Mount Sinai Health System, New York, New York, USA (M.K., K.M.); School of Medicine, SUNY Downstate Health Sciences University, New York, New York, USA (P.J.); Department of Neurological Surgery, University of Miami Miller School of Medicine, Miami, Florida, USA (L.D., A.H.S., R.J.K.); Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, Florida, USA (A.H.S., R.J.K.)

References

1. Ostrom QT, Price M, Neff C, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019. *Neuro-Oncology*. 2022;24(Supplement_5):v1–v95.
2. Ostrom QT, Shoaf ML, Cioffi G, et al. National-level overall survival patterns for molecularly-defined diffuse glioma types in the United States. *Neuro-Oncol*. 2023;25(4):799–807.

3. Wen PY, Kesari S. Malignant gliomas in adults. *N Engl J Med*. 2008;359(5):492–507.
4. Wick W, Osswald M, Wick A, Winkler F. Treatment of glioblastoma in adults. *Ther Adv Neurol Disord*. 2018;11:175628641879045.
5. Vanderbeek AM, Rahman R, Fell G, et al. The clinical trials landscape for glioblastoma: is it adequate to develop new treatments? *Neuro-Oncol*. 2018;20(8):1034–1043.
6. Stupp R, Taillibert S, Kanner A, et al. Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: a randomized clinical trial. *JAMA*. 2017;318(23):2306–2316.
7. Tewarie IA, Senders JT, Kremer S, et al. Survival prediction of glioblastoma patients—are we there yet? A systematic review of prognostic modeling for glioblastoma and its clinical potential. *Neurosurg Rev*. 2021;44(4):2047–2057.
8. Chen L, Ma J, Zou Z, et al. Clinical characteristics and prognosis of patients with glioblastoma: a review of survival analysis of 1674 patients based on SEER database. *Medicine (Baltimore)*. 2022;101(47):e32042.
9. Chandra A, Lopez-Rivera V, Dono A, et al. Comparative analysis of survival outcomes and prognostic factors of supratentorial versus cerebellar glioblastoma in the elderly: does location really matter? *World Neurosurg*. 2021;146:e755–e767.
10. Sheikh S, Radivoyevitch T, Barnholtz-Sloan JS, Vogelbaum M. Long-term trends in glioblastoma survival: implications for historical control groups in clinical trials. *Neurooncol Pract*.. 2020;7(2):158–163.
11. Ostrom QT, Rubin JB, Lathia JD, Berens ME, Barnholtz-Sloan JS. Females have the survival advantage in glioblastoma. *Neuro-Oncology*. 2018;20(4):576–577.
12. Adams H, Chaichana KL, Avendaño J, et al. Adult cerebellar glioblastoma: understanding survival and prognostic factors using a population-based database from 1973 to 2009. *World Neurosurg*. 2013;80(6):e237–e243.
13. Johnson DR, O'Neill BP. Glioblastoma survival in the United States before and during the temozolomide era. *J Neurooncol*. 2012;107(2):359–364.
14. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8–17.
15. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2007;2:59–77.
16. Khader F, Kather JN, Müller-Franzes G, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Sci Rep*. 2023;13(1):10666.
17. Luckett PH, Olufawo M, Lamichhane B, et al. Predicting survival in glioblastoma with multimodal neuroimaging and machine learning. *J Neurooncol*. 2023;164(2):309–320.
18. Kim Y, Kim KH, Park J, Yoon HI, Sung W. Prognosis prediction for glioblastoma multiforme patients using machine learning approaches: development of the clinically applicable model. *Radiother Oncol*. 2023;183:109617.
19. Senders JT, Staples P, Mehrtash A, et al. An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning. *Neurosurgery*. 2020;86(2):E184–E192.
20. Mallin K, Browner A, Palis B, et al. Incident cases captured in the national cancer database compared with those in U.S. population based central cancer registries in 2012–2014. *Ann Surg Oncol*. 2019;26(6):1604–1612.
21. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13(1):1.
22. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18(12):e323.
23. Ladha KS, Zhao K, Quraishi SA, et al. The deyo-charlson and elixhauser-van walraven comorbidity indices as predictors of mortality in critically ill patients. *BMJ Open*. 2015;5(9):e008990.
24. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak*. 2016;16(3):74.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.
26. Hollmann N, Müller S, Eggensperger K, Hutter F. TabPFN: a transformer that solves small tabular classification problems in a second [published online May 7]. *arXiv preprint*. 2023. doi:10.48550/arXiv.2207.01848
27. Arik SO, Pfister T. TabNet: attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence. 2021;35(8):6679–6687.
28. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA; 2016:785–794. doi:10.1145/2939672.2939785
29. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. Vol. 30. New York, NY: Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
30. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
31. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework [published online July 25]. 2019. doi:10.48550/arXiv.1907.10902
32. Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project [published online]. *arXiv preprint*. 2013. doi:10.48550/ARXIV.1309.0238
33. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif*. 1999;10(3):61–74.
34. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning—ICML'05. Bonn, Germany: ACM Press; 2005:625–632. doi:10.1145/1102351.1102430
35. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.
36. Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. *Biom J*. 2005;47(4):458–472.
37. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
38. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. Vol. 30. New York, NY: Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
39. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat*. 2015;24(1):44–65.
40. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145–1159.
41. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
42. Schaul T, Schmidhuber J. Metalearning. *Scholarpedia*. 2010;5(6):4650.
43. Müller S, Hollmann N, Arango SP, Grabocka J, Hutter F. Transformers can do Bayesian inference [published online]. *arXiv preprint*. 2021. doi:10.48550/ARXIV.2112.10510

44. Braithwaite D, Hedges J, Smithe TSC. The compositional structure of Bayesian inference [published online]. *arXiv preprint*. 2023. doi:10.48550/ARXIV.2305.06112
45. Pease M, Gersey ZC, Ak M, et al. Pre-operative MRI radiomics model non-invasively predicts key genomic markers and survival in glioblastoma patients. *J Neurooncol*. 2022;160(1):253–263.
46. Jia X, Zhai Y, Song D, et al. A multiparametric MRI-based radiomics nomogram for preoperative prediction of survival stratification in glioblastoma patients with standard treatment. *Front Oncol*. 2022;12:758622.
47. Wang S, Xu X. An immune-related gene pairs signature for predicting survival in glioblastoma. *Front Oncol*. 2021;11:564960.
48. Cheng L, Yuan M, Li S, et al. Identification of an IFN-β-associated gene signature for the prediction of overall survival among glioblastoma patients. *Ann Transl Med*. 2021;9(11):925.
49. Li XY, Zhang LY, Li XY, Yang XT, Su LX. A pyroptosis-related gene signature for predicting survival in glioblastoma. *Front Oncol*. 2021;11:697198.
50. Zeng WJ, Cao YF, Li H, et al. A novel thrombosis-related signature for predicting survival and drug compounds in glioblastoma. *J Oncol*. 2022;2022:6792850.
51. Yu Z, Du M, Lu L. A novel 16-genes signature scoring system as prognostic model to evaluate survival risk in patients with glioblastoma. *Biomedicines*. 2022;10(2):317.
52. Jin Y, Wang Z, Xiang K, et al. Comprehensive development and validation of gene signature for predicting survival in patients with glioblastoma. *Front Genet*. 2022;13:900911.
53. Koscielny S. Why most gene expression signatures of tumors have not been useful in the clinic. *Sci Transl Med*. 2010;2(14):14ps2–14ps2.
54. Kuo CC, Monteiro A, Lim J, et al. An online calculator using machine learning for predicting survival in pediatric patients with medulloblastoma. *J Neurosurg Pediatr*. 2023;33(1):85–94.
55. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-Oncology*. 2021;23(8):1231–1251.