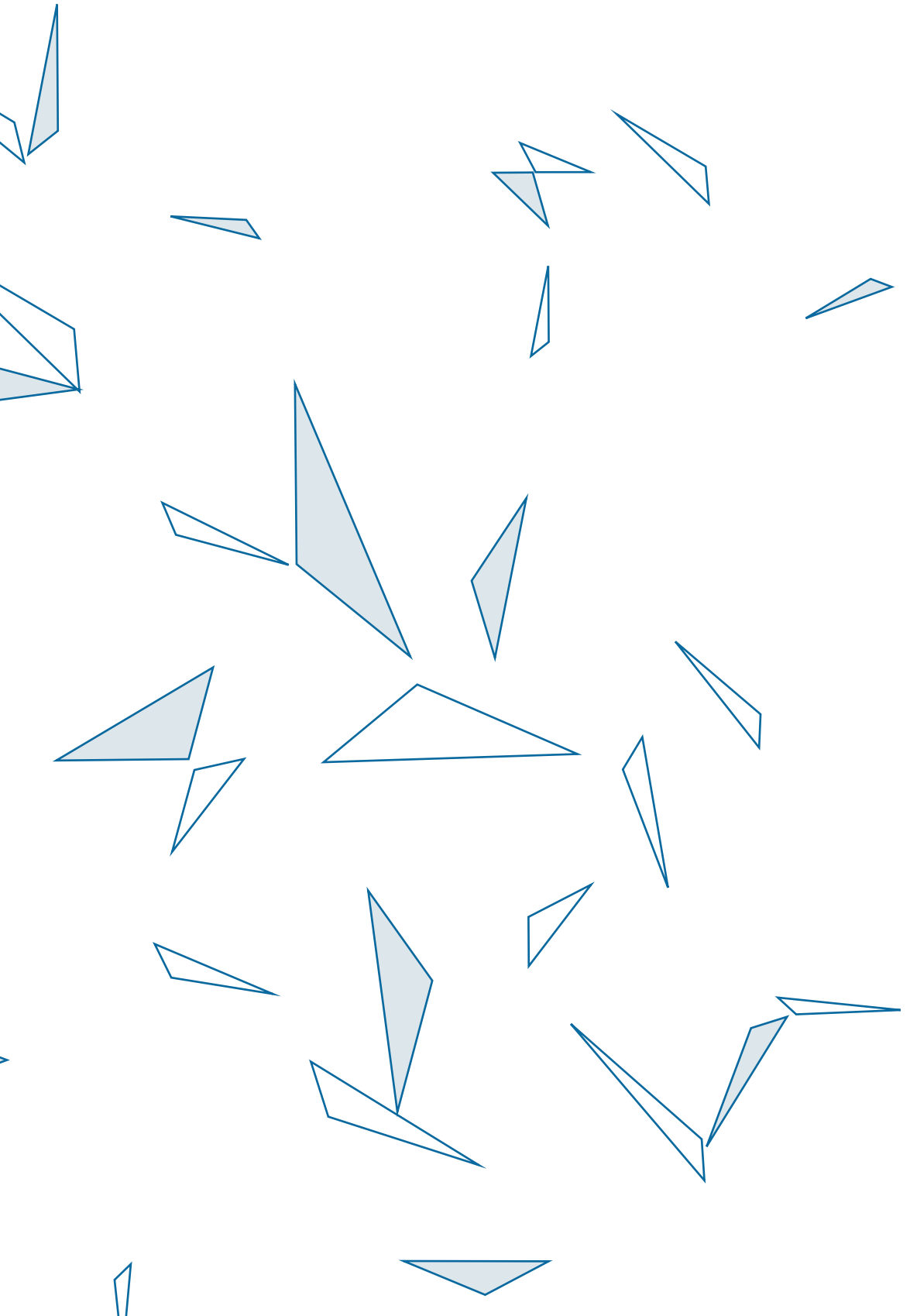
Applied machine learning in neurosurgical oncology

Senders, J.T.

PART II

Predictive analytics in neurosurgical oncology



5

An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning

Joeky T. Senders, Patrick Staples, Alireza Mehrtash, David J. Cote, Martin J.B. Taphoorn, David A. Reardon, William B. Gormley, Timothy R. Smith, Marike L. Broekman*, Omar Arnaout*

*These authors contributed equally and share last authorship

NEUROSURGERY 2020 FEB 1;86(2):E184-E192

Abstract

Introduction

Although survival statistics in patients with glioblastoma are well-defined at the group level, predicting individual-patient survival remains challenging due to significant variation within strata. The aim of this study was to compare statistical and machine learning algorithms in their ability to predict survival in glioblastoma patients and deploy the best performing model as an online survival calculator.

Methods

Patients undergoing an operation for a histopathologically confirmed glioblastoma were extracted from the Surveillance Epidemiology and End Results (SEER) database (2005-2015) and split into a training and hold-out test set in an 80/20 ratio. Fifteen statistical and machine learning algorithms were trained based on 13 demographic, socio-economic, clinical, and radiographic features to predict overall survival, one-year survival status, and compute personalized survival curves.

Results

In total, 20,821 patients met our inclusion criteria. The accelerated failure time model demonstrated superior performance in terms of discrimination (concordance-index=0.70), calibration, interpretability, predictive applicability, and computational efficiency compared to Cox proportional hazards regression and other machine learning algorithms. This model was deployed through a free, publicly available software interface (https://cnoc-bwh.shinyapps.io/gbmsurvivalpredictor/).

Conclusion

The development and deployment of survival prediction tools require a multimodal assessment rather than a single metric comparison. This study provides a framework for the development of prediction tools in cancer patients, as well as an online survival calculator for patients with glioblastoma. Future efforts should improve the interpretability, predictive applicability, and computational efficiency of existing machine learning algorithms, increase the granularity of population-based registries, and externally validate the proposed prediction tool.

Introduction

Glioblastoma is the most common primary malignant brain tumor with almost 12,000 new cases per year in the United States and a median survival of only a year after diagnosis.[1] Adequate survival prognostication is essential for informing clinical and personal decision-making. Although survival statistics are well-defined at the group-level, predicting individual patient survival remains challenging due to the heterogenous nature of the disease and significant variation in survival within strata.

In recent years, numerous statistical and machine learning algorithms have emerged that can learn from examples to make patient-level predictions of survival. These algorithms can be particularly useful for tailoring clinical care to the needs of the individual glioblastoma patient.

This study aims to compare the most commonly used statistical and machine learning algorithms in their ability to predict individual-patient survival in glioblastoma patients. In order to promote the reproducibility of the current study and facilitate external validation and implementation of the developed models, we deployed the best performing model as an online calculator that provides interactive, online, and graphical representations of personalized survival estimates.

Methods

Data and study population

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement was used for the reporting of this study.[2] Data was extracted from the Surveillance Epidemiology and End Results (SEER) database (2005-2015).[3] The SEER registry compiles cancer incidence and survival data of 18 registries and covers 28% of the U.S. population from academic and nonacademic hospitals, and as such, is broadly representative of the U.S. population as a whole.[4] Patients who underwent surgery for a histopathologically confirmed diagnosis of a glioblastoma (International Classification of Diseases for Oncology-Third Edition [ICD-O-3] codes 9440, 9441, 9442) were included in the analysis. Patients were excluded from the analysis if they died in the direct postoperative period (≤30 days after surgery). Our institutional review board has exempted the SEER database from review and waived the need for informed consent due to the retrospective nature of this study.

Outcome and input features

Although machine learning provides a variety of predictive algorithms, most of them are developed to accommodate binary or continuous outcomes instead of censored survival outcomes (i.e., time-to-event data). To facilitate a vis-à-vis comparison between traditional statistical and novel machine learning algorithms, we compared all algorithms in their ability to predict one or more of the following survival outcomes: (i) continuous: overall survival from diagnosis to death in months, (ii) binary: one-year survival probability, and (iii) censored: subject-level Kaplan-Meier survival curves. All demographic, socio-economic, radiographic, and therapeutic characteristics available at individual patient-level in the SEER registry were included as input features. Continuous variables included age at diagnosis (years) and maximal enhancing tumor diameter in any dimension (millimeters). Categorical variables included sex, race (White, Black, Asian, other), ethnicity (Hispanic, non-Hispanic), marital status (married, non-married), insurance status (insured, uninsured/Medicaid), tumor laterality (left, right, midline), tumor location (frontal, temporal, parietal, occipital, cerebellum, brainstem, ventricles, overlapping lesion), tumor extension (confined to primary location, ventricle involvement, midline crossing), surgery type (biopsy, sub-total resection, gross-total resection), and administration of any form of postoperative chemotherapy and/or radiotherapy. Data on input features and survival outcomes were collected by independent, trained data collectors.

Statistical analysis

Missing data was multiple imputed by means of a random forest algorithm.[5] The total cohort was randomly split into a training and hold-out test set based on an 80/20 ratio. The Cox proportional hazards regression (CPHR) and the Accelerated Failure Time (AFT) algorithms allow for inferential analysis on censored survival data. Therefore, both approaches were also utilized to provide insight into the independent association between covariates and survival. Interactions between age, sex, surgery type, radiotherapy, and chemotherapy were modeled in both approaches. The Benjamini-Hochberg procedure based on 41 comparisons (26 parameters plus 15 two-way interactions) was used to adjust for multiple testing. The proportional hazards assumption of the CPHR model was assessed by means of the Schoenfeld Residuals Test, and the distribution assumption of the AFT by means of a quantile-quantile plot. All covariates that were statistically significantly associated with survival in the inferential analysis were included in the predictive analysis.

For the predictive analysis, 15 machine learning and statistical algorithms were trained including AFT, bagged decision trees, boosted decision trees, boosted

decision trees survival, CPHR, extreme boosted decision trees, k-nearest neighbors, generalized linear models, lasso and elastic-net regularized generalized linear models, multilayer perceptron, naïve Bayes, random forests, random forest survival, recursive partitioning, and support vector machines.[6–8] Among these, only the AFT, boosted decision trees survival, CPHR, random forest survival, and recursive partitioning algorithms were capable of modeling time-to-event data. Five-fold cross-validation was used on the training set for preprocessing optimization and hyperparameter tuning. Hyperparameters were model-specific, such as the number of trees in a random forest model and the number of layers or nodes per layer in a neural network. The algorithms were subsequently trained with optimized hyperparameter settings on the full training set and evaluated on the hold-out test set, which has not been used for preprocessing and hyperparameter tuning in any form.

Metrics of predictive performance

Discrimination and calibration were used as metrics for prediction performance. Discrimination reflects the ability of a model to separate observations, whereas calibration measures the agreement between the observed and predicted outcomes.[9] Discrimination was quantified according to the concordance index (C-index). The C-index represents the probability that for any two patients chosen at random, the patient who had the event first is rated as being more at risk of the event according to the model. Therefore, the C-index takes into account the occurrence of the event, as well as the length of follow-up, and is particularly well-suited for right-censored survival analysis.[10] For the subject-level survival curves produced by time-to-event models, the C-index was evaluated per time point weighted according to the survival distribution in the test set and integrated over time. The relationship between predicted one-year survival probability and observed survival rate was graphically assessed in a calibration plot.

Secondary metrics

In addition to prediction performance, we evaluated additional metrics that pose significant pragmatic challenges to the deployment and implementation of prediction models in clinical care. These metrics include model interpretability, predictive applicability, and computational efficiency. Lack of interpretability is an important concern for the implementation of many machine learning models, which are typically referred to as “black-boxes” and sometimes cited as a weakness compared to classical statistical methods. Inferential utility is a traditional hallmark of model interpretability and therefore included as a model assessment measure. Predictive applicability refers to the type of outcome classes to be predicted (binary, continuous, or time-to-event), as

well as the generated output of the fitted models (class probability, numeric estimate, or subject-level survival curve, respectively). Computational efficiency was measured in terms of model size, loading time, and computation time to produce a prediction. For models that do not provide natural prediction confidence intervals, model predictions were bootstrapped 100 times with replacement to provide such estimates.

We also developed an online, interactive, and graphical tool based on the overall best performing model. Statistical analyses were conducted using R (version 3.5.1, R Core Team, Vienna, Austria).[11] All machine learning modeling was performed using the Caret package,[12] and the application was built and deployed using the Shiny package and server.[13]

Results

Patient demographics and clinical characteristics

In total, 20,821 patients met our inclusion criteria. Missing data was multiply imputed for insurance status (16.7% missingness), tumor size (14.3%), tumor laterality (12.0%), tumor location (6.6%), marital status (3.8%), tumor extension (1.6%), surgery type (1.3%), and race (0.2%). Survival time was censored for 3,745 patients (18.0%). The estimated median survival time in the total cohort was 13 months (95%-CI 12-13 months). The total cohort was split into a training and hold-out test set of 16,656 and 4,165 patients, respectively (Table 1).

TABLE 1. Baseline characteristics for the training and hold-out test set.

		Training set (n = 16,656)		Hold-out test set (n = 4,165)		
Characteristic	**Definition**	**n**	**%**	**n**	**%**	***p***
Age (years)	<50	2900	17.4	695	16.7	0.505
	50-70	9781	58.7	2456	59.0	
	>70	3975	23.9	1014	24.3	
	Mean ± SD	60.5 ± 13.8		60.7 ± 13.9		0.179
Sex	Female	6872	41.3	1717	41.2	0.982
	Male	9784	58.7	2448	58.8	
Race	White	14821	89.0	3710	89.1	0.509
	Black	1018	6.1	238	5.7	
	Asian	741	4.4	201	4.8	
	Other	76	0.5	16	0.4	
Hispanic	No	14993	90.0	3735	89.7	0.533
	Yes	1663	10.0	430	10.3	
Married	No	5535	33.2	1305	31.3	0.021
	Yes	11121	66.8	2860	68.7	
Insurance	Insured	14503	87.1	3636	87.3	0.717
	Uninsured/Medicaid	2153	12.9	529	12.7	
Laterality	Left	7779	46.7	1901	45.6	0.469
	Right	8714	52.3	2222	53.3	
	Midline	163	1.0	42	1.0	
Location	Frontal lobe	5001	30.0	1219	29.3	0.377
	Temporal lobe	4901	29.4	1270	30.5	
	Parietal lobe	3071	18.4	770	18.5	
	Occipital lobe	875	5.3	206	4.9	
	Ventricle, NOS	79	0.5	14	0.3	
	Cerebellum, NOS	125	0.8	39	0.9	
	Brain stem	75	0.5	12	0.3	
	Overlapping lesion of brain	2529	15.2	635	15.2	
Tumor extension	Confined to primary location	14007	84.1	3536	84.9	0.295
	Ventricles	653	3.9	144	3.5	
	Midline crossing	1996	12.0	485	11.6	
Tumor size (mm)	<25	1539	9.2	382	9.2	0.387
	25-50	9380	56.3	2393	57.5	
	>50	5737	34.4	1390	33.4	

TABLE 1. Continued

		Training set (n = 16,656)		Hold-out test set (n = 4,165)		
Characteristic	**Definition**	**n**	**%**	**n**	**%**	***p***
Tumor size	Median [IQR]	45 [35-55]	45 [35-55]	45 [35-55]	45 [35-55]	0.986
Surgery type	Biopsy	3882	23.3	977	23.5	0.894
	Sub-total resection	5888	35.4	1456	35.0	
	Gross-total resection	6886	41.3	1732	41.6	
Radiotherapy	No	2667	16.0	662	15.9	0.871
	Yes	13989	84.0	3503	84.1	
Chemotherapy	No	3647	21.9	883	21.2	0.341
	Yes	13009	78.1	3282	78.8	

Abbreviations: IQR=interquartile range; mm=millimeters; n=number; SD=standard deviation

Inferential analysis

The Schoenfeld residuals test demonstrated that the assumption of proportionality was violated for all variables except sex and ethnicity in the CPHR model (all p<.006 and global test p<.001; Supplementary Table S1). The quantile-quantile plot demonstrated a valid log-logistic distribution assumption for the (AFT) model (Supplementary Figure S1). For these reasons, we present the inferential results of the AFT model. The AFT allows for uncomplicated interpretation, as it provides acceleration factors (γ), which represent the relative survival duration of a strata compared to the reference group. For example, a γ of 1.5 reflects an expected survival duration that is 50% longer compared to the reference group. Multivariable AFT analysis identified older age (γ=0.75 per 10 years increase, p<.001), male sex (γ=0.93, p<.001), uninsured insurance status or insurance by Medicaid (γ=0.87, p<.001), midline tumors (γ=0.79, p=.004), tumors primarily located in the parietal lobe (γ=0.91, p<.001), brain stem (γ=0.44, p<.001), or multiple lobes (γ=0.88, p<.001), tumors extending to the ventricles (γ=0.90, p<.001) or across the midline (γ=0.73, p<.001), and larger sized tumors (γ=0.99 per cm, p<.001) as independent predictors of shorter survival (Figure 1). Asian race (γ=1.14, p=.001), Hispanic ethnicity (γ=1.08, p=.007), married marital status (γ=1.15, p<.001), gross-total resection (γ=1.19, p<.001), radiotherapy (γ=1.27, p<.001), and chemotherapy (γ=1.49, p<.001) were identified as independent predictors of longer survival.

The AFT model with interaction terms demonstrated that age interacted with extent of resection (γ>1.03 per 10 years increase, p<.02), as well as radiotherapy (γ=1.04 per 10 years increase, p=.03) (Supplementary Table S2).

Variable		N	Estimate	p
Age at diagnosis (y)		16656	0.97 (0.97, 0.97)	<0.001
Sex	Female	6872	Reference	
	Male	9784	0.93 (0.90, 0.96)	<0.001
Race	White	14821	Reference	
	Black	1018	1.05 (0.99, 1.12)	0.229
	Asian	741	1.14 (1.06, 1.23)	0.001
	Other	76	0.91 (0.73, 1.13)	0.778
Hispanic	No	14993	Reference	
	Yes	1663	1.08 (1.03, 1.14)	0.007
Married	No	5535	Reference	
	Yes	11121	1.15 (1.12, 1.19)	<0.001
Insurance	Insured	14503	Reference	
	Uninsured/Medicaid	2153	0.87 (0.83, 0.91)	<0.001
Laterality	Left	7779	Reference	
	Right	8714	1.00 (0.97, 1.03)	1.000
	Midline	163	0.79 (0.69, 0.91)	0.004
Location	Frontal lobe	5001	Reference	
	Temporal lobe	4901	0.98 (0.94, 1.02)	0.590
	Parietal lobe	3071	0.91 (0.87, 0.95)	<0.001
	Occipital lobe	875	0.97 (0.91, 1.04)	0.801
	Ventricle, NOS	79	0.97 (0.78, 1.20)	1.000
	Cerebellum, NOS	125	1.03 (0.87, 1.22)	1.000
	Brain stem	75	0.44 (0.35, 0.54)	<0.001
	Overlapping lesion of brain	2529	0.88 (0.84, 0.92)	<0.001
Tumor extension	Confined to Primary Location	14007	Reference	
	Ventricles	653	0.90 (0.83, 0.97)	0.013
	Midline Crossing	1996	0.73 (0.70, 0.76)	<0.001
Tumor size (cm)		16656	0.99 (0.99, 1.00)	0.002
Surgery type	Biopsy	3882	Reference	
	Sub–Total Resection	5888	0.96 (0.93, 1.00)	0.089
	Gross–Total Resection	6886	1.19 (1.15, 1.24)	<0.001
Radiotherapy	No	2667	Reference	
	Yes	13989	1.27 (1.21, 1.33)	<0.001
Chemotherapy	No	3647	Reference	
	Yes	13009	1.49 (1.42, 1.55)	<0.001

FIGURE 1. Forest plot for the accelerated failure time model characterizing the association between the individual predictors and survival. In the inferential analysis, the estimates for age and tumor size were presented per ten years and ten millimeters increase, respectively, to reflect the incremental relative survival duration of clinically meaningful intervals. The p-value was corrected for multiple testing by means of the Benjamini-Hochberg procedure.

Predictive analysis

The discriminatory performance on the hold-out test set as measured by the C-index set ranged between 0.66-0.70 and between 0.67-0.70 across all models for predicting overall survival and one-year survival status, respectively (Table 2). Among the time-to-event models, the integrated C-index ranged between 0.68-0.70 for predicting subject-level Kaplan-Meier survival curves. The AFT model based on a log-logistic distribution demonstrated the highest discriminatory performance for computing personalized survival curves. Compared to all continuous and binary models, the AFT model demonstrated similar or better discrimination for predicting overall survival and one-year survival probability, respectively. Model calibration varied significantly across all models (Supplementary Figure S2). The traditional CPHR model systematically underestimated survival in the 0.5-0.75 one-year survival probability range, whereas the AFT model showed better calibration, particularly in this clinically relevant interval (Figure 2).

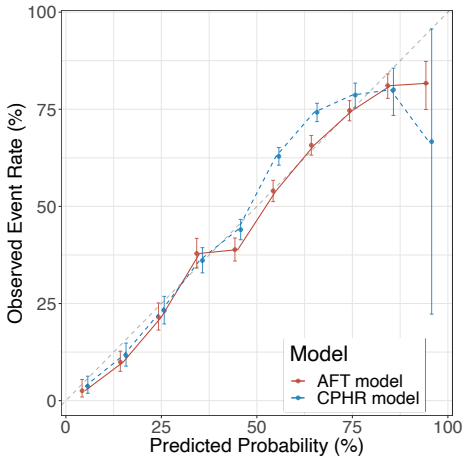


FIGURE 2. Calibration plot demonstrating a systematic underestimation of survival by the Cox proportional hazards regression model in the 0.5 to 0.75 one-year survival probability range and a well-calibrated accelerated failure time model. Abbreviations: AFT=accelerated failure time; CPHR=Cox proportional hazards regression.

TABLE 2. Discriminatory performance for all time-to-event, continuous, and binary survival models according to the (integrated) concordance index.

	C-index (95%-CI)		
	Overall survival	**1Y-survival status**	**Integrated C-index**
Time-to-event Models			
Accelerated Failure Time	0.70 (0.70-0.70)	0.70 (0.70-0.70)	0.70 (0.70-0.70)
CPHR	0.69 (0.69-0.70)	0.69 (0.69-0.70)	0.69 (0.69-0.70)
Boosted Decision Tree Survival	0.69 (0.69-0.70)	0.69 (0.69-0.70)	0.69 (0.69-0.70)
Random Forest Survival	0.68 (0.68-0.68)	0.69 (0.69-0.69)	0.68 (0.68-0.68)
Recursive Partitioning	0.68 (0.68-0.68)	0.68 (0.68-0.68)	0.68 (0.68-0.68)
Continuous and binary Models			
Gradient Boosting	0.70 (0.70-0.70)	0.70 (0.70-0.70)	NA
Regularized GLM	0.70 (0.70-0.70)	0.70 (0.70-0.70)	NA
GLM	0.70 (0.70-0.70)	0.70 (0.70-0.70)	NA
Support Vector Machines	0.70 (0.70-0.70)	0.69 (0.69-0.69)	NA
Multilayer Perceptron	0.61 (0.61-0.61)	0.69 (0.69-0.69)	NA
Naïve Bayes[a]	NA	0.69 (0.69-0.69)	NA
Random Forest	0.69 (0.69-0.69)	0.69 (0.69-0.69)	NA
Extreme Gradient Boosting	0.68 (0.68-0.68)	0.68 (0.68-0.68)	NA
K-Nearest Neighbors	0.67 (0.67-0.67)	0.68 (0.67-0.68)	NA
Bagging	0.67 (0.66-0.67)	0.66 (0.66-0.66)	NA

Abbreviations: 1Y=one year; C-index=concordance index; CI=confidence interval; CPHR=cox proportional hazards regression; GLM=generalized linear models; NA=not available
[a] Naïve Bayes fits to categorical data only.

Secondary metrics

Secondary metrics related to model deployment and clinical implementation varied across all models (Table 3). AFT, CPHR, and (regularized) generalized linear models were the only models with inferential utility. AFT, CPHR, boosted decision trees survival, recursive partitioning, and random forest survival were the only models that can analyze time-to-event data and thus compute subject-level survival curves. The application loading time varied between 0.2 seconds and 45 minutes. The 100-fold bootstrapped prediction time varied between 1.9 seconds and four minutes on a single central processing unit.

TABLE 3. Secondary metrics for model performance and deployment.

	Interpretability		Predictive Applicability			Computational Efficiency[a]		
Model	**Inference**	**Prediction**	**Binary**	**Continuous**	**Survival Curves**	**Size (Mb)**	**Load Time (s)**	**Prediction Time (s)**
AFT	X	X	X	X	X	20	0.9	1.9
Bagging	-	X	X	X	-	16,380	1,335	31.8
Blackboost	-	X	X	X	X	36,790	2,455	234.3
CPHR	X	X	X	X	X	37	1.7	7.5
Recursive Partitioning	-	X	X	X	X	490	52.1	3.4
BDT	-	X	X	X	-	300	8.2	2.1
GLM	X	X	X	X	-	1	0.2	1.7
GLMnet	X	X	X	X	-	109	6.7	2.3
K-Nearest Neighbors	-	X	X	X	-	91	5.6	1.9
Multilayer Perceptron	-	X	X	X	-	45	1.4	17.4
Naïve Bayes	-	X	X	-	-	82	2.9	13.0
Random forest	-	X	X	X	-	1,100	41.4	10.1
Random Forest Survival	-	X	X	X	X	6,350	65.7	139.0
Support Vector Machine	-	X	X	X	-	111	4.8	4.4
XBDT	-	X	X	X	-	92	2.4	1.5

Abbreviations: AFT=accelerated failure time; CPHR=Cox proportional hazards regression; GLM(net)= (Lasso and elastic-net regularized) generalized linear models; Mb = megabyte; s = seconds; TTE=time-to-event; (X)BDT= (extreme) boosted decision trees

[a] *Based on a 100-fold bootstrapped model.*

Deployment

Although the AFT model demonstrated similar to superior performance in terms of discrimination and calibration, it outperformed competing statistical and machine learning algorithms in terms of interpretability, predictive applicability, and computational efficiency. Therefore, it was selected as back end for the online survival prediction tool. (https://cnoc-bwh.shinyapps.io/gbmsurvivalpredictor/). The estimated survival profile for a hypothetical patient is shown in Figure 3.

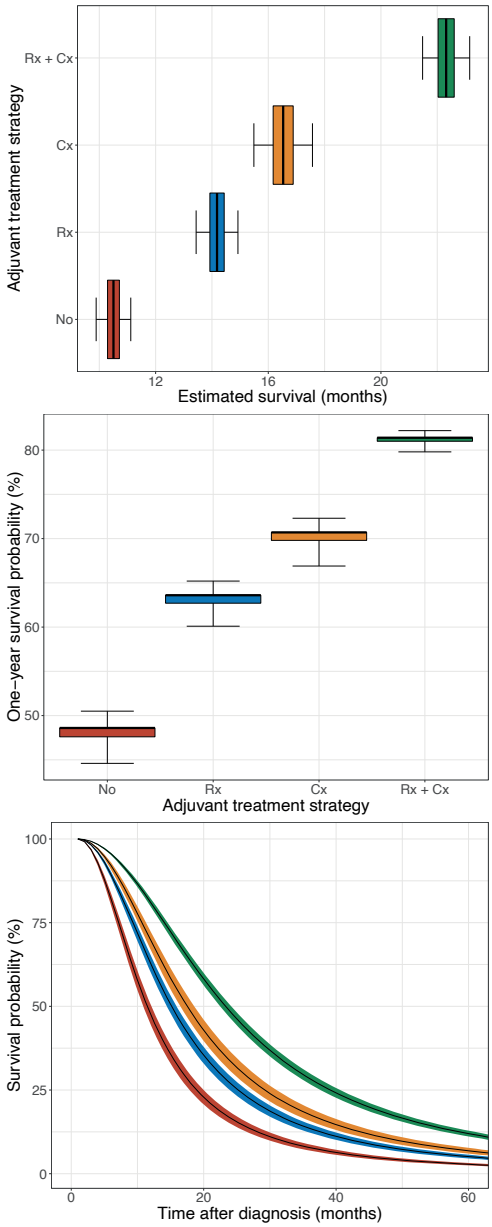


FIGURE 3. Estimated survival profile of a hypothetical patient (male, 50-years old, white, non-Hispanic, married, insured, left-sided, frontal lobe, confined to its primary location, 50mm in size, gross-total resection), plotted per adjuvant treatment strategy. Personalized estimates of overall survival in months (upper left), one-year survival probability (upper right), and five-year survival curves (lower) as predicted by the accelerated failure time model. The boxes and whiskers in the boxplots represent the 50% and 95% confidence interval, respectively. The ribbons in the survival curves represent the 95% confidence intervals. Abbreviations: Rx=Radiotherapy; Cx=Chemotherapy.

Discussion

This manuscript and the accompanying online prediction tool provide a framework for individualized survival modeling in patients with glioblastoma that is generalizable to other cancer and neurosurgical patients. Although prior investigation in this area tends to focus on metrics of prediction performance, we advocate a multimodal assessment when constructing and implementing clinical prediction models. The online prediction tool provides interactive, online, and graphical representations of expected survival in glioblastoma patients.

Few other groups have developed an online survival prediction tool for glioblastoma patients.[14–16] Gorlia et al. developed multiple nomograms based on a secondary analysis of trial data using age at diagnosis, World Health Organization performance score (WPS), extent of resection, Mini-Mental State Examination (MMSE) score, and O6-methylguanine–DNA methyltransferase (MGMT) methylation status as input features, thereby achieving a maximum C-index of 0.66.[14] Gittleman et al. developed similar nomograms including sex as an input feature and Karnofsky Performance Status (KPS) score as a measure of functional status. However, model discrimination remained similar (C-index 0.66).[15] Marko et al. developed a model in which extent of resection was modeled as a continuous covariate. This group also utilized an AFT model to account for the violated proportional hazards assumption and achieved a C-index of 0.69.[16] Higher discriminatory performance (C-index 0.63-0.77) was achieved in studies that used machine learning algorithms to analyze complex, high-dimensional data structures, such as genomic, imaging, and health-related quality of life data.[17–25] Although many machine learning algorithms are ideally suited for superior prediction performance by utilizing these high-dimensional data structures, increasing model complexity may incur other costs in terms of interpretability, ease of use, computation speed, and external generalizability.

Limitations

Due to the retrospective nature of the data acquisition, it cannot be excluded that adjuvant therapy was administered at outside hospitals and not corresponded back to the reporting hospital. However, because of the short survival period in this patient population, the percentage of patients with complete survival follow-up is exceptionally high. Although clinically essential features were included to mitigate the risk of confounding, the possibility of influence from unmeasured confounders cannot be excluded. Randomized data would be ideal; however, it is practically and financially infeasible to establish a cohort on this scale, and it has become ethically unjustifiable

to randomize newly diagnosed patients to a placebo arm now that a proven, effective adjuvant treatment for glioblastoma has emerged.[26] Predictive modeling on this scale remains therefore bound to observational data, thereby highlighting the need for exploring analytical approaches to mitigate confounding.

On average, 3.3% of all data points were missing in the total data set, which was multiply imputed by means of a random forest algorithm to mitigate the risk of systematic bias associated with a complete-case analysis. Nonetheless, survival performance in the current study is limited by the type and number of features included in the SEER registry. As a result, KPS score, isocitrate dehydrogenase 1 (IDH1) mutation, 1p/19q co-deletion, and MGMT methylation status were not included in the current iteration of the prediction model. Despite these limitations, the performance of the current proposed prediction tool exceeds that of the currently available prediction tools and even approximates the performance of many complex radiogenomic models,[17–25] yet with the ease, speed, accessibility, interpretability, and generalizability of clinical prediction tools. Furthermore, this study presents a framework that can be updated and reiterated when novel variables are added to the SEER registry or when novel large-scale multicenter glioblastoma registries are assembled. Because these models are trained on data from thousands of patients from numerous hospitals across the U.S., we expect the fitted models to be less prone to overfitting to data from a single institution and plausibly more generalizable to patients from diverse geographic regions undergoing a variety of clinical treatments.

Implications

Survival prognostication is critical for clinical and personal decision-making in glioblastoma patients. Although our current prediction tool provides an interactive interface for survival modeling with potential clinical utility, it is designed as a research tool and should not be implemented in clinical practice prior to prospective validation on multiple heterogenous cohorts. Using a population-based registry might be more representative of the typical glioblastoma patient in the US; however, testing the current model on single institutional or multicenter data might be essential to confirm its prognostic value at point-of-care. Furthermore, predictive models should inform rather than direct clinical decision-making. We advocate a multidimensional approach for survival prognostication, in which model predictions are adjusted and balanced against complementary information that is available including clinical experience, neuropsychological testing, imaging data, and genomic information.

Many statistical and machine learning algorithms allow for the analysis of historical patient cohorts to predict survival in new patients. However, prediction performance,

interpretability, clinical utility, computational efficiency, and their associated limitations vary widely across different models due to their mathematical underpinnings. CPHR has emerged as the cornerstone of survival analysis but is limited by the assumption of proportionality, which assumes that the relationship between covariate and outcome is constant over time. In the real world, this association is often dynamic, and the assumption of proportionality is effectively violated. The AFT model does allow for increasing or decreasing covariate risk contribution over time, which is particularly useful in individualizing survival predictions. The AFT model has been shown to be a valuable alternative to CPHR in simulation studies,[27] as well as survival studies on glioblastoma patients.[16]

Molecular markers (e.g., IDH1 mutation, 1p19q codeletion, and MGMT methylation status), as well as functional status (e.g., KPS, MMSE), have been demonstrated to impact survival in glioblastoma patients and are commonly used for stratifying patient cohorts in clinical decision-making. However, they have not yet been included in large-scale, multicenter registries. Inclusion of these variables would improve individual patient survival modeling. Furthermore, granular information with regards to the healthcare setting (e.g., academic versus non-academic) and provided clinical care (e.g., volumetric measurements of tumor size and extent of resection, as well as the timing, type, dose, and sequence of adjuvant treatment) would be valuable to further improve model performance. If addition of any of these variables improves model performance only slightly, however, it may be preferable to exclude some predictors for ease of use at the point of care. Another method to overcome the lack of large-scale granular data sets could be to explore the concept of transfer learning, a common machine learning approach of updating a pre-trained model on novel data sources or even different outcomes.[28] In the context of glioblastoma survival prediction, this could mean developing a base model on population-based data, which is further trained on institutional data to fit institutional patterns and include relevant institutional parameters not available in population-based registries.

Although many machine learning algorithms show great predictive performance, their utility is often limited to continuous and binary models, which merely provide point estimates of overall survival and one-year survival probability at a given point in time, respectively. Transferring the predictive power of these algorithms to time-to-event models allows for the computation of subject-level survival curves, thereby enabling more granular insight into expected survival. Furthermore, time-to-event models can be trained on patients with either complete or incomplete follow-up, which mitigates the systematic bias associated with exclusion of the latter group. Although many machine learning models demonstrate high performance in the

academic realm,[29] lack of interpretability and computational inefficiency hinders their deployment in the clinical realm. When evaluating models for clinical deployment, we recommend evaluating fitted models on several criteria rather than a singular focus on prediction performance since factors unrelated to prediction performance (such as interpretability or applicability) can exclude high-performing models from clinical deployment. Although the AFT model was selected due its high overall performance, the difference in prediction performance was not always clinically meaningful, thereby emphasizing the importance of taking into account these secondary metrics as well. Furthermore, the prediction performance can change as the number and nature of the input features change. For example, the assembly of multimodal data including radiogenomics data might call for alternative analytical approaches in the near future.

Prognostication is and always has been aimed at a moving target and future factors impacting clinical course cannot be modeled, most importantly advances in clinical care. Prediction performance therefore remains an asymptotic ideal for which perfection will never be reached. Future research should focus on developing clinically meaningful and interpretable prediction tools. Improving the end-user transparency regarding the underlying predictive mechanisms and the inherent limitations allows for a safe and reliable implementation of survival prediction tools in clinical care.

Conclusion

This study provides a framework for the development of survival prediction tools in cancer patients, as well as an online calculator for predicting survival in glioblastoma patients. Future efforts should focus on developing additional algorithms that can train on right-censored survival data, improve the granularity of population-based registries, and externally validate the proposed prediction tool.

Supplementary material

Supplementary tables and figures available online at:

https://academic.oup.com/neurosurgery/article/86/2/E184/5581744#supplementary-data

References

1. Ostrom QT, Gittleman H, Liao P, et al. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. Neuro-Oncology. 2017;19(suppl_5):v1-v88. doi:10.1093/neuonc/nox158
2. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. doi:10.1136/bmj.g7594
3. Mohanty S, Bilimoria KY. Comparing national cancer registries: The National Cancer Data Base (NCDB) and the Surveillance, Epidemiology, and End Results (SEER) program. J Surg Oncol. 2014;109(7):629-630. doi:10.1002/jso.23568
4. Altekruse SF, Rosenfeld GE, Carrick DM, et al. SEER Cancer Registry Biospecimen Research: Yesterday and Tomorrow. Cancer Epidemiol Biomarkers Prev. 2014;23(12):2681-2687. doi:10.1158/1055-9965.EPI-14-0490
5. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3(8). doi:10.1136/bmjopen-2013-002847
6. Senders JT, Arnaout O, Karhade AV, et al. Natural and Artificial Intelligence in Neurosurgery: A Systematic Review. Neurosurgery. 2017. doi:10.1093/neuros/nyx384
7. Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. Vol 1857. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000:1-15. doi:10.1007/3-540-45014-9_1
8. Zare A, HOSSEINI M, MAHMOODI M, MOHAMMAD K, ZERAATI H, HOLAKOUIE NAIENI K. A Comparison between Accelerated Failure-time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients. Iran J Public Health. 2015;44(8):1095-1102.
9. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
10. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. Stat Med. 2011;30(10):1105-1117. doi:10.1002/sim.4154
11. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/. Published 2008. Accessed June 11, 2018.
12. Kuhn M. Building Predictive Models in R Using the caret Package | Kuhn | Journal of Statistical Software. Journal of Statisticla Software. 2008. doi:10.18637/jss.v028.i05
13. Chang W, Cheng J, Allaire JJ, et al. Shiny: Web Application Framework for R.; 2018. https://CRAN.R-project.org/package=shiny. Accessed June 11, 2018.
14. Gorlia T, Bent MJ van den, Hegi ME, et al. Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. The Lancet Oncology. 2008;9(1):29-38. doi:10.1016/S1470-2045(07)70384-4
15. Gittleman H, Lim D, Kattan MW, et al. An independently validated nomogram for individualized estimation of survival among patients with newly diagnosed glioblastoma: NRG Oncology RTOG 0525 and 0825. Neuro Oncol. 2017;19(5):669-677. doi:10.1093/neuonc/now208
16. Marko NF, Weil RJ, Schroeder JL, Lang FF, Suki D, Sawaya RE. Extent of Resection of Glioblastoma Revisited: Personalized Survival Modeling Facilitates More Accurate Survival Prediction and Supports a Maximum-Safe-Resection Approach to Surgery. JCO. 2014;32(8):774-782. doi:10.1200/JCO.2013.51.8886

17. Hilario A, Sepulveda JM, Perez-Nuñez A, et al. A Prognostic Model Based on Preoperative MRI Predicts Overall Survival in Patients with Diffuse Gliomas. American Journal of Neuroradiology. 2014;35(6):1096-1102. doi:10.3174/ajnr.A3837
18. Cui Y, Ren S, Tha KK, Wu J, Shirato H, Li R. Volume of high-risk intratumoral subregions at multi-parametric MR imaging predicts overall survival and complements molecular analysis of glioblastoma. Eur Radiol. 2017;27(9):3583-3592. doi:10.1007/s00330-017-4751-x
19. Mazurowski MA, Desjardins A, Malof JM. Imaging descriptors improve the predictive power of survival models for glioblastoma patients. Neuro Oncol. 2013;15(10):1389-1394. doi:10.1093/neuonc/nos335
20. Cui Y, Tha KK, Terasaka S, et al. Prognostic Imaging Biomarkers in Glioblastoma: Development and Independent Validation on the Basis of Multiregion and Quantitative Analysis of MR Images. Radiology. 2015;278(2):546-553. doi:10.1148/radiol.2015150358
21. Kickingereder P, Burth S, Wick A, et al. Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. Radiology. 2016;280(3):880-889. doi:10.1148/radiol.2016160845
22. Lao J, Chen Y, Li Z-C, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. Scientific Reports. 2017;7(1):10353. doi:10.1038/s41598-017-10649-8
23. Li Q, Bai H, Chen Y, et al. A Fully-Automatic Multiparametric Radiomics Model: Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme. Scientific Reports. 2017;7(1):14331. doi:10.1038/s41598-017-14753-7
24. Mauer MEL, Taphoorn MJB, Bottomley A, et al. Prognostic Value of Health-Related Quality-of-Life Data in Predicting Survival in Patients With Anaplastic Oligodendrogliomas, From a Phase III EORTC Brain Cancer Group Study. JCO. 2007;25(36):5731-5737. doi:10.1200/JCO.2007.11.1476
25. Gómez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Palacios-Corona R, Trevino V. Integration and comparison of different genomic data for outcome prediction in cancer. BioData Mining. 2015;8(1):32. doi:10.1186/s13040-015-0065-1
26. Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. New England Journal of Medicine. 2005;352(10):987-996. doi:10.1056/NEJMoa043330
27. Chiou SH, Kang S, Yan J. Fitting Accelerated Failure Time Models in Routine Survival Analysis with R Package aftgee. Journal of Statistical Software. 2014;61(11). doi:10.18637/jss.v061.i11
28. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191
29. Senders JT, Staples PC, Karhade AV, et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. World Neurosurgery. 2018;109:476-486.e1. doi:10.1016/j.wneu.2017.09.149