

Machine learning-based survival prediction in glioma using large scale registry data—the importance of chemotherapy and radiation therapy management as predictive features

Zhao R¹, Zhuge Y², Camphausen K², Krauze A^{1,3}

¹University of British Columbia, Faculty of Medicine, 317 - 2194 Health Sciences Mall, Vancouver, BC V6T 1Z3.

²Radiation Oncology Branch, Center for Cancer Research, National Cancer Institute, NIH, 9000 Rockville Pike, Building 10, CRC Bethesda, MD 20892, USA.

³BC Cancer Surrey, 13750 96 Ave, Surrey, BC V3V 1Z2.

ARTICLE INFO

Article history:
Received
Received in revised form
Accepted
Available online

Keywords:
Artificial intelligence,
Machine learning,
Cancer registry,
Glioma,
Survival prediction,
Treatment planning;

ABSTRACT

Gliomas are the most common central nervous system tumors exhibiting poor survival, quality of life and neurological outcomes prompting significant discussion surrounding optimisation of the aggressiveness of management. The ability to estimate prognosis is crucial for both patients and providers in order to select the most appropriate treatment. Previous attempts at predicting survival outcomes have relied on clinical parameters (age, KPS, gender) and resection or methylation status and statistical models to create prognostic groups limiting survival prediction due to selection bias and tumor heterogeneity. Machine learning (ML) allows for more sophisticated approaches to survival prediction amalgamating real world clinical, molecular and imaging data. We wanted to examine clinical parameters needed to achieve superior predictive accuracy in order to help advance guidelines for the creation and maintenance of robust large-scale glioma registries.

2021 Sciforce Publications. All rights reserved.

*Corresponding author. E-mail: andra.krauze@bccancer.bc.ca

Introduction

Gliomas are the most common central nervous system tumors. Gliomas are typically managed by maximal safe resection followed by radiation therapy, chemotherapy or in rare cases observation depending on the histology and clinical context [1,2,5]. The survival of glioma remains overall extremely poor with a 5-year overall survival less than 35% [3]. The ability to estimate prognosis is crucial for both patients and providers in order to select the most appropriate treatment that is sufficiently aggressive to allow for tumor control while minimizing adverse long term normal tissue changes but also appropriately de-escalated when prognosis is poor and emphasis is on patient quality of life and best supportive care. Multiple attempts have been made to design robust scoring systems predictive of outcome for both low [4] and high- grade glioma [5,6]. Mostly, these have relied on clinical parameters (age, Karnofsky Performance Status (KPS), gender) and resection or methylation status as well as statistical models to create prognostic groups [7,9,10] with survival prediction lacking

generalisability secondary to: 1) small cohorts of patients, 2) the inclusion of (mostly) trial patients and 3) management of these patients at tertiary academic centers. The current approaches present limitations as: 1) most glioma patients are treated off study; 2) outside of centers of excellence; 3) do not necessarily benefit from expert pathology review or molecular analysis, and 4) significant tumor heterogeneity further undermines the ability to predict survival. Existing evidence already suggests that patients falling outside of these settings may have poorer outcomes [11,12] and therefore existing scoring systems may not necessarily reflect their prognosis. Machine learning (ML) can allow for more sophisticated approaches to clinical, molecular and imaging data to predict risk and survival [13-18]. In this study we aimed to explore the effectiveness of both ML and statistical approaches to predict survival in glioma patients using a set of commonly available clinical features in a real-world evidence cohort using a larger glioma dataset representative of a high volume publicly funded system – the BC Cancer registry which includes patients of all glioma histological subtypes

treated largely off trial over the course of nearly 20 years in the province of British Columbia, Canada.

2. Material and Methods

2.1 Study cohort

Data from 3907 glioma patients diagnosed between 2000 and 2018 was obtained from the BC Cancer Registry following research ethics board approval. Patients who received treatment out of province (14) or for whom any of the features necessary for the analysis were not captured (317), were excluded. Only patients with a pathological diagnosis of glioma were included and uncommon glioma histologies with less than 0.5% recorded cases in the dataset were excluded.

Overall, 3462 patients were included in the analysis.

2.2 Training and Test datasets

The overall dataset was split into two into two mutually exclusive datasets with a 7:3 training data to test data ratio by random sampling. The same training and test dataset were used for all models. Each dataset contained the following features: age, sex, administration of chemotherapy, surgical resection, administration of radiation therapy, tumor histology and tumor site.

2.3 Modeling and Prediction

Three modeling methods were implemented using open source python libraries, *scikit-survival* by Pölsterl et al [19-21]. Each model was trained using the entire training set and was applied to predict a risk score and survival function for each patient in the test set. Predicted median survival time was the time at 0.5 survival probability derived from the survival function. The accuracy of the survival prediction was evaluated by Concordance Index (c-index), calculated using the python package *Lifelines* (<https://doi.org/10.5281/zenodo.1252342>).

2.3.1 Cox Proportional hazards (CPH) model

The CPH model is the linear regression model most widely used in survival studies to predict the risk of an outcome based on multiple variables [24]. We built a CPH model using clinical features in the training data as covariates.

2.3.2 Support Vector Machine (SVM) model

We optimized a linear SVM classifier through a hyperparameter search to find the best regularization hyperparameter which was used to train the classifier using the entire training set.

2.3.3 Random Forest (RF) Model

The random forest consists of 1000 decision trees trained using the training dataset. Risk score prediction were the average across all trees in the forest [25]. The feature importance score of each feature was calculated by the decrease in concordance index of the test dataset if it was made unavailable by assigning random value to it for all patients [26].

2.3 Kaplan Meier (KM) Survival curves

KM survival curves were plotted to compare training and test datasets. The log rank-test [22, 23] was used to determine if there is a significant difference between the survival distributions of the training and test dataset, using the Python package *Lifelines* (<https://doi.org/10.5281/zenodo.1252342>).

While there is currently no established consensus on how to approach representation of this type of predictive survival analysis in figures, KM survival curves were plotted to compare the median survival time predicted by the model and the clinically recorded survival time of the patients in the test dataset as other authors have employed similar approaches [27-29]. Patients who were still alive at the time of the analysis were removed for plotting these two KM curves since model predictions would not include censoring status at the end of the study and a KM curve of predicted median survival for those patients whose censoring status is not known would not be appropriate. Therefore, we removed all censored pts from the recorded test data and predicted survival data. In order to show potential clinical application, we used both c-index (which includes both censored and uncensored patients as it is based on event risk ranking) and log rank test (resulting in survival time predicted vs recorded survival time where we only used uncensored data).

3. Results

3.1 Clinical characteristics

3462 patients with a diagnosis of glioma treated between 2000 and 2018 were included in the analysis. 2113 (61%) were male and 1349 (39%) female. Histological distribution was: glioblastoma 1555(45%), astrocytoma 926 (27%), oligodendroglioma 299 (9%), mixed glioma 267 (8%), anaplastic oligodendroglioma 130 (4%), glioma malignant 118 (3%), anaplastic astrocytoma 70 (2%), other glioma histologies (2%) (Table 1). 1119 (33%), 795 (23%) and 117 (3%) of tumors originated in the frontal lobe, temporal and occipital lobe respectively. 2410 (70%) had maximal safe surgical resection whilst the remainder, 1052 (30%) had biopsy only. 2730 (79%) total patients received RT. At the time of the analysis 1831 (53%) had not received chemotherapy, 1515 (44%) had received chemotherapy immediately following diagnosis, 81 (2%) received subsequent chemotherapy and 35 (1%) received both initial and subsequent chemotherapy. Molecular characterization including MGMT status and patient performance status were not captured in the BC Cancer registry data.

3.2 Training and testing datasets

The training and testing datasets were created using random sampling of the overall dataset in a 7:3 ratio and there was no statistically significant difference in survival between the training and testing datasets (log rank test $p=0.99$) (Figure 1) and minimal difference in c-index between training and test dataset across all models (Figure 2).

3.3 ML models for survival prediction

C-index is a commonly used method comparing the ranking of survival time recorded in a clinical dataset to the ranking of predicted risk for death. A score c-index of 0.5 is expected from random prediction and 1.0 is expected if two rankings are in perfect concordance [22]. Concordance index (c-index) adjusted for right censoring was calculated for the test dataset using risk score predicted by each model training different combinations of features numbered as 1) Age, 2) Sex, 3) Tumor Histology, 4) Tumor site, 5) Tumor resection, 6) Radiation therapy (RT), 7) Chemotherapy. The prediction accuracy was lowest when the model did not take into account information on management (features 6 and 7 representative of administration of RT and chemotherapy respectively) (Figure 3). The highest survival prediction accuracy was obtained using a model that takes into account information on patient characteristics, tumor characteristics and cancer management with CI of 0.767, 0.771 and 0.757 for CPH, SVM and RF models respectively (Figure 3).

3.4 Clinical features predictive importance

The variables available in the dataset and employed in the analysis were age, sex, administration of chemotherapy, surgical resection, administration of radiation therapy, tumor histology and tumor site. The predictive value of each variable in the CPH model, calculated as the c-index of the test dataset obtained from a univariate CPH analysis using this variable, ranging from 0.5 indicating random prediction, for gender and 0.69 for age (Table 2). In the RF model, each feature was assigned a feature importance score calculated by the decrease in the concordance index of the test dataset predicted by RF model if this feature were not available (Table 3). Both models show that chemotherapy, followed by RT are more predictive than any features other than age.

3.4 Survival models

All three models CPH, SVM and RF performed reasonably well (Figure 4 A, B, C) as seen in the predicted survival probability for 3 sample patients (figure simplified to include only 3 sample patients for ease of interpretation) (entire patient sample supplemental Figure 1). The clinical information for each patient is as follows: Patient 1: 67-year-old male diagnosed with oligodendroglioma NOS in overlapping areas of the brain (site = Brain, overlapping lesion) who received chemotherapy and no surgical resection or radiation. Patient 2: 83-year-old female diagnosed with glioblastoma located in the cerebrum managed with surgical resection only (no chemotherapy or radiation). Patient 3 was a 69-year-old male diagnosed with anaplastic oligodendroglioma located in the front lobe who passed away 7 days after diagnosis not having received any therapy. Only uncensored patients were used for generation the KM survival curve as the model predictions do not include censoring status. There was no statistically significant difference between the recorded survival time distribution and the predicted median survival time distribution using the CPH and RF models for the

test dataset, $p = 0.07$ and $p = 0.61$ respectively (Figure 4 D and F). The difference between SVM predicted median survival distribution was statically significant from the recorded survival time, $p < 0.005$ (Figure 4E).

4. Discussion

ML as a tool towards superior prediction of clinical outcomes has increased in popularity in all domains of medicine including oncology [16, 17 30, 31, 32, 33] driven by the need to rapidly harness clinically relevant results when prospective data is unavailable and impossible to obtain such as in the context of the COVID-19 pandemic.

Using a large retrospective glioma patient cohort originating in the BC Cancer registry in British Columbia, Canada, we explore the ability to predict survival while employing exclusively non radiomic, non-molecular data features generally available in most high volume cancer centers treating gliomas. We achieved excellent survival prediction with c-index ranging from 0.757 (RF model) to 0.771 (SVM model) while including the following features: 1) Age, 2) Sex, 3) Tumor Histology, 4) Tumor site, 5) Tumor resection, 6) Radiation therapy (RT), 7) Chemotherapy, the lowest common denominators embedded in most large brain tumor registries.

Most ML survival prediction studies aimed at patients with glioma center around MRI radiomics or histological features as a result involving smaller patient populations as proof of concept (Tan et al 2019 MRI radiomics, $n = 147$), Papp et al 2018 (PET, $n = 70$) Mobadersany et al. 2018 (histology and genomics $n = 769$), Mizutani et al 2019 (radiation dosimetry, $n = 35$). Our patient characteristics and tumor management features are more similar to large retrospective registry studies such as the SEER database which using traditional statistical analysis has been employed to develop nomograms in the context of low grade glioma (Zhao Y et al, 2019) (3732 patients), oligodendroglioma (2689 patients) (Brandel et al., 2017), high-grade glioma (6395 patients) (Yang et al., 2020) and glioblastoma [16].

The CPH model has been the gold standard for survival analysis involving a semi-parametric statistical modelling approach where the survival outcome is a linear combination of predictive variables. Although popular, CPH operates on the underlying assumption that predictive variables are independent and do not interact, and their impact on survival do not change over time. We hypothesized that these assumptions are unlikely to hold true when considering: 1) the large number of predictive features potentially available in cancer patients and 2) that fact that these features are likely to interact with each other in an unforeseen manner [16]. Therefore, we selected two ML methods Support Vector Machine (SVM) and Random Forest (RF) that can deal with predictive features that have potential interactions and are easy to interpret and generalizable in terms of presence in medical literature [16,17]. The SVM approach assigns weight to each predictive feature to produce a score that maximizes concordance between predicted survival ranking and recorded survival time ranking [28]. By contrast the RF

approach takes the average of a collection of decision trees where the branches are split based on values of the predictive features [29]. All three models ultimately produced equal or indeed superior c-index in comparison with the literature [16, 30-36].

We achieved a higher c-index using the ranking based SVM model as compared to the other two models. However, our SVM model exhibited a difference between SVM predicted median survival time distribution and the recorded survival time. This was likely secondary to our selection of a rank based SVM approach which optimizes risk ranking. A regression based SVM model can be explored in further analysis for potential better survival time prediction [27]. We found some parallels with Nemati et al. who employed real world data to predict hospital discharge for COVID-19 patients [17] and Senders et al. who employed 20821 glioblastoma patients originating in SEER database to predict survival at 1 year following diagnosis. Both employed c-index as a performance metric and focused on risk factor analysis and 1 year survival classification respectively but stopped short of comparing predicted and recorded discharge or survival time as c-index is solely based on ranking of times of events in all possible pairs [16,17]. To enhance clinical applicability, it was important to us to compare the predicted survival time to actual survival time and we employed log rank test to accomplish this using uncensored data. Further analysis alternatives could include using alternate weighting methods to include censored data and comparison with accelerated failure time (AFT) algorithms [16,19,37].

Similar to other studies employing large scale retrospective data, the current study is limited by lack of information with respect to patient performance status, molecular features and the detailed timing of chemotherapy administration in relationship to diagnosis and RT administration, all currently not being collected as part of the BC Cancer registry. Additional limitations are posed by the lack of vital status information for some patients and the possibility that abrupt events not directly related to patient characteristics, histology or management that may have affected outcome. Whilst we do have information on the intention to have administered RT or chemotherapy, the current analysis does not take into account whether the treatment was in fact ultimately administered or completed as intended.

Ultimately both ML methods achieved good predictive ability on par with the gold standard (CPH) in a large dataset but similarly to other studies [16,17], they did not outperform the CPH statistical model acknowledging that in the context of additional highly complex interacting features (radiomic, RT dosimetry, detailed genomic and pharmaceutical data), machine and deep learning models are likely to perform better [13-15]. The robust capture and inclusion of the above features comprise future directions in the field of oncology.

The patient population and outcomes of our population are similar to other large series [16, 30-33] and we determined that management as a feature was crucial in achieving superior predictive capability for all models. Our study is a first step

towards future investigations into the potential of involving ML models in personalized treatment planning where model predicted survival times for different treatment options can be taken into consideration when determining the optimal management plan for each patient especially in cases such as glioma management, where the intricacies of administration of chemotherapy can be a source of clinical debate (concurrent versus sequential, number of cycles and patient selection). The fact that these aspects of management are often incompletely captured and hence often used as a dichotomy chemotherapy (yes/no) should be remedied. Future studies are required to address the issue of how ML encapsulates the a priori complexity of clinical decision making and the implications for patient outcomes juxtaposed with the ability to create clinically meaningful ML models that appropriately disentangle the multiple factors involved.

Our efforts in this study highlight both the need to create reliable clinician/ML connections as much as the need for increasingly robust datasets that capture the intricacies of patient management in large scale registries. This means more clinical oversight of data coding in registries as well as quality assurance of patient management as is now increasingly performed via peer review in tertiary care institutions. The ability to work from a platform of consensus will allow for meaningful conclusions based on ML eventually on par with those currently obtained from prospective trials. Ongoing efforts and future directions involve in depth survival modelling aimed specifically at the management and outcomes of elderly patients with a glioma diagnosis as well as that of patients with lower grade gliomas and incorporation of large-scale systemic management data into existing models.

Materials and Methods: We employed three approaches: Cox Proportional hazards (CPH) model, Support Vector Machine (SVM) model, Random Forest (RF) model in a large glioma dataset (3462 patients, diagnosed 2000-2018) originating in the BC Cancer Registry to explore the most optimal approach to survival prediction. Training and testing datasets were created using random sampling in a 7:3 ratio with no statistically significant difference in survival between the training and testing sets. Features employed were age, sex, surgical resection, tumor histology and tumor site, administration of radiation therapy (RT) and chemotherapy. Concordance index (c-index) (CI) was employed to compare the ranking of survival time recorded in clinical dataset to the ranking of predicted risk for death and adjusted for right censoring using risk score predicted by each model training different combinations of features where: 1) Age, 2) Sex, 3) Tumor Histology, 4) Tumor site, 5) Tumor resection, 6) RT, 7) Chemotherapy.

Results: 2113 (61%) of patients were male and 1349 (39%) female. Histological distribution was glioblastoma 1555 (45%), astrocytoma 926 (27%), oligodendroglioma 299 (9%), mixed glioma 267 (8%), anaplastic oligodendroglioma 130 (4%), glioma malignant 118 (3%), anaplastic astrocytoma 70 (2%), other glioma histologies (2%). 2410 (70%) had maximal safe surgical resection, 1052 (30%) had biopsy only. 2730 (79%)

total patients received RT. 1631 (48%) overall received chemotherapy (1515 (44%) immediately following diagnosis, 81 (2%) subsequent chemotherapy, 35 (1%) both initial and subsequent). There was no statistically significant difference between the recorded and predicted median survival time distribution using the CPH and RF models for the test dataset, $p = 0.07$ and $p = 0.61$ respectively. The difference between SVM predicted median survival distribution was statistically significant from the recorded survival time, $p < 0.005$. All three models performed well with prediction accuracy highest (CI 0.757, 0.767, 0.771 for RF, CPH, SVM models respectively) when taking incorporating RT and chemotherapy administration features.

5. Conclusions

We achieved superior survival prediction performance with the aforementioned ML studies as compared to other ML and non-ML approaches in the literature while employing exclusively widely available clinical sets of features. The administration of chemotherapy and RT emerged as a key features raising questions as to the potential for superior results that may be achieved through further optimisation and clinical oversight of large-scale real world datasets to allow for clinically relevant results to be generated by ML approaches.

Table 1. Characteristics of training and testing sets.

		Training Set (n=2423)	Test Set (n=1039)	Total (n=3462)
Age (SD)		15	15	15
Sex	Male	1483 (61.2%)	630 (60.6%)	2113 (61.0%)
	Female	940 (38.8%)	409 (39.4%)	1349 (39.0%)
Histology	Glioblastoma, NOS	1085 (44.8%)	470 (45.2%)	1555 (44.9%)
	Astrocytoma, NOS	636 (26.2%)	290 (27.9%)	926 (26.7%)
	Oligodendroglioma, NOS	220 (9.1%)	79 (7.6%)	299 (8.6%)
	Mixed Glioma	185 (7.6%)	82 (7.9%)	267 (7.7%)
	Anaplastic Oligodendroglioma	98 (4.0%)	32 (3.1%)	130 (3.8%)
	Glioma, malignant	86 (3.5%)	32 (3.1%)	118 (3.4%)
	Anaplastic astrocytoma	47 (1.9%)	23 (2.2%)	70 (2.0%)
	Giant cell glioblastoma	19 (0.8%)	11 (1.1%)	30 (0.9%)
	Gliosarcoma	20 (0.8%)	7 (0.7%)	27 (0.8%)
	Pilocytic Astrocytoma	16 (0.7%)	8 (0.8%)	24 (0.7%)
	Fibrillary astrocytoma	11 (0.5%)	5 (0.5%)	16 (0.5%)
Site	Frontal lobe	781 (32.2%)	338 (32.5%)	1119 (32.3%)
	Temporal lobe	558 (23.0%)	237 (22.8%)	795 (23.0%)
	Brain, overlapping lesion	408 (16.8%)	187 (18.0%)	595 (17.2%)
	Parietal lobe	372 (15.4%)	140 (13.5%)	512 (14.8%)
	Occipital lobe	70 (2.9%)	47 (4.5%)	117 (3.4%)
	Brain, unspecified	90 (3.7%)	28 (2.7%)	118 (3.4%)
	Cerebrum	69 (2.8%)	25 (2.4%)	94 (2.7%)
	Brain stem	38 (1.6%)	16 (1.5%)	54 (1.6%)
	Spinal cord	13 (0.5%)	10 (1.0%)	23 (0.7%)
	Cerebellum, NOS	17 (0.7%)	7 (0.7%)	24 (0.7%)
	Ventricle, NOS	7 (0.3%)	4 (0.4%)	11 (0.3%)
Surgery	Surgical Resection	1666 (68.8%)	744 (71.6%)	2410 (69.6%)
	No Surgical Resection	757 (31.2%)	295 (28.4%)	1052 (30.4%)
Radiation	Radiation Therapy	1906 (78.7%)	814 (78.3%)	2730 (78.6%)
	No Radiation Therapy	517 (21.3%)	225 (21.7%)	742 (21.4%)
Chemotherapy	No Chemotherapy	1274 (52.6%)	557 (53.6%)	1831 (52.9%)
	Concurrent Chemotherapy	1072 (44.2%)	443 (42.6%)	1515 (43.8%)
	Subsequent Chemotherapy	53 (2.2%)	28 (2.7%)	81 (2.3%)
	Initial and Subsequent Chemotherapy	24 (1.0%)	11 (1.1%)	35 (1.0%)

Table 2. Predictive value of each variable under *Cox Proportional hazards (CPH) model*. The predictive value of

each variable is the concordance index of the test dataset using predictions made by the model containing only this variable.

Variable	Predictive value
Patient characteristic	
Age	0.69
Sex	0.50
Chemotherapy	0.63
Radiation Therapy	0.60
Surgical resection	0.57
Tumor Histology	0.60
Tumor Site	0.55

List of Abbreviations

CPH - Cox Proportional Hazards Model
 SVM - Support Vector Machine Model
 RF - Random Forest Model
 BC - British Columbia
 RT- Radiation Therapy
 KPS - Karnofsky Performace Status
 DNET- Dysembryoplastic neuroepithelial tumors
 KM - Kaplan Meier

Table 3. Most important features as identified in the **Random Forest (RF) model**. The feature importance score of each feature

is shown as calculated by the decrease in the concordance index of the test dataset predicted by RF model if this feature were not available. Features with feature importance score less than 0.001 and Histology/Site features applying only to relatively small portions of the data are not shown.

Feature	Feature Importance Score
Age	0.080
Chemotherapy	0.031
Radiation therapy	0.020
Histology = Glioblastoma	0.020
Histology =	
Oligodendroglioma	0.011
Histology = Astrocytoma	0.010
Surgical Resection	0.006
Tumor site =	
Frontal lobe	0.003
Tumor site=	
Temporal lobe	0.001

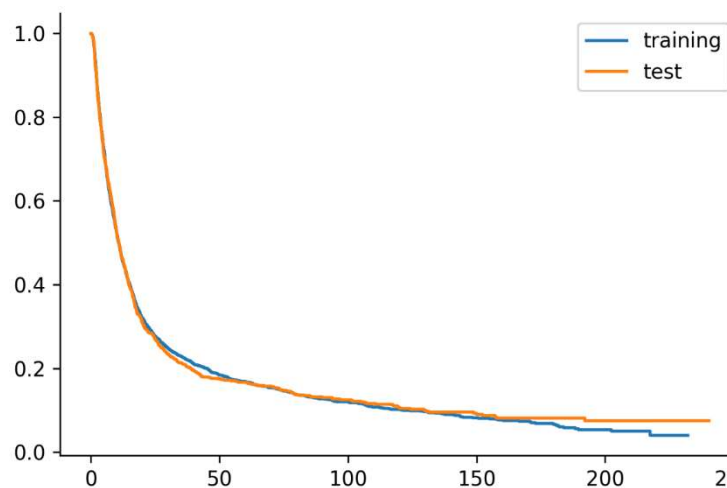


Figure 1. Kaplan Meier overall survival for the training and testing datasets.

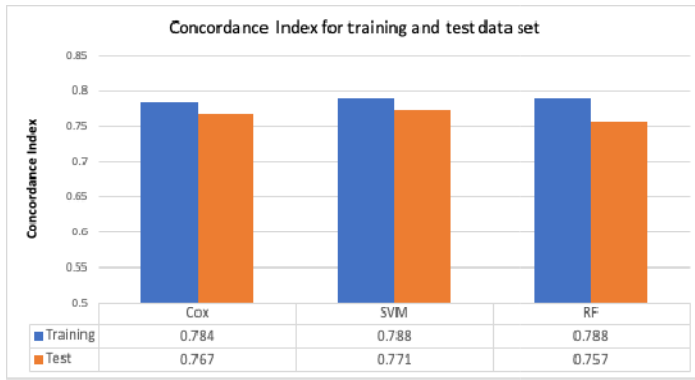


Figure 2. Survival prediction accuracy of each model for training and test dataset.

The difference in c-index between training and test dataset is minimal across all models, suggesting that no overfitting of the training data. The model predictions were generalizable to the unused test data. Cox = Cox Proportional hazards (CPH) model. SVM = Support Vector Machine (SVM) model, RF = Random Forest (RF) model.

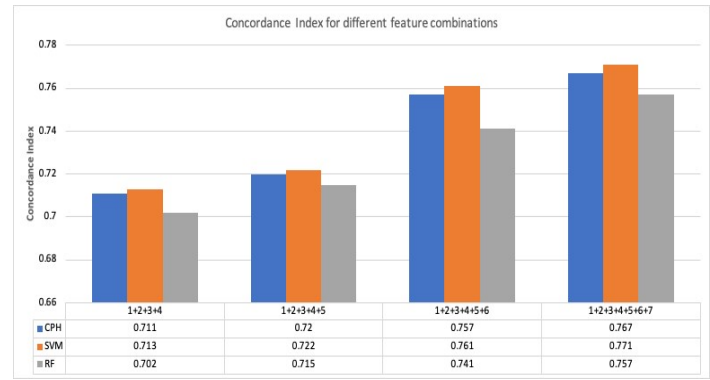


Figure 3. Survival prediction accuracy of each model using different combinations of features where: 1) Age, 2) Sex, 3) Tumor Histology, 4) Tumor site, 5) Tumor resection, 6) Radiation therapy (RT), 7) Chemotherapy. CPH = Cox Proportional hazards model. SVM = Support Vector Machine (SVM) model, RF = Random Forest (RF) model.

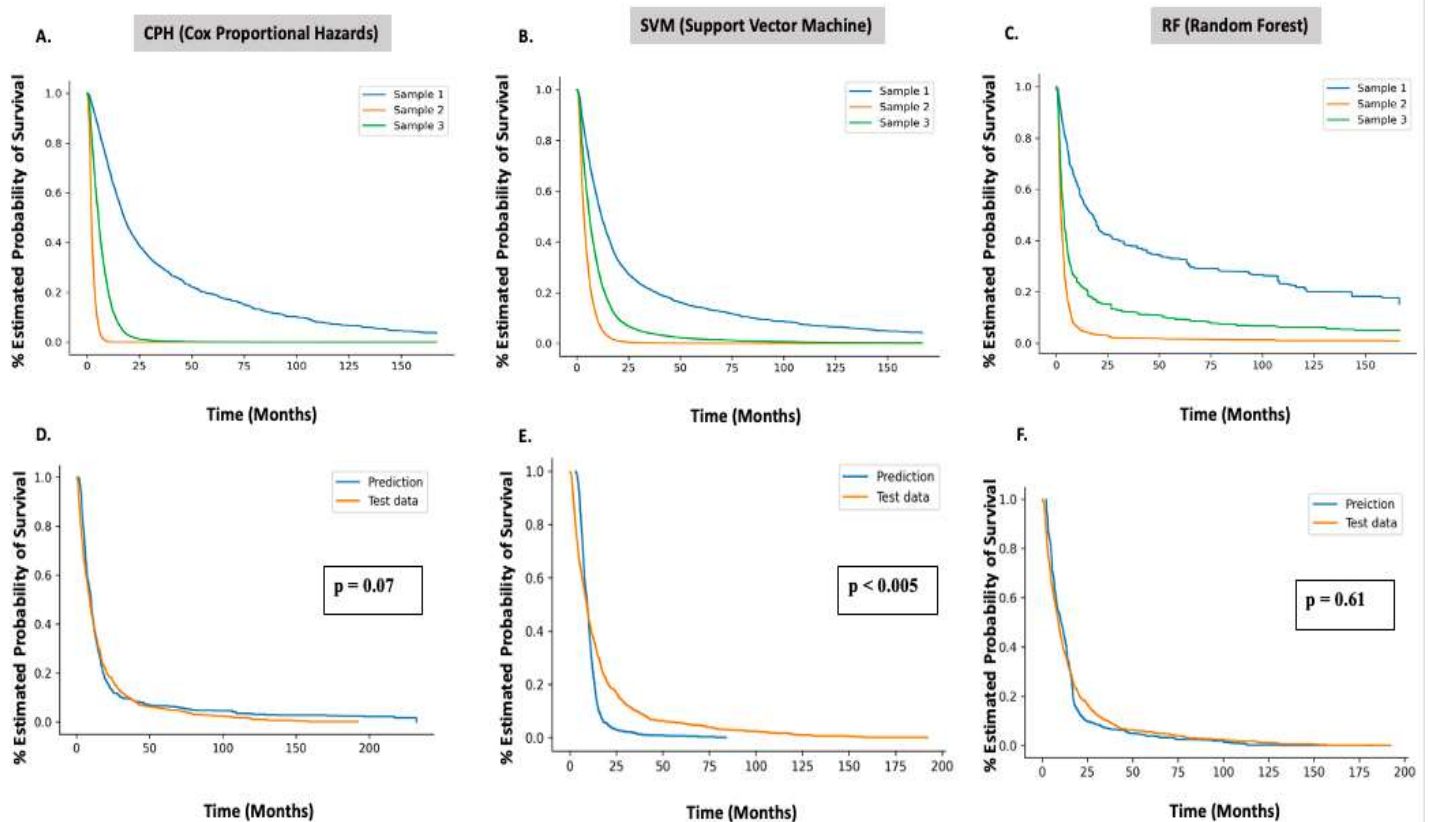


Figure 4. Predicted survival curves for three patients in the test dataset using **A.** CPH model. **B.** SVM model. **C.** RF model respectively. Corresponding KM survival curves for recorded survival times in test dataset and KM survival curve for predicted median survival times using **D.** CPH model. **E.** SVM model. **F.** RF model. CPH = Cox Proportional hazards model. SVM = Support Vector Machine (SVM) model, RF = Random Forest (RF) model.

Availability of data and material

The datasets supporting the conclusions of this article were obtained from the BC Cancer Registry following approved Research Ethics Board Review.

Funding

This work was supported by the Porte-Hungerford Neuro-Oncology Grant held by Dr. A. Krauze and the BC Cancer Summer Studentship Program.

Competing Interests

The authors declare that they have no competing interests.

Authors' contributions

RZ conceived the study, organized patient registry data, created and optimized non-machine and machine learning models and co-drafted the manuscript.

YZ reviewed the manuscript and conceived the original joint study of machine learning based glioma patient outcomes analysis employing large scale data.

KC reviewed the manuscript and conceived the original joint study of machine learning based glioma patient outcomes analysis employing large scale data.

AVK conceived the study, participated in its design and coordination, collected patient data and co-drafted the manuscript.

Conclusions

The administration of chemotherapy and RT emerged as key features in achieving superior survival prediction in this large real-world dataset of glioma patients. This finding should prompt stricter clinician oversight over registry data accuracy through capture of quality assurance and peer review in clinical decision making possibly combined with central review akin to that of prospective data as we move towards meaningful predictive ability using ML approaches in glioma.

Acknowledgements

We acknowledge the contribution of Dr. N. Coleman who kindly reviewed the manuscript and provided suggestions for its relevance to the field. We also acknowledge the contributions of Ryan Proulx and the staff at Safe Software and Dr. Iulian Badragan, medical physicist at BC Cancer Surrey for their assistance with some of the data related technical aspects involved in this analysis.

References

1. Morshed RA, Young JS, Hervey-Jumper SL, Berger MS. The management of low-grade gliomas in adults. *J Neurosurg Sci*. 2019 Aug;63(4):450-457.

2. van den Bent MJ, Smits M, Kros JM, Chang SM. Diffuse Infiltrating Oligodendroglioma and Astrocytoma. *J Clin Oncol*. 2017 Jul 20;35(21):2394-2401.
3. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. *Lancet*. 2018;392:432-46.
4. Franceschi E, Mura A, Lamberti G, De Biase D, Tosoni A, Di Battista M, Argento C, Visani M, Paccapelo A, Bartolini S, Brandes AA. Concordance between RTOG and EORTC prognostic criteria in low-grade gliomas. *Future Oncol*. 2019 Aug;15(22):2595-2601.
5. Geurts M, van den Bent MJ. On high-risk, low-grade glioma: What distinguishes high from low? *Cancer*. 2019 Jan 15;125(2):174-176.
6. Corso CD, Bindra RS, Mehta MP. The role of radiation in treating glioblastoma: here to stay. *J Neurooncol*. 2017 Sep;134(3):479-485.
7. Curran WJ, Jr, Scott CB, Horton J, et al. Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *J Natl Cancer Inst*. 1993 May 5;85(9):704-710.
8. Li J, Wang M, Won M, et al. Validation and simplification of the Radiation Therapy Oncology Group recursive partitioning analysis classification for glioblastoma. *Int J Radiat Oncol Biol Phys*. 2011 Nov 1;81(3):623-630.
9. Mirimanoff RO, Gorlia T, Mason W, et al. Radiotherapy and temozolomide for newly diagnosed glioblastoma: recursive partitioning analysis of the EORTC 26981/22981-NCIC CE3 phase III randomized trial. *J Clin Oncol*. 2006 Jun 1;24(16):2563-2569.
10. Bell EH, Pugh SL, McElroy JP, Gilbert MR, Mehta M, Klimowicz AC, Magliocco A, Bredel M, Robe P, Grosu AL, Stupp R, Curran W Jr, Becker AP, Salavaggione AL, Barnholtz-Sloan JS, Aldape K, Blumenthal DT, Brown PD, Glass J, Souhami L, Lee RJ, Brachman D, Flickinger J, Won M, Chakravarti A. Molecular-Based Recursive Partitioning Analysis Model for Glioblastoma in the Temozolomide Era: A Correlative Analysis Based on NRG Oncology RTOG 0525. *JAMA Oncol*. 2017 Jun 1;3(6):784-792.
11. Harrison RA, Anderson MD, Cachia D, Kamiya-Matsuoka C, Weathers SS, O'Brien BJ, Penas-Prado M, Yung WKA, Wu J, Yuan Y, de Groot JF. Clinical trial participation of patients with glioblastoma at The University of Texas MD Anderson Cancer Center. *Eur J Cancer*. 2019 May;112:83-93.
12. Krauze AV, Mackey M, Cooley-Zgela T, Mathen P, Shih JH, et al. The Addition of Valproic Acid to Concurrent Radiation Therapy and Temozolomide Improves Patient Outcome: A Correlative Analysis of RTOG 0525, SEER and A Phase II NCI Trial. 2020. *Cancer Stud Ther J Volume* 5(1): 1-8.
13. Peeken JC, Goldberg T, Pyka T, Bernhofer M, Wiestler B, Kessel KA, Tafti PD, Nüsslin F, Braun AE, Zimmer C, Rost B, Combs SE. Combining multimodal imaging and treatment features improves machine learning-based

- prognostic assessment in patients with glioblastoma multiforme. *Cancer Med.* 2019 Jan;8(1):128-136.
14. Booth TC, Williams M, Luis A, Cardoso J, Ashkan K, Shuaib H. Machine learning and glioma imaging biomarkers. *ClinRadiol.* 2020 Jan;75(1):20-32.
15. Lu CF, Hsu FT, Hsieh KL, Kao YJ, Cheng SJ, Hsu JB, Tsai PH, Chen RJ, Huang CC, Yen Y, Chen CY. Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas. *Clin Cancer Res.* 2018 Sep 15;24(18):4429-4436.
16. Senders JT, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA, Gormley WB, Smith TR, Broekman ML, Arnaout O. An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery.* 2020 Feb 1;86(2):E184-E192.
17. Nemati M, Ansary J, Nemati N. Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. *Patterns (N Y).* 2020 Aug 14;1(5):100074. doi: 10.1016/j.patter.2020.100074. Epub 2020 Jul 4. PMID: 32835314; PMCID: PMC7334917.
18. Han D, Kolli KK, Gransar H, Lee JH, Choi SY, Chun EJ, Han HW, Park SH, Sung J, Jung HO, Min JK, Chang HJ. Machine learning based risk prediction model for asymptomatic individuals who underwent coronary artery calcium score: Comparison with traditional risk prediction approaches. *J CardiovascComputTomogr.* 2020 Mar-Apr;14(2):168-176. doi: 10.1016/j.jcct.2019.09.005. Epub 2019 Sep 23.
19. Pölsterl, S., Navab, N., and Katouzian, A., Fast Training of Support Vector Machines for Survival Analysis. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, Lecture Notes in Computer Science*, vol. 9285, pp. 243-259 (2015).
20. Pölsterl, S., Navab, N., and Katouzian, A., An Efficient Training Algorithm for Kernel Survival Support Vector Machines. *4th Workshop on Machine Learning in Life Sciences*, 23 September 2016, Riva del Garda, Italy.
21. Pölsterl, S., Gupta, P., Wang, L., Conjeti, S., Katouzian, A., and Navab, N., Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. *F1000Research*, vol. 5, no. 2676 (2016).
22. Chen, H.C., Kodell, R.L., Cheng, K.F. and Chen, J.J., 2012. Assessment of performance of survival prediction models for cancer prognosis. *BMC medical research methodology*, 12(1), p.102.
23. Bewick, Viv, Liz Cheek, and Jonathan Ball. "Statistics review 12: survival analysis." *Critical care* 8.5 (2004): 389.
24. Cox, David R (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society, Series B.* 34 (2): 187–220. JSTOR 2985181. MR 0341758.
25. Ishwaran, Hemant, et al. "Random survival forests." *The annals of applied statistics* 2.3 (2008): 841-860.
26. Paul, Jérôme, and Pierre Dupont. "Inferring statistically significant features from random forests." *Neurocomputing* 150 (2015): 471-480.
27. Bellera, Carine A., et al. "Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer." *BMC medical research methodology* 10.1 (2010): 20.
28. Van Belle, Vanya, et al. "Support vector methods for survival analysis: a comparison between ranking and regression approaches." *Artificial intelligence in medicine* 53.2(2011): 107-118.
29. Mogensen, Ulla B., Hemant Ishwaran, and Thomas A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software* 50.11 (2012): 1.
30. Tan, Yan, et al. "Improving survival prediction of high-grade glioma via machine learning techniques based on MRI radiomic, genetic and clinical risk factors." *European journal of radiology* 120 (2019): 108609.
31. Papp, László, et al. "Glioma survival prediction with combined analysis of in vivo 11C-MET PET features, ex vivo features, and patient features by supervised machine learning." *Journal of Nuclear Medicine* 59.6 (2018): 892-899.
32. Mobadersany, Pooya, et al. "Predicting cancer outcomes from histology and genomics using convolutional networks." *Proceedings of the National Academy of Sciences* 115.13 (2018): E2970-E2979.
33. Mizutani, Takuya, et al. "Optimization of treatment strategy by using a machine learning model to predict survival time of patients with malignant glioma after radiotherapy." *Journal of Radiation Research* 60.6 (2019): 818-824.
34. Zhao YY, et al. A Nomogram for Predicting Individual Prognosis of Patients with Low-Grade Glioma. *World Neurosurg.* 2019.
35. Brandel MG, et al. Survival trends of oligodendroglial tumor patients and associated clinical practice patterns: a SEER-based analysis. *J Neurooncol.* 2017.
36. Yang Y, et al. Prognostic Nomograms for Primary High-Grade Glioma Patients in Adult: A Retrospective Study Based on the SEER Database. *Biomed Res Int.* 2020.
37. Vock, David M., et al. "Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting." *Journal of biomedical informatics* 61 (2016): 119-131.